

Noeuds de modélisation 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. À 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 Introduction à la modélisation 8

Création du flux	10
Navigation dans le modèle	15
Évaluation du modèle	20
Scoring des enregistrements	24
Récapitulatif	25

3 Présentation de la modélisation 26

Description des noeuds de modélisation	26
Création de modèles découpés	33
Découpage et partitionnement	36
Noeuds de modélisation prenant en charge les modèles découpés	37
Fonctions affectées par le découpage	38
Options de champs des noeuds de modélisation	39
Utilisation des champs d'effectif et de pondération	42
Options d'analyse des noeuds de modélisation	44
Scores de propension	47

Nuggets de modèle	48
Liens de modèle	49
Remplacement d'un modèle	51
Palette Modèles	52
Navigation dans les nuggets de modèle	55
Récapitulatif /Informations sur les nuggets de modèle	56
Importance des valeurs prédites	57
Modèles pour des ensembles	60
Nuggets de modèle pour les modèles découpés	68
Utilisation de nuggets de modèle dans les flux	70
Régénération d'un noeud de modélisation	72
Importation et exportation de modèles au format PMML	72
Publication des modèles pour un adaptateur de scoring	76
Modèles bruts	77

4 Modèles de filtrage 78

Filtrage des champs et des enregistrements	78
Noeud Sélection de fonction	78
Paramètres des modèles Sélection de fonction	79
Options de sélection des caractéristiques	81
Nuggets de modèles Sélection de fonction	83
Résultats du modèle Sélection de fonction	83
Sélection des champs en fonction de leur importance	85
Génération d'un filtre à partir d'un modèle Sélection de fonction	85
Noeud Détection des anomalies	86
Options des modèles Détection des anomalies	88
Options expert de détection d'anomalies	89
Nuggets de modèles Détection des anomalies	91
Détails relatifs aux modèles Détection des anomalies	92
Récapitulatif du modèle Détection des anomalies	93
Paramètres des modèles Détection des anomalies	94

5 Noeuds de modélisation automatisés 96

Noeud de modélisation automatisé - Paramètres d'algorithme	97
Noeuds de modélisation automatisés - Règles d'arrêt	98
Noeud Classificateur automatique	99
Noeud Classificateur automatique - Options du modèle	101
Noeud Classificateur automatique - Options expert	103

Coûts de mauvaise réaffectation	106
Noeud Classificateur automatique - Options de suppression	107
Noeud Classificateur automatique - Options des paramètres	108
Noeud Numérisation automatique	110
Noeud Numérisation automatique - Options du modèle	111
Noeud Numérisation automatique - Options expert	114
Noeud Numérisation automatique - Options des paramètres	116
Noeud Classification automatique	117
Noeud Classification - Options du modèle	118
Noeud Classification - Options expert	120
Noeud Classification - Options de suppression	121
Nuggets de modèle automatisés	122
Génération de noeuds et de modèles	125
Génération de graphiques Evaluation	125
Graphiques d'évaluation	126

6 Arbres décision

128

Modèles d'arbre décision	128
Générateur d'arbres interactifs	131
Développement et élagage de l'arbre	133
Définition de divisions personnalisées	134
Détails et substitutions d'une division	137
Personnalisation de l'affichage d'un arbre	138
Gains	140
Risques	149
Enregistrement de modèles d'arbre et de résultats	149
Génération de noeuds Filtrer et Sélectionner	154
Génération d'un noeud Ensemble de règles à partir d'un noeud Arbre décision	154
Création directe d'un modèle d'arbre	155
Noeuds Arbre de décision	157
Noeud Arbre C&RT	158
Noeud CHAID	159
Noeud QUEST	160
Options des champs du noeud Arbre décision	161
Options de création du noeud Arbre décision	162
Noeud Arbre décision - Options du modèle	177
Noeud C5.0	180
Noeud C5.0 - Options du modèle	181

Nuggets de modèle Arbre de décision	184
Nuggets de modèle d'arbre unique	185
Nuggets de modèle pour l'amélioration, le regroupement et les ensembles très volumineux	197
Nugget Modèle d'ensemble de règles.	198
Onglet Modèle d'ensemble de règles.	200
Importation de projets à partir d'AnswerTree 3.0	201
7 Modèles de réseau Bayésien	203
Noeud Réseau Bayésien.	203
Options du modèle de noeud Réseau Bayésien	205
Noeud Réseau Bayésien - Options expert	207
Nuggets de modèle de Réseau Bayésien	209
Paramètres de modèle de Réseau Bayésien	210
Récapitulatif du modèle de Réseau Bayésien.	212
8 Réseaux de neurones	213
Le modèle des réseaux de neurones.	214
Utilisation des réseaux de neurones avec les flux hérités	215
Objectifs	215
Bases	217
Règles d'arrêt	218
Ensembles	219
Avancés	220
Options du modèle	221
Récapitulatif du modèle	222
Importance des valeurs prédites	223
Valeurs prédites en fonction des valeurs observées	224
Classification	224
Réseau	226
Paramètres	227
9 Liste de décision	229
Options du modèle Liste de décision.	235
Noeud Liste de décision - Options expert	237

Nugget du modèle Liste de décision	238
Paramètres du nugget du modèle Liste de décision	240
Decision List Viewer	240
Panneau du modèle de travail	241
Onglet Alternatives	243
Onglet Instantanés	245
Utilisation de l'Decision List Viewer	247

10 Modèles statistiques

267

Noeud linéaire	268
Modèles linéaires	269
Objectifs	270
Bases	272
Choix du modèle	273
Ensembles	275
Avancé	276
Options de modèle	276
Récapitulatif de modèle	277
Préparation automatique des données	278
Importance des variables indépendantes	279
Valeurs prévues en fonction des valeurs observées	280
Résidus	281
Valeurs éloignées	282
Effets	283
Coefficients	284
Moyennes estimées	286
Récapitulatif de création de modèle	287
Paramètres	288
Noeud Logistique	288
Noeud Logistique - Options du modèle	290
Ajout de caractéristiques à un modèle de régression logistique	295
Noeud Logistique - Options expert	297
Options de convergence de la régression logistique	298
Sorties avancées du noeud Régression logistique	299
Régression logistique - Options pas à pas	301
Nugget de modèle logistique	302
Modèle de nugget logistique - Détails	303
Nugget de modèle logistique - Récapitulatif	305
Nugget de modèle logistique - Paramètres	306
Nugget de modèle Logistique - Sorties avancées	308

Noeud ACP/Analyse factorielle	310
Noeud ACP/Analyse factorielle – Options du modèle	311
Noeud ACP/Analyse factorielle – Options expert	312
Noeud ACP/Analyse factorielle – Options de rotation	313
Nugget de modèle ACP/Analyse factorielle	314
Equations de nugget de modèle ACP/Analyse factorielle	315
Nugget de modèle ACP/Analyse factorielle - Récapitulatif	315
Nugget de modèle ACP/Analyse factorielle - Sorties avancées	317
Noeud discriminant	318
Noeud discriminant - Options de modèle	319
Noeud discriminant - Options Expert	320
Noeud discriminant - Options de sortie	321
Noeud discriminant - Options pas à pas	323
Nugget du modèle discriminant	324
Sortie avancée du nugget du modèle discriminant	325
Paramètres du nugget du modèle discriminant	325
Récapitulatif du nugget du modèle discriminant	326
Noeud Modèles linéaires généralisés	327
Noeud Modèles linéaires généralisés - Options de champs	329
Noeud Modèles linéaires généralisés - Options du modèle	330
Noeud Modèles linéaires généralisés - Options expert	332
Modèles linéaires généralisés - Itérations	335
Modèles linéaires généralisés - Sorties avancées	337
Modèle de nugget Modèles linéaires généralisés	339
Nugget de modèle Modèles linéaires généralisés - Sorties avancées	340
Nugget de modèle Modèles linéaires généralisés - Paramètres	340
Modèle de nugget Modèles linéaires généralisés - Récapitulatif	341
Noeud MMLG	342
Modèles mixtes linéaires généralisés	342
Cible	345
Effets fixes	348
Effets aléatoires	351
Pondération et décalage	354
Options de création	355
Général	356
Moyennes estimées	357
Vue du modèle	358
Noeud de Cox	371
Noeud de Cox - Options de champs	372
Noeud de Cox - Options de modèle	373
Noeud de Cox - Options expert	376
Options des paramètres du noeud de Cox	379

Nugget de modèle Cox	380
Paramètres de sortie de la régression de Cox	381
Sorties avancées du noeud Régression de Cox	381

11 Modèles de classification 382

Noeud Kohonen	383
Noeud Kohonen - Options du modèle	385
Noeud Kohonen - Options expert	387
Nuggets de modèle Kohonen	389
Récapitulatif du modèle Kohonen.	389
Noeud K-Means	390
Noeud K-Means - Options du modèle.	391
Noeud K-Means - Options expert.	392
Nuggets du modèle K-Means	393
Récapitulatif du modèle K-Means	394
Noeud Classification TwoStep	395
Noeud Classification TwoStep - Options du modèle	396
Nuggets de modèle de classification TwoStep	398
Récapitulatif du modèle TwoStep.	398
Le visualiseur de classes	399
Visualiseur de classes - Onglet Modèle	400
Navigation dans le visualiseur de classes	410
Génération de diagrammes à partir de modèles de classe	412

12 Règles d'association 415

Données tabulaires et données transactionnelles	416
Noeud Apriori	417
Noeud Apriori - Options du modèle	418
Options expert du noeud Apriori.	419
Noeud CARMA	421
Noeud CARMA - Options des champs	422
Noeud CARMA - Options du modèle.	425
Noeud CARMA - Options expert.	426
Nuggets du modèle de règle d'association	427
Détails du nugget de modèle de règle d'association.	427
Paramètres des nuggets de modèles de règle d'association	434
Récapitulatif du nugget du modèle de règle d'association	436

Génération d'un ensemble de règles à partir d'un nugget de modèle d'association	437
Génération d'un modèle filtré.	438
Scoring des règles d'association	439
Déploiement des modèles d'association	440
Noeud Séquence	443
Noeud Séquence - Options de champs	444
Noeud Séquence - Options modèle	446
Noeud Séquence - Options expert	447
Nuggets de modèles de séquences	449
Détails du nugget de modèle de séquence.	451
Paramètres du nugget de modèle de séquence	453
Récapitulatif du nugget de modèle de séquence	454
Génération d'un super noeud Règle à partir d'un nugget de modèle de séquence	455

13 Modèles de séries temporelles

458

Prévoir, à quoi ça sert ?	458
Séries temporelles.	458
Caractéristiques des séries temporelles	459
Fonctions d'autocorrélation et d'autocorrélation partielle	463
Transformations de série	464
Série de valeurs prédites	465
Noeud de modélisation Séries temporelles	465
Conditions requises	466
Options du modèle de séries temporelles.	469
Critères du modélisateur expert de séries temporelles.	471
Critères du lissage exponentiel des séries temporelles	473
Critères ARIMA pour les séries temporelles.	474
Fonctions de transfert	476
Gestion des valeurs éloignées	478
Génération de modèles de séries temporelles.	479
Génération de plusieurs modèles.	479
Utilisation des modèles de séries temporelles à des fins de prévision	479
Nouvelle estimation et prévision	480
Nugget du modèle Séries temporelles	480
Paramètres du modèle Séries temporelles.	484
Résidus du modèle de séries temporelles	485
Récapitulatif du modèle de séries temporelles.	486
Paramètres du modèle de séries temporelles.	487

14 Modèles de noeud Réponse en auto-apprentissage 488

Noeud MRAA	488
Noeud MRAA - Options de champs	489
Noeud MRAA - Options du modèle	490
Noeud MRAA - Options de paramètres	492
Nuggets de modèle MRAA	493
Paramètres du modèle MRAA	495

15 Modèles Support Vector Machine 498

A propos de SVM	498
Fonctionnement de SVM	498
Affinement d'un modèle SVM	500
Noeud SVM	501
Noeud SVM - Options modèle	502
Noeud SVM - Options expert	502
Nugget de modèle SVM	504
Paramètres du modèle SVM	505

16 Modèles d'agrégation suivant le saut minimum 507

Noeud KNN	507
Noeud KNN - Options des objectifs	508
Paramètres de noeud KNN	509
Modèle de nugget KNN	519
Vue du modèle	519
Paramètres du modèle KNN	527

Annexe

A Remarques 529

Index 532

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 sur p. 26.](#)

- **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*.

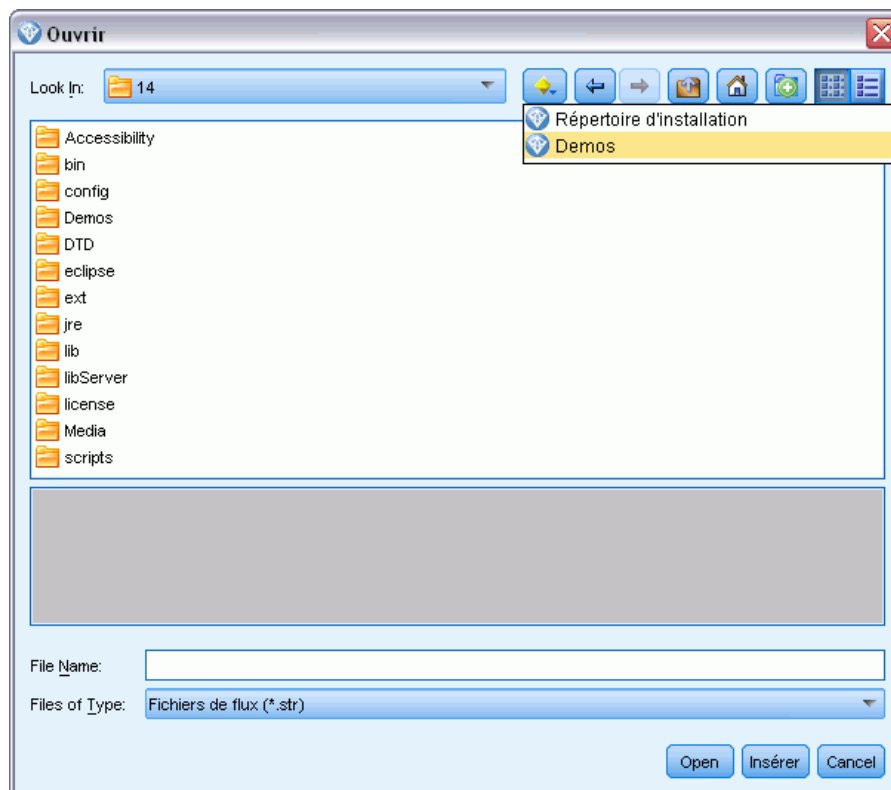
Dossier Demos

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

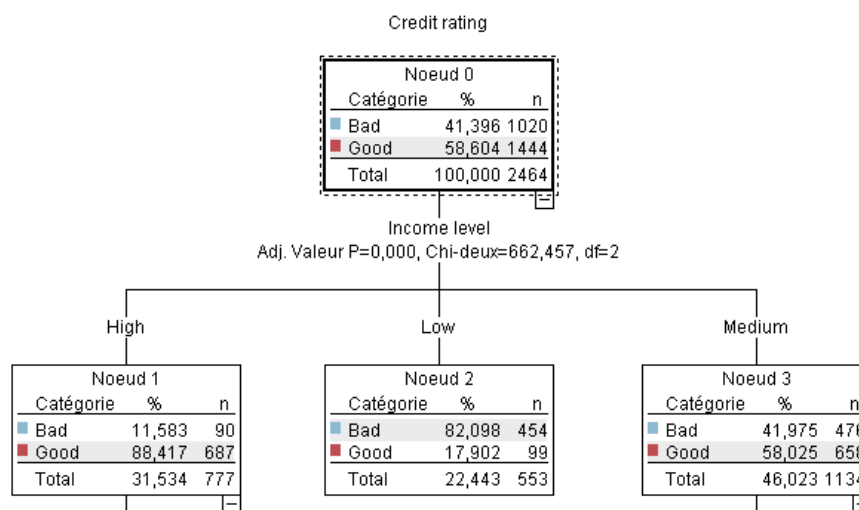


Introduction à la modélisation

Un modèle est un ensemble de règles, de formules, ou d'équations pouvant être utilisées pour prédire un résultat en fonction d'un ensemble de champs ou de variables d'entrée. Par exemple, une institution financière peut utiliser un modèle pour prédire si les emprunteurs représentent un risque important ou peu de risque, en fonction des informations déjà connues sur le passé de ces emprunteurs.

La capacité à prédire un résultat est l'objectif central de l'analyse prédictive, et la compréhension du processus de modélisation est essentielle pour l'utilisation de IBM® SPSS® Modeler.

Figure 2-1
Modèle d'arbre décision simple



Cet exemple utilise un modèle d'**arbre décision** qui classe les enregistrements (et prédit une réponse) à l'aide d'une série de règles de décisions, par exemple :

SI revenu = Moyen
ET cartes <5
ALORS -> 'Bon'

Bien que cet exemple utilise un modèle CHAID (Chi-Squared Automatic Interaction Detection), il est destiné à fournir une introduction générale, et la plupart des concepts s'appliquent globalement aux autres types de modélisation dans SPSS Modeler.

Pour comprendre tous les modèles, vous devez d'abord comprendre les données qu'ils contiennent. Les données de cet exemple contiennent des informations sur les clients d'une banque. Les champs suivants sont utilisés :

Nom de champ	Description
Conditions_crédit	Conditions de crédit : 0=Mauvaises, 1=Bonnes, 9=valeurs manquantes
Age	Age en années
Revenu	Niveau de revenu : 1=bas, 2=moyen, 3=élevé
Cartes_crédit	Nombre de cartes de crédit possédées : 1=Moins de cinq, 2=Cinq ou plus
Éducation	Niveau d'éducation : 1=Université, 2=Lycée
Prêts_voiture	Nombre de prêts voiture en cours : 1=aucun ou un, 2=Plus de deux

La banque gère une base de données contenant des informations sur les clients qui ont contracté un prêt, notamment sur le respect de leur engagement de remboursement (conditions de crédit = bonnes) ou le non-respect de leur engagement (conditions de crédit = mauvaises). À l'aide de ces données, la banque peut créer un modèle qui lui permettra de prédire les probabilités de remboursement des futurs emprunteurs.

À partir d'un modèle d'arbre de décision, vous pouvez analyser les caractéristiques de deux groupes de clients et prédire les risques de non-remboursement.

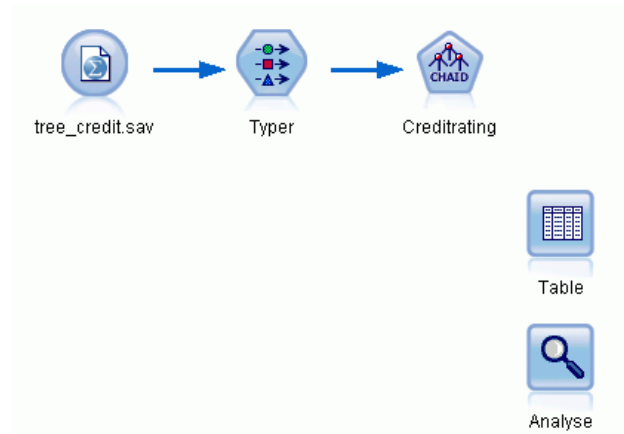
Cet exemple utilise le flux nommé *modelingintro.str*, disponible dans le dossier *Demos* du sous-dossier des *flux*. Le fichier de données est *tree_credit.sav*. [Pour plus d'informations, reportez-vous à la section Dossier Demos dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Regardons le flux de plus près.

- ▶ Dans le menu principal, sélectionnez les options suivantes :
Fichier > Ouvrir un flux
- ▶ Cliquez sur l'icône de la pépite d'or dans la barre d'outils de la boîte de dialogue Ouvrir et choisissez le dossier Demos.
- ▶ Double-cliquez sur le dossier des *flux*.
- ▶ Double-cliquez sur le fichier *modelingintro.str*.

Création du flux

Figure 2-2
Modélisation du flux



Pour construire un flux qui va créer un modèle, vous avez besoin d'au moins trois éléments :

- Un noeud source qui lit les données issues d'une source externe, dans ce cas un fichier de données IBM® SPSS® Statistics.
- Un noeud source ou Typer qui spécifie les propriétés des champs, telles que le niveau de mesure (le type de données contenues dans le champ) et le rôle de chaque champ en tant que cible ou entrée dans la modélisation.
- Un noeud de modélisation qui génère un nugget de modèle lors de l'exécution du flux.

Dans cet exemple, nous utilisons un noeud de modélisation CHAID. CHAID, ou Chi-Squared Automatic Interaction Detection, est une méthode de classification qui crée des arbres de décision à l'aide d'un type de statistiques spécifique connu sous le nom de statistiques du Khi-deux et qui permet de définir les meilleurs endroits auxquels opérer le découpage dans l'arbre de décision.

Si les niveaux de mesure sont spécifiés dans le noeud source, le noeud Typer distinct peut être éliminé. D'un point de vue fonctionnel, le résultat est le même.

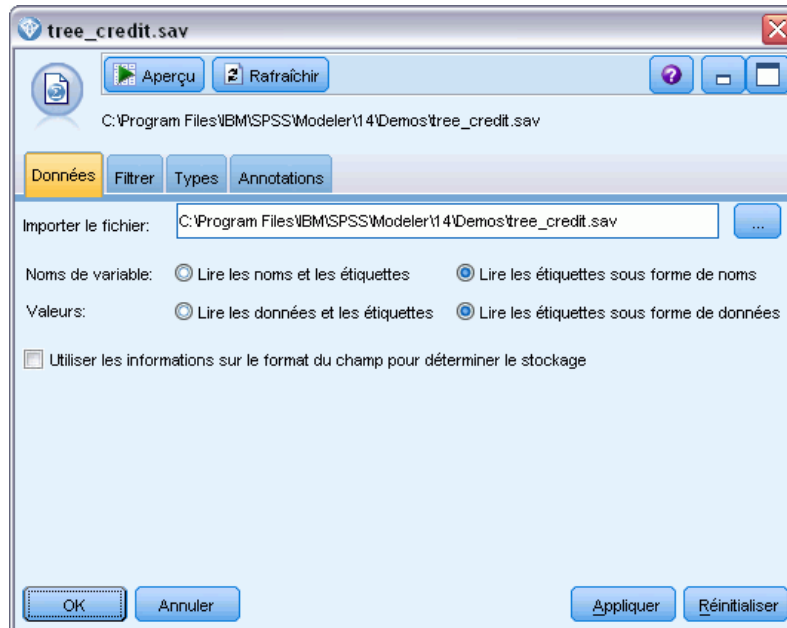
Ce flux comporte également des noeuds Table et Analyse qui seront utilisés pour afficher les résultats de scoring après la création du nugget de modèle et son ajout au flux.

Le noeud Statistics lit les données au format SPSS Statistics à partir du fichier de données *tree_credit.sav*, qui est installé dans le dossier *Demos*. (Une variable spéciale nommée *\$CLEO_DEMOS* est utilisée pour faire référence à ce dossier sous l'installation IBM® SPSS®)

Modeler actuelle. Ainsi, le chemin sera toujours valide, quelque soit le dossier d'installation actuel ou la version.)

Figure 2-3

Lecture des données avec un noeud source Statistics



Le noeud Typer définit le **niveau de mesure** pour chaque champ. Le niveau de mesure est une catégorie qui indique le type de données du champ. Notre fichier de données source utilise trois niveaux de mesure différents.

Un champ **Continu** (comme le champ *Age*) contient des valeurs numériques continues, alors qu'un champ **Nominal** (comme le champ *Conditions de crédit*) contient deux valeurs distinctes minimum, par exemple *Mauvaises*, *Bonnes*, ou *Pas d'antécédents de crédit*. Un champ **Ordinal**

(comme le champ *Niveau de revenu*) décrit les données avec différentes valeurs distinctes ayant un ordre inhérent—dans ce cas *Bas*, *Moyen* et *Elevé*.

Figure 2-4

Définition des champs cibles et des champs d'entrées avec le noeud *Typer*



Pour chaque champ, le noeud *Typer* spécifie également un **rôle**, qui indique le rôle que joue chaque champ dans la modélisation. Le rôle est défini sur *Cible* pour le champ *Conditions de crédit*, qui indique si un client donné a remboursé ou non son prêt. Il s'agit de la **cible**, soit le champ dont vous souhaitez prédire la valeur.

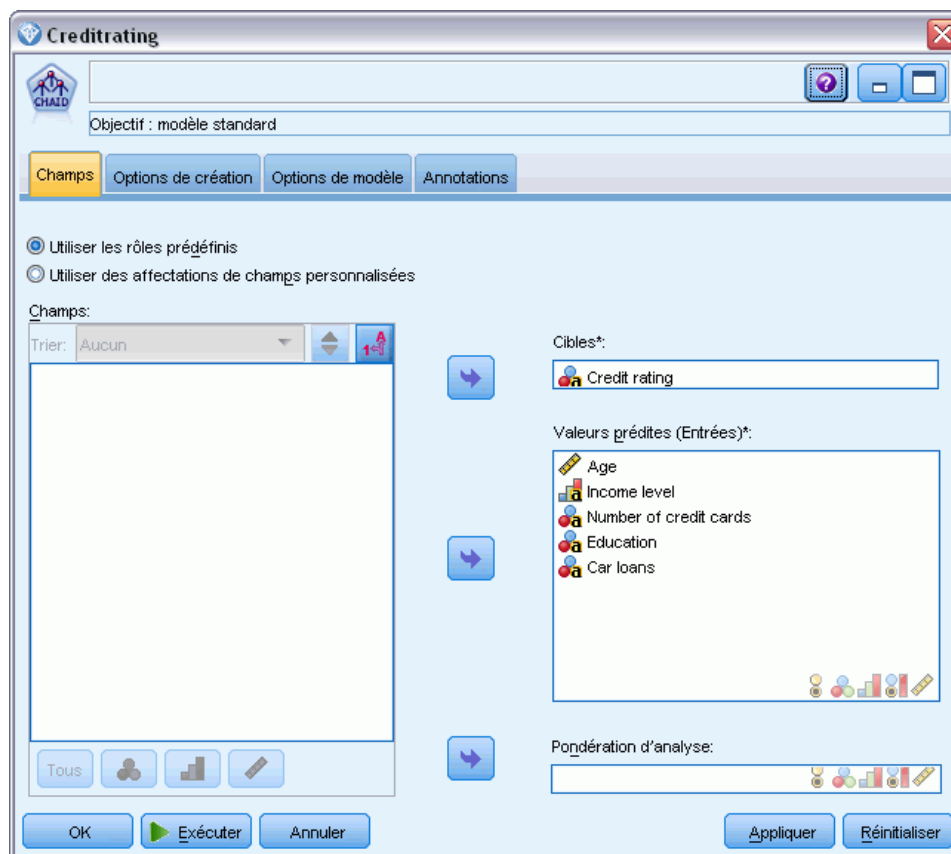
Le rôle est défini sur *Entrée* pour les autres champs. Les champs d'entrée sont quelquefois désignés sous le nom de **variables indépendantes**, ou champs dont les valeurs sont utilisées par l'algorithme de modélisation afin de prévoir la valeur du champ cible.

Le noeud de modélisation CHAID génère le modèle.

Sur l'onglet Champs du noeud de modélisation, l'option Utiliser les rôles prédéfinis est sélectionnée, ce qui signifie que la cible et les entrées indiquées dans le noeud *Typer* seront utilisées. Vous pouvez modifier les rôles de champ à ce stade, mais pour cet exemple, nous les utiliserons tels quels.

- Cliquez sur l'onglet Options de création.

Figure 2-5
Noeud de modélisation CHAID - Onglet Champs



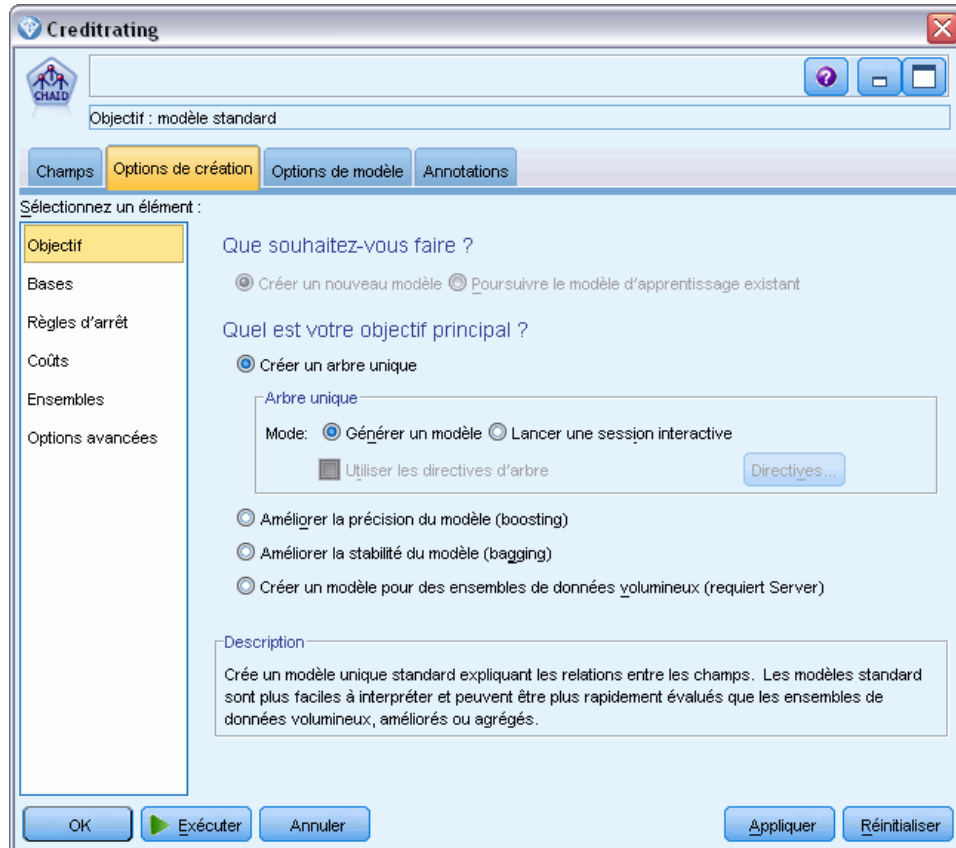
Plusieurs options sont disponibles ici dans lesquelles vous pouvez spécifier le type de modèle que vous voulez créer.

Nous voulons un tout nouveau modèle, donc nous utiliserons l'option par défaut Créer un nouveau modèle.

Nous voulons également un seul modèle d'arbre décision standard sans aucune amélioration, donc nous conserverons l'option d'objectif par défaut Créer un seul arbre.

Bien que vous puissiez lancer une session de modélisation interactive qui vous permet d'ajuster le modèle, cet exemple génère simplement un modèle à l'aide du paramètre de mode par défaut Générer le modèle.

Figure 2-6
Noeud de modélisation CHAID - Onglet Options de création



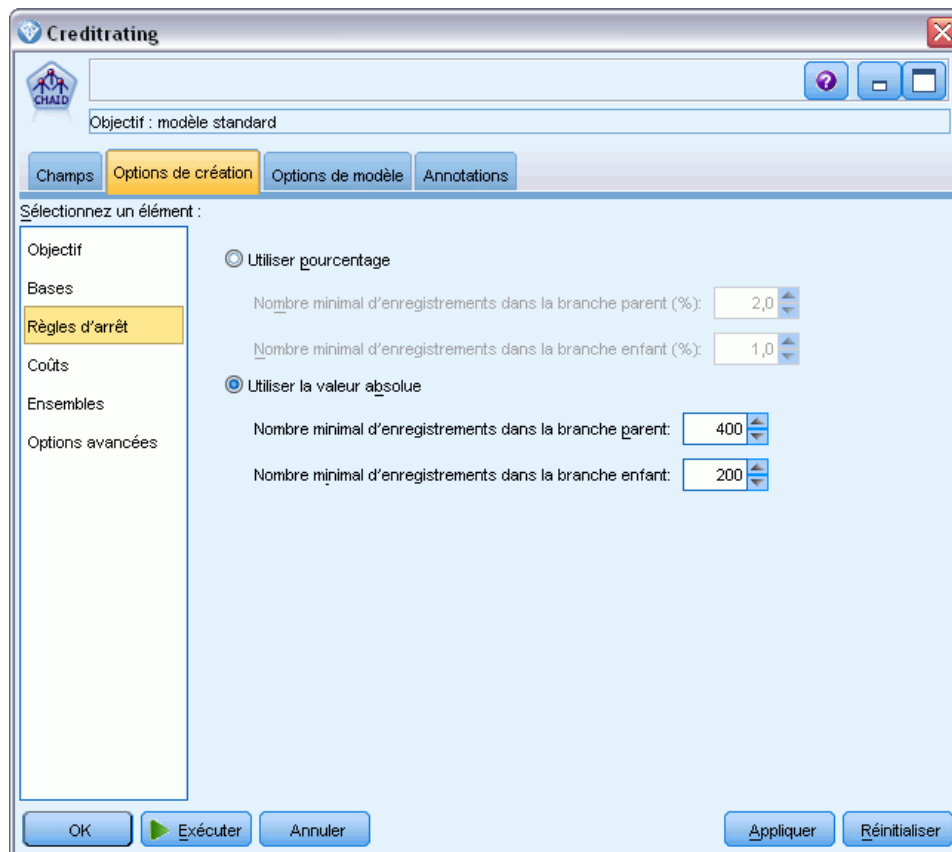
Dans cet exemple, pour que l'arbre reste simple, nous limiterons sa croissance en augmentant le nombre minimum d'observations pour les noeuds parents et enfants.

- ▶ Dans l'onglet Options de création, sélectionnez Règles d'arrêt dans le panneau de gauche du navigateur.
- ▶ Sélectionnez l'option Utiliser la valeur absolue.
- ▶ Définissez Enregistrements minimum dans la branche parent sur 400.

- Définissez Enregistrements minimum dans la branche enfant sur 200.

Figure 2-7

Définition des critères d'arrêt pour la création d'un arbre de décision



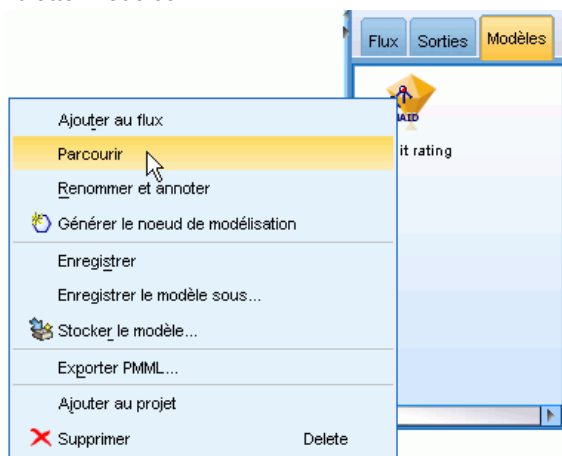
Nous pouvons utiliser toutes les autres options par défaut pour cet exemple, par conséquent, cliquez sur Exécuter pour créer le modèle. (Vous pouvez également cliquer avec le bouton droit de la souris sur le noeud et choisir Exécuter dans le menu contextuel ou sélectionner le noeud et choisir Exécuter dans le menu Outils.)

Navigation dans le modèle

Lorsque l'exécution se termine, le nugget de modèle est ajouté à la palette Modèles dans le coin supérieur droit de la fenêtre de l'application, et est aussi placé dans l'espace de travail du flux avec un lien vers le noeud de modélisation à partir duquel il a été créé. Pour consulter les détails du

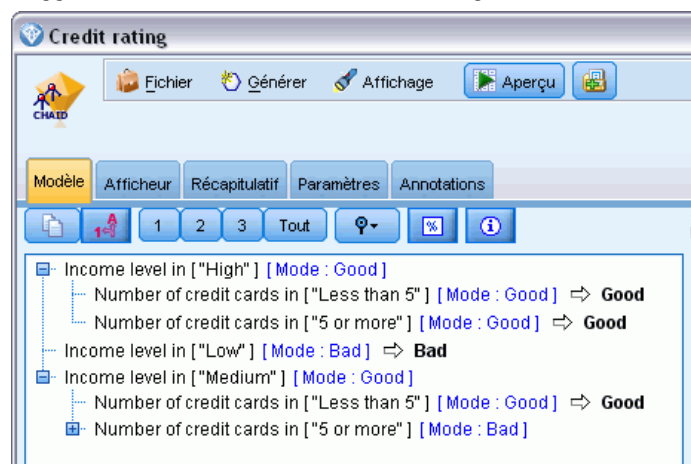
modèle, cliquez avec le bouton droit de la souris sur le nugget de modèle Parcourir (dans la palette des modèles) ou Modifier (dans l'espace de travail).

Figure 2-8
Palette Modèles



Dans le cas du nugget CHAID, l'onglet Modèle affiche les détails sous la forme d'un ensemble de règles. Il s'agit essentiellement d'une série de règles pouvant être utilisées pour affecter des enregistrements individuels à des noeuds enfant, en fonction des valeurs des différents champs d'entrée.

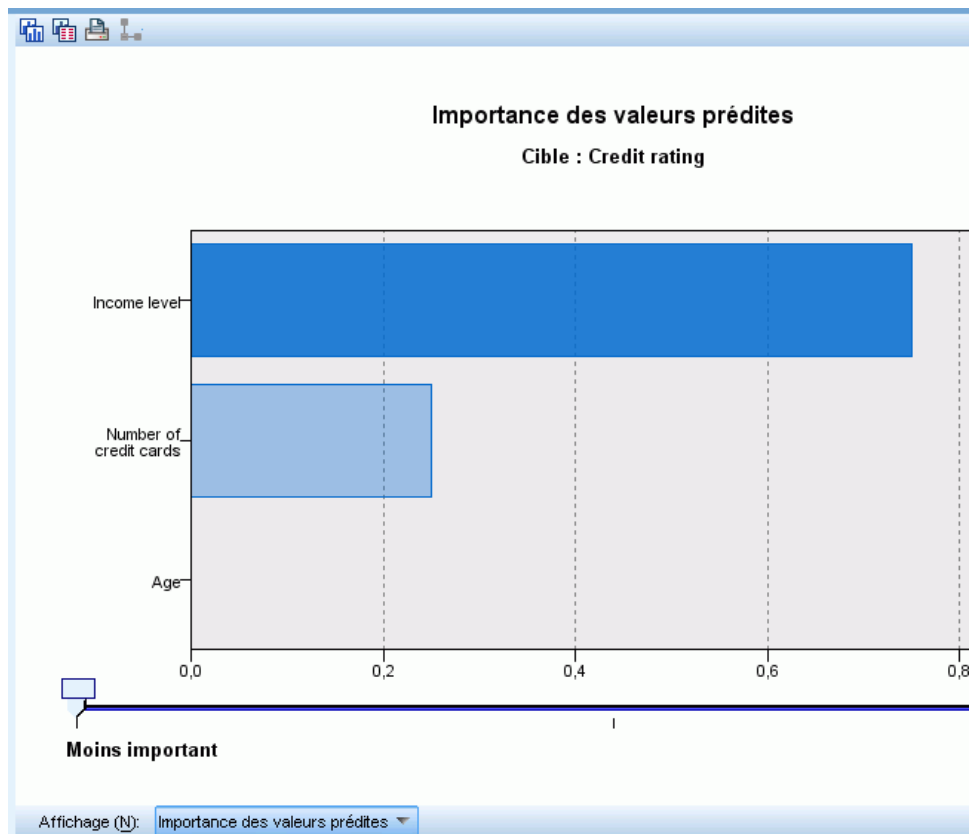
Figure 2-9
Nugget de modèle CHAID, ensemble de règles



Pour chaque noeud de terminal d'arbre de décision, c'est-à-dire ces noeuds Arbre qui ne sont pas plus divisés, une prévision de *Bon* ou *Mauvais* est renvoyée. Dans chaque cas la prédiction est déterminée par le **noeud**, ou par la réponse la plus courante pour les enregistrements qui sont compris dans ce noeud.

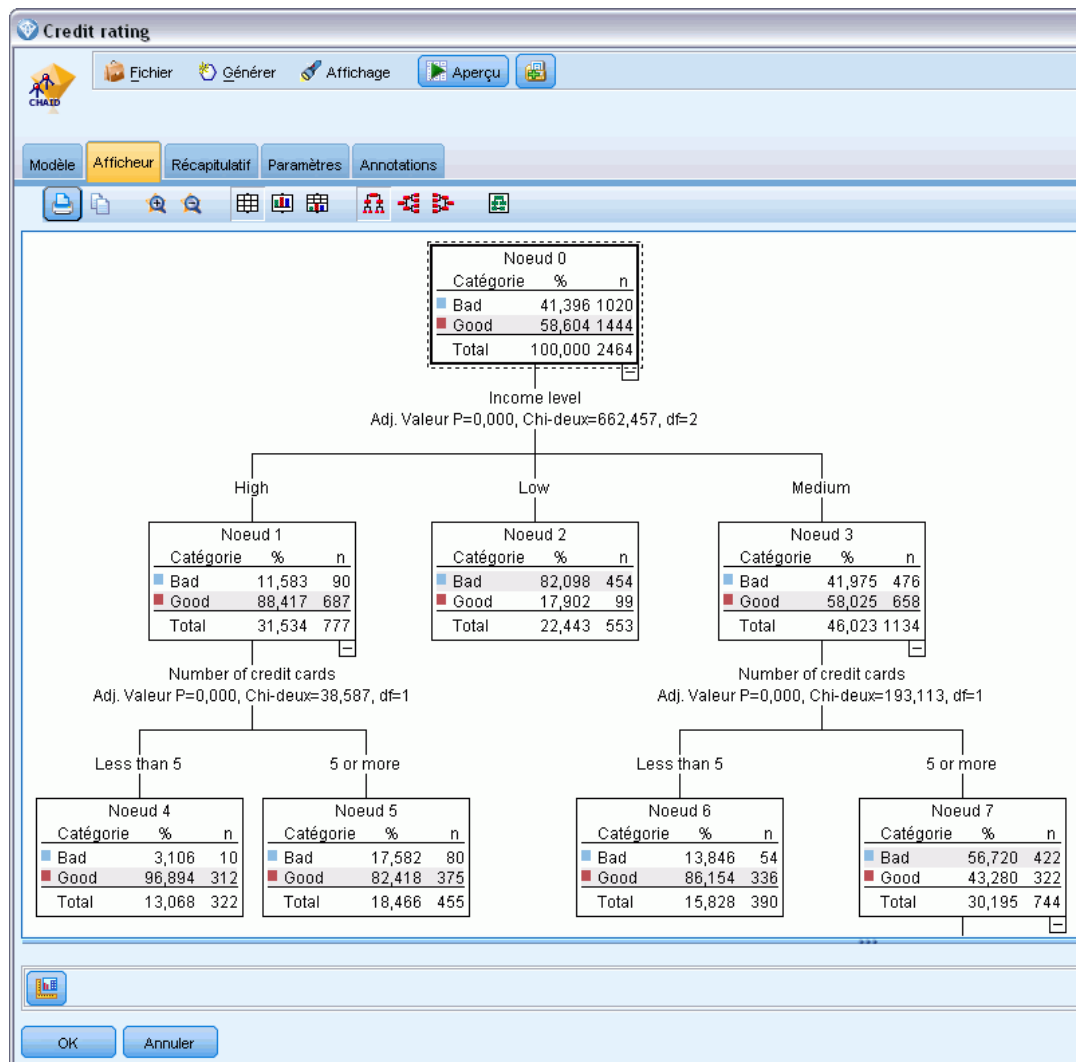
À droite de l'ensemble de règles, l'onglet Modèle affiche le tableau d'importance des variables indépendantes qui montre l'importance relative de chaque variable indépendante dans l'estimation du modèle. Nous pouvons observer que le *niveau de revenu* est le critère plus important dans ce cas et que le seul autre facteur intéressant est le *Nombre de cartes de crédit*.

Figure 2-10
Graphique de l'importance des variables indépendantes



L'onglet Afficheur dans le nugget de modèle affiche le même modèle sous la forme d'un arbre, avec un noeud à chaque point de décision. Utilisez les commandes du Zoom sur la barre d'outils pour effectuer un zoom avant sur un noeud spécifique ou un zoom arrière pour afficher une plus grande partie de l'arbre.

Figure 2-11
Onglet Afficheur dans le nugget de modèle, avec zoom arrière sélectionné



Si l'on regarde la partie supérieure de l'arbre, le premier noeud (Noeud 0) propose un récapitulatif de tous les enregistrements dans l'ensemble de données. Un peu plus de 40 % des observations de cet ensemble de données sont classées comme risquées. Il s'agit d'une proportion élevée. Voyons si l'arbre peut nous donner des informations sur les facteurs responsables.

Nous pouvons observer que la première division se situe au niveau du *Niveau de revenu*. Les enregistrements dans lesquels le niveau de revenu se trouve dans la catégorie *Bas* sont affectés au Noeud 2 et il n'est pas surprenant de voir que cette catégorie contient le plus fort pourcentage

de non-remboursement de prêts. Il est évident qu'accorder un prêt aux clients de cette catégorie présente un risque élevé.

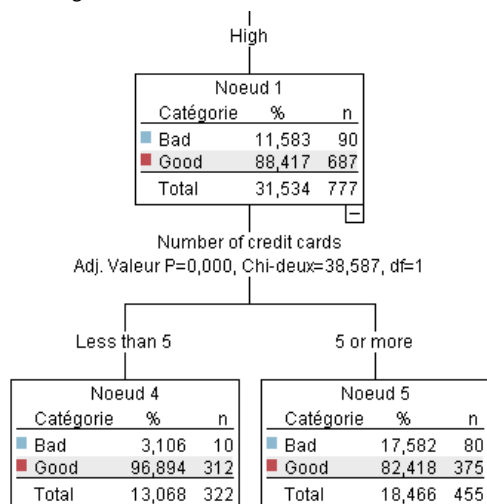
Cependant, 16% des clients de cette catégorie ont, en réalité; remboursé leur prêt. Par conséquent, cette prévision n'est pas toujours exacte. Aucun modèle ne peut réellement prédire toutes les réponses, mais un bon modèle doit vous permettre de prédire la réponse *la plus probable* pour chaque enregistrement, sur la base des données disponibles.

De la même façon, si l'on observe les clients avec un revenu élevé (Noeud 1), on s'aperçoit que la grande majorité (89 %) présente un risque peu élevé. Mais plus de 1 clients sur 10 n'a pas remboursé son prêt. Est-il possible d'affiner nos critères de prêt pour diminuer le risque ?

Veillez noter que le modèle a divisé ces clients en deux sous-catégories (noeuds 4 et 5), en fonction du nombre de cartes de crédit possédées. Pour les clients à revenu élevé, si nous prêtons uniquement à ceux possédant moins de 5 cartes de crédit, nous pouvons faire passer notre taux de succès de 89% à 97%, soit un résultat encore plus satisfaisant.

Figure 2-12

Affichage sous forme d'arbre des clients à revenu élevé

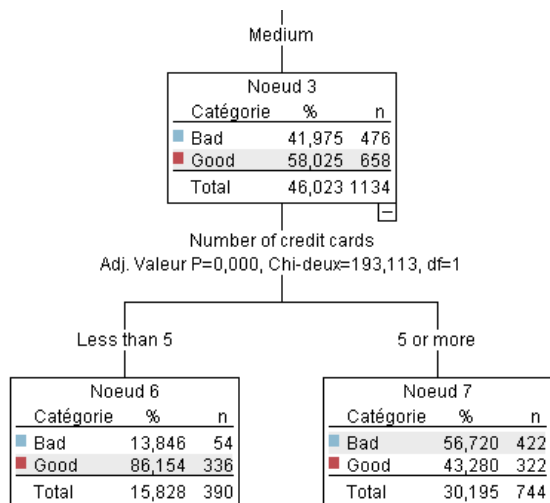


Mais qu'en est-il des clients appartenant à la catégorie Revenu moyen (Noeud 3) ? Ils sont encore plus fortement divisés entre les conditions Bonnes et Mauvaises.

De nouveau, les sous-catégories (Noeuds 6 et 7 dans ce cas) peuvent nous aider. Cette fois, prêter uniquement aux clients avec des revenus moyens et possédant moins de 5 cartes de crédit fait passer le pourcentage de conditions Bonnes de 58% à 85%, soit une augmentation importante.

Figure 2-13

Affichage sous forme d'arbre des clients à revenu moyen



Nous avons appris que chaque enregistrement contenu dans ce modèle sera attribué à un noeud spécifique et recevra une prévision *Bonne* ou *Mauvaise* en fonction des réponses les plus courantes de ce noeud.

Ce processus consistant à affecter des prédictions à des enregistrements individuels s'appelle le **scoring**. En effectuant le scoring des mêmes enregistrements utilisés pour estimer le modèle, il est possible d'évaluer sa précision sur les données d'apprentissage, données dont nous connaissons le résultat. Examinons comment effectuer cette opération.

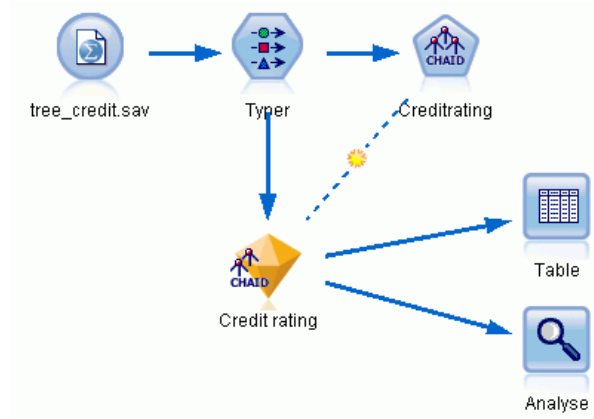
Evaluation du modèle

Nous avons parcouru le modèle pour comprendre le fonctionnement du scoring. Mais pour évaluer sa *précision*, nous devons déterminer le score de certains enregistrements et comparer les réponses prédites par le modèle aux résultats réels. Nous allons déterminer le score des mêmes

enregistrements qui ont été utilisés pour estimer le modèle, ce qui nous permet de comparer les réponses observées et les réponses prédites.

Figure 2-14

Liez le nugget de modèle au noeuds de sortie pour l'évaluation du modèle.



- Pour voir les scores ou les prédictions, attachez le noeud Table au nugget de modèle, double-cliquez sur le noeud Table et cliquez sur Exécutez.

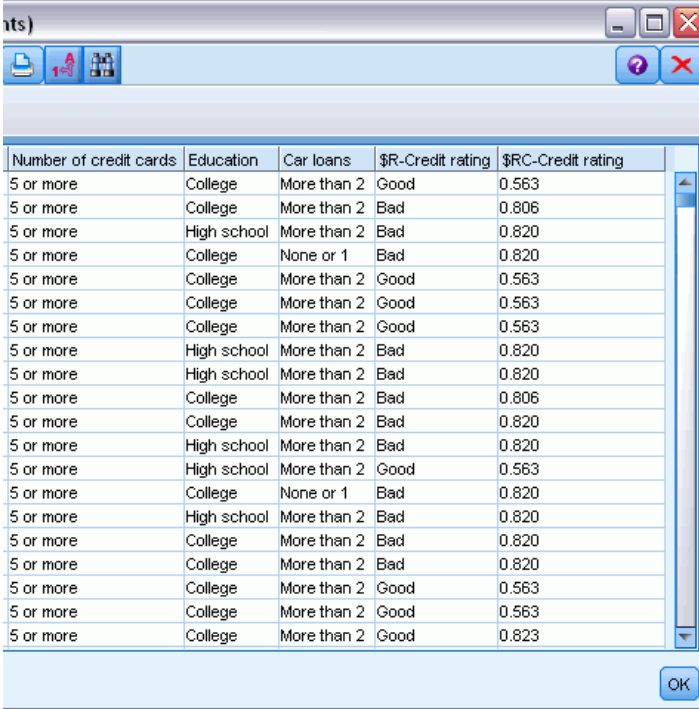
La table affiche les scores prédits dans un champ nommé *\$R-Credit rating*, qui a été créé par le modèle. Nous pouvons comparer ces valeurs au champ *Conditions de crédit* d'origine qui contient les réponses réelles.

Par convention, les noms des champs générés au cours du scoring sont déterminés en fonction du champ cible, mais avec un préfixe standard tel que *\$R-* pour les prédictions ou *\$RC-* pour les valeurs de confiance. Différents types de modèles utilisent différents ensembles de préfixes.

Une **valeur de confiance** est la propre estimation du modèle, sur une échelle de 0,0 à 1,0, de la précision de chaque valeur prédite.

Figure 2-15

Table affichant les scores générés et les valeurs de confiance



Number of credit cards	Education	Car loans	\$R-Credit rating	\$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823

Comme prévu, la valeur prédite correspond aux réponses réelles pour de nombreux enregistrements mais pas pour tous. La raison à cela est que chaque noeud terminal CHAID comporte un ensemble de réponses. La prédiction correspond à la réponse la *plus courante*, mais elle sera fautive pour toutes les autres réponses de ce noeud. (Pensez à la minorité de 16% de clients à faible revenu qui ont remboursé leur prêt).

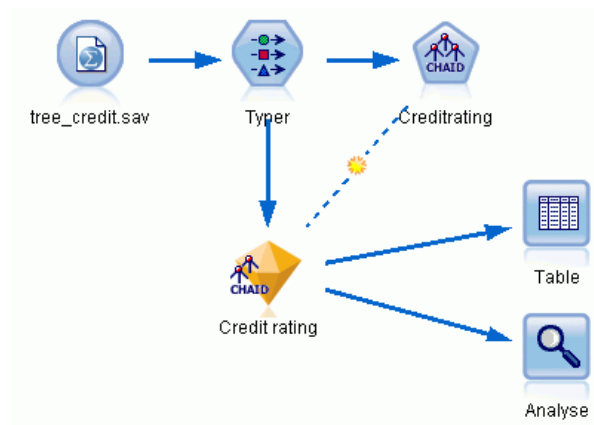
Pour éviter ceci, nous pouvons continuer à diviser l'arbre en branches de plus en plus petites, jusqu'à ce que chaque noeud soit pur à 100%, autrement dit qu'il ne comporte que des *Bonnes* ou des *Mauvaises* sans réponses mixtes. Mais un tel modèle serait extrêmement compliqué et serait probablement difficile à étendre à d'autres ensembles de données.

Pour connaître précisément le nombre de prévisions correctes, nous pouvons lire la table et compter le nombre d'enregistrements où la valeur du champ prédit *\$R-Credit rating* correspond à la valeur des *Conditions de crédit*. Heureusement, il y a beaucoup plus simple : nous pouvons utiliser le noeud Analyse, qui effectue automatiquement cette opération.

- Connectez le nugget de modèle au noeud Analyse.

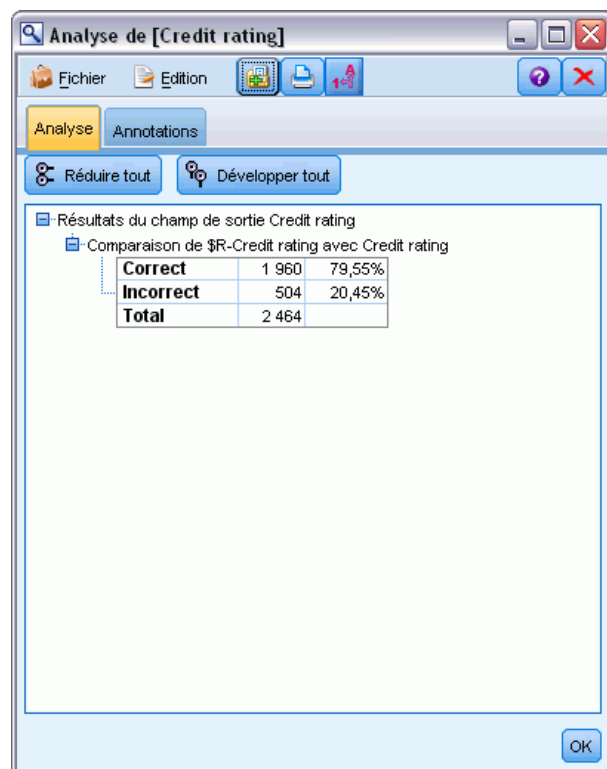
- Double-cliquez sur le noeud Analyse, puis cliquez sur Exécuter.

Figure 2-16
Ajout d'un noeud Analyse



L'analyse montre que pour 1899 enregistrements sur 2464 - un peu plus de 77% - la valeur prédite par le modèle correspondait à la réponse réelle.

Figure 2-17
Résultats d'analyse comparant les réponses observées et les réponses prédites



Ce resultat est limité parce que les enregistrements auxquels un score est donné sont les mêmes que ceux utilisés pour évaluer le modèle. Dans la réalité, vous pourriez utiliser un noeud Partitionner pour diviser les données en échantillons distincts pour l'apprentissage et l'évaluation.

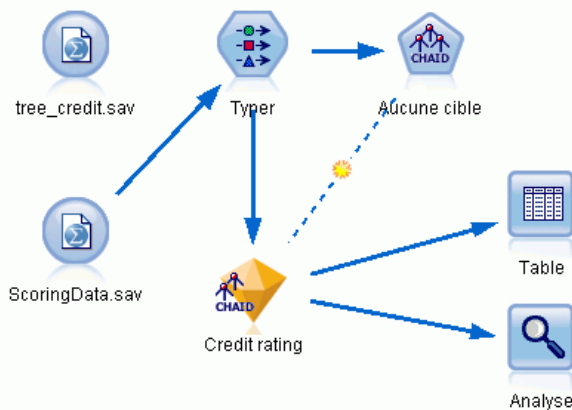
L'utilisation d'un échantillon de partition pour la génération du modèle et d'un autre échantillon pour le tester vous permet d'avoir une bien meilleure indication de la manière dont il peut s'étendre à d'autres ensembles de données.

Le noeud Analyse nous permet de tester le modèle sur les enregistrements pour lesquels nous connaissons déjà le résultat réel. L'étape suivante illustre la façon dont nous pouvons utiliser le modèle pour évaluer les enregistrements dont nous ne connaissons pas le résultat. Par exemple, cela peut comprendre les gens qui ne sont pas des clients de la banque, mais qui sont des cibles potentielles pour un publipostage promotionnel.

Scoring des enregistrements

Auparavant, nous avons évalué les mêmes enregistrements utilisés pour estimer le modèle afin de connaître la précision du modèle. À présent, nous allons voir comment évaluer un ensemble d'enregistrements différent de ceux utilisés pour créer le modèle. Il s'agit de l'objectif de la modélisation avec un champ cible : étudier les enregistrements pour lesquels vous connaissez le résultat pour identifier des schémas qui vous permettront de prédire les résultats que vous ne connaissez pas encore.

Figure 2-18
Association de nouvelles données pour le scoring



Vous pouvez mettre à jour le noeud source Statistics pour qu'il pointe vers un fichier de données différent ou vous pouvez ajouter un nouveau noeud source qui lit dans les données que vous voulez évaluer. Dans les deux méthodes, le nouvel ensemble de données doit contenir les mêmes champs d'entrée utilisés par le modèle (*Age, Niveau de revenu, Education, etc.*) mais pas le champ cible *Conditions de crédit*.

Vous pouvez également ajouter le nugget de modèle à tout flux contenant les champs d'entrée attendus. Qu'il soit lu à partir d'un fichier ou d'une base de données, le type de source n'importe pas du moment que les noms et les types des champs correspondent à ceux utilisés par le modèle.

Vous pouvez également enregistrer le nugget de modèle en tant que fichier distinct, exporter le modèle au format PMML pour une utilisation avec d'autres applications qui prennent en charge ce format ou stocker le modèle dans un répertoire IBM® SPSS® Collaboration and Deployment Services, ce qui permet le déploiement, le scoring et la gestion des modèles à l'échelle de l'entreprise.

Quelque soit l'infrastructure utilisée, le modèle proprement dit fonctionne de la même manière.

Récapitulatif

Cet exemple décrit la procédure standard de création, d'évaluation et de scoring d'un modèle.

- Le noeud de modélisation estime le modèle en étudiant les enregistrements pour lesquels le résultat est connu et crée un nugget de modèle. On parle parfois d'apprentissage du modèle.
- Le nugget de modèle peut être ajouté à n'importe quel flux contenant les champs attendus pour évaluer les enregistrements. En effectuant le scoring des enregistrements pour lesquels vous connaissez déjà le résultat (les clients existants par exemple), vous pouvez évaluer la performance du modèle.
- Une fois que vous êtes satisfait de la performance du modèle, vous pouvez effectuer un scoring de nouvelles données (des clients potentiels par exemple) pour prédire leur réponse.
- Les données utilisées pour l'apprentissage ou l'estimation du modèle peuvent être appelées données analytiques ou historiques; les données de scoring peuvent également être appelées données opérationnelles.

Présentation de la modélisation

Description des noeuds de modélisation

IBM® 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.

Le *Guide des applications SPSS Modeler* fournit des exemples pour de nombreuses méthodes, ainsi qu'une présentation générale du processus de modélisation. Ce guide est disponible en tant que tutoriel en ligne ainsi qu'au format PDF. [Pour plus d'informations, reportez-vous à la section Exemples d'application dans le chapitre 1 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Les méthodes de modélisation sont divisées en trois catégories :

- Classification
- Association
- Segmentation

Modèles de classification

Les *modèles de classification* utilisent les valeurs d'un ou de plusieurs champs d'**entrée** afin de prédire la valeur d'un ou de plusieurs champs de résultat ou **cible**. Certains exemples de ces techniques sont : Les arbres de décision (algorithmes C&R Tree, QUEST, CHAID et C5.0), la régression (algorithmes linéaires, logistiques, linéaires généralisés et régression de Cox), les réseaux de neurones, les machines à vecteurs de prise en charge et les réseaux Bayésiens.

Les modèles de classification permettent aux entreprises de prévoir un résultat connu, par exemple si un client va acheter ou ne pas acheter un produit, ou si une transaction entre dans le cadre d'une fraude connue. Les techniques de modélisation comprennent le processus d'apprentissage automatique, l'induction de règles, l'identification de sous-groupes, les méthodes statistiques et la génération de modèles multiples.

noeuds Classification



Le noeud Classificateur automatique crée et compare les résultats binaires de plusieurs modèles différents (oui ou non, avec ou sans attrition, etc.), ce qui vous permet de choisir la meilleure approche pour une analyse donnée. Plusieurs algorithmes de modélisation sont pris en charge. Vous pouvez alors sélectionner les méthodes que vous souhaitez utiliser, les options spécifiques pour chacune d'elles et le critère de comparaison des résultats. Le noeud génère un ensemble de modèles basé sur les options spécifiées et classe les meilleurs candidats en fonction des critères indiqués. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique dans le chapitre 5 sur p. 99.](#)



Le noeud Numérisation automatique évalue et compare des modèles pour des résultats d'intervalle numérique continus par le biais de différentes méthodes. Le noeud fonctionne de la même manière que le noeud Classificateur automatique, vous permettant ainsi de choisir les algorithmes à utiliser et à tester avec différentes combinaisons d'options en un seul passage de modélisation. Les algorithmes pris en charge comprennent les réseaux neuraux, l'arborescence C&R Tree, CHAID, la régression linéaire, la régression linéaire généralisée et support vector machines (SVM). Les modèles peuvent être comparés selon la corrélation, l'erreur relative ou le nombre de variables utilisées. [Pour plus d'informations, reportez-vous à la section Noeud Numérisation automatique dans le chapitre 5 sur p. 110.](#)



Le noeud Arbre Classification et Regression, (C&R) Tree, génère un arbre décision qui vous permet de prévoir ou de classifier les observations futures. La méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments en réduisant l'index d'impureté à chaque étape, un noeud de l'arbre étant considéré comme "pur" si 100 % de ses observations appartiennent à une catégorie spécifique du champ cible. Les champs cible et les champs d'entrée peuvent être des champs d'intervalle numériques ou des champs catégoriels numériques (nominal, ordinal ou booléen). Toutes les divisions sont binaires (deux sous-groupes uniquement). [Pour plus d'informations, reportez-vous à la section Noeud Arbre C&RT dans le chapitre 6 sur p. 158.](#)



Le noeud QUEST est une méthode de classification supervisée binaire permettant de créer des arbres décision, développée pour réduire le temps de traitement nécessaire aux analyses C&R Tree importantes, tout en limitant la tendance, observée parmi les méthodes d'arbre de classification, à favoriser les entrées autorisant un nombre supérieur de divisions. Les champs d'entrée peuvent être des intervalles numériques (continues) mais les champs cible doivent être catégoriels. Toutes les divisions sont binaires. [Pour plus d'informations, reportez-vous à la section Noeud QUEST dans le chapitre 6 sur p. 160.](#)



Le noeud CHAID génère des arbres décision à l'aide des statistiques du Chi-deux pour identifier les séparations optimales. Contrairement aux noeuds C&R Tree et QUEST, CHAID peut générer des arbres non binaires, ce qui implique que certaines divisions possèdent plusieurs branches. Les champs cibles et les champs d'entrée peuvent être d'intervalle numérique (continu) ou catégoriels. La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en détail toutes les divisions possibles, mais dont les calculs sont plus longs. [Pour plus d'informations, reportez-vous à la section Noeud CHAID dans le chapitre 6 sur p. 159.](#)



Le noeud C5.0 crée un arbre décision ou un ensemble de règles. Le fonctionnement de ce modèle repose sur un découpage de l'échantillon basé sur le champ qui fournit le gain d'informations le plus important à chaque niveau. Le champ cible doit être catégoriel. Les divisions multiples en plus de deux sous-groupes sont autorisées. [Pour plus d'informations, reportez-vous à la section Noeud C5.0 dans le chapitre 6 sur p. 180.](#)



Le noeud Liste de décision identifie les sous-groupes, ou les segments, qui présentent une probabilité plus élevée ou plus faible d'un résultat binaire donné par rapport à la population globale. Vous pouvez, par exemple, rechercher les clients qui ont une faible probabilité d'attrition ou ceux qui ont une plus forte probabilité de répondre favorablement à une campagne. Vous pouvez incorporer votre connaissance du marché dans le modèle en ajoutant vos propres segments personnalisés et en prévisualisant des modèles alternatifs côte à côte de façon à comparer les résultats. Les modèles Liste de décision se composent d'une liste de règles dans laquelle chaque règle présente une condition et un résultat. Les règles sont appliquées dans l'ordre et la première règle correspondante détermine le résultat. [Pour plus d'informations, reportez-vous à la section Liste de décision dans le chapitre 9 sur p. 229.](#)



Les modèles de régression linéaire prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites. [Pour plus d'informations, reportez-vous à la section Modèles linéaires dans le chapitre 10 sur p. 269.](#)



Le noeud ACP/Analyse factorielle propose des techniques de factorisation puissantes qui vous permettent de réduire la complexité de vos données. L'analyse en composantes principales (ACP) recherche les combinaisons linéaires des champs d'entrée qui permettent de capturer au mieux la variance dans l'ensemble de champs, où les composantes sont orthogonales (perpendiculaires) les unes par rapport aux autres. L'analyse factorielle a pour but d'identifier les facteurs sous-jacents qui expliquent la tendance des corrélations dans un ensemble de champs observés. Quelle que soit l'approche choisie, le but consiste à trouver un nombre limité de champs dérivés récapitulant les informations contenues dans l'ensemble de champs d'origine. [Pour plus d'informations, reportez-vous à la section Noeud ACP/Analyse factorielle dans le chapitre 10 sur p. 310.](#)



Le noeud Sélection de fonction filtre les champs d'entrée en vue de leur suppression, en fonction d'un ensemble de critères donné (tel que le pourcentage de valeurs manquantes) ; il classe ensuite les entrées restantes selon leur importance par rapport à la cible indiquée. Si l'on prend, par exemple, un ensemble de données comportant des centaines d'entrées potentielles, quelles sont celles susceptibles d'être les plus utiles dans la modélisation des résultats de patients ? [Pour plus d'informations, reportez-vous à la section Noeud Sélection de fonction dans le chapitre 4 sur p. 78.](#)



L'analyse discriminante crée des hypothèses plus strictes que la régression logistique mais peut constituer une alternative ou un complément précieux à une analyse de régression logistique lorsque ces hypothèses sont réunies. [Pour plus d'informations, reportez-vous à la section Noeud discriminant dans le chapitre 10 sur p. 318.](#)



La régression logistique est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est similaire à la régression linéaire. [Pour plus d'informations, reportez-vous à la section Noeud Logistique dans le chapitre 10 sur p. 288.](#)



La procédure Modèles linéaires généralisés développe le modèle linéaire général de sorte que la variable dépendante soit linéairement reliée aux facteurs et covariables via une fonction de lien précise. En outre, le modèle permet à la variable dépendante de suivre une distribution non normale. Il couvre les fonctionnalités d'un grand nombre de modèles statistiques, notamment le modèle de régression linéaire, le modèle de régression logistique, le modèle log-linéaire pour les données d'effectif et le modèle de survie avec censure par intervalle. [Pour plus d'informations, reportez-vous à la section Noeud Modèles linéaires généralisés dans le chapitre 10 sur p. 327.](#)



Un modèle mixte linéaire généralisé (MMLG) élargit le modèle linéaire de sorte que la cible puisse avoir une distribution non normale, qu'elle soit liée linéairement aux facteurs et covariables via une fonction de lien spécifiée, et que les observations puissent être corrélées. Les modèles mixtes linéaires généralisés couvrent une large variété de modèles, depuis les modèles de régression linéaire simple aux modèles multi-niveaux complexes destinés aux données longitudinales non normales. [Pour plus d'informations, reportez-vous à la section Noeud MMLG dans le chapitre 10 sur p. 342.](#)



Le noeud de régression de Cox vous permet de créer un modèle de survie pour les données de durée jusqu'à l'événement en présence d'enregistrements censurés. Ce modèle produit une fonction de survie qui prédit la probabilité que l'événement en question se soit produit à un moment (t) pour des valeurs données des variables d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud de Cox dans le chapitre 10 sur p. 371.](#)



Le noeud Support Vector Machine (SVM) vous permet de classer les données dans l'un de deux groupes sans sur-ajustement. SVM fonctionne bien avec les grands ensembles de données, comme ceux qui disposent d'un très grand nombre de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud SVM dans le chapitre 15 sur p. 501.](#)



Le noeud Réseau Bayésien permet de créer un modèle de probabilité en combinant les preuves observées et enregistrées avec les connaissances réelles pour établir la probabilité des occurrences. Le noeud est axé sur le Tree Augmented Naïve Bayes (TAN) et sur les réseaux Couverture de Markov qui servent principalement à la classification. [Pour plus d'informations, reportez-vous à la section Noeud Réseau Bayésien dans le chapitre 7 sur p. 203.](#)



Le noeud Modèle de réponse en auto-apprentissage (SLRM) vous permet de créer un modèle dans lequel une nouvelle observation unique, ou un petit nombre de nouvelles observations, peuvent être utilisés pour réestimer un modèle sans qu'un recyclage de toutes les données soit nécessaire. [Pour plus d'informations, reportez-vous à la section Noeud MRAA dans le chapitre 14 sur p. 488.](#)



Le noeud Séries temporelles estime les modèles de lissage exponentiel, d'ARIMA (Autoregressive Integrated Moving Average) univariable et d'ARIMA multivariable (ou fonction de transfert) pour les données de séries temporelles et génère des prévisions d'une performance future. Un noeud Séries temporelles doit toujours être précédé d'un noeud Intervalles de temps. [Pour plus d'informations, reportez-vous à la section Noeud de modélisation Séries temporelles dans le chapitre 13 sur p. 465.](#)



Le noeud k -Voisin le plus proche (KNN) associe une nouvelle observation à la catégorie ou à la valeur des objets k les plus proches dans l'espace du prédicteur, où k est un entier. Les observations semblables sont proches l'une de l'autre et les observations dissemblables sont éloignées l'une de l'autre. [Pour plus d'informations, reportez-vous à la section Noeud KNN dans le chapitre 16 sur p. 507.](#)

Modèles d'association

Les *modèles d'association* recherchent des modèles dans vos données où une ou plusieurs entités (telles que des événements, des achats ou des attributs) sont associées à une ou plusieurs autres entités. Les modèles établissent des ensembles de règles qui définissent ces relations. Ici, les champs au sein des données peuvent se comporter à la fois comme des entrées et comme des cibles. Vous ne pouvez pas découvrir ces associations manuellement, mais des algorithmes de règles font bien plus et plus rapidement, de sorte que vous êtes en mesure d'explorer des modèles plus complexes. Les modèles Apriori et Carma sont des exemples de l'utilisation de tels algorithmes. Un autre type de modèle d'association est un modèle de détection de séquences qui recherche des motifs séquentiels dans des données à structure temporelle.

Les modèles d'association sont particulièrement utiles pour prévoir des résultats multiples ; par exemple, les clients qui ont acheté le produit X ont également acheté les produits Y et Z. Les modèles d'association associent une conclusion particulière (telle que la décision d'acheter un produit) à un ensemble de conditions. L'avantage des algorithmes de règles d'association par rapport aux algorithmes d'arbre décision standard (C5.0 et Arbre C&RT) est le fait qu'il puisse exister des associations entre tous les attributs. Un algorithme d'arbre décision peut construire une règle uniquement avec une seule conclusion. En revanche, les algorithmes d'association tentent d'en trouver plusieurs, chaque règle pouvant avoir une conclusion différente.

Noeuds Association



Le noeud Apriori extrait des données un ensemble de règles et retient les règles contenant la plus grande quantité d'informations. Le noeud Apriori fournit cinq méthodes de sélection de règles et utilise un modèle d'indexation sophistiqué pour traiter efficacement les volumes de données importants. Pour les problèmes importants, l'apprentissage du noeud Apriori est généralement plus rapide ; il n'existe aucune limite quant au nombre de règles pouvant être conservées et il peut prendre en charge des règles faisant l'objet de 32 pré-conditions. Le noeud Apriori exige que les champs d'entrée et de sortie soient tous catégoriels, mais fournit de meilleures performances car il est optimisé de ce type de données. [Pour plus d'informations, reportez-vous à la section Noeud Apriori dans le chapitre 12 sur p. 417.](#)



Le modèle CARMA extrait un ensemble de règles des données sans que vous ayez à définir les champs d'entrée ou les champs cible. Au contraire du noeud Apriori le noeud CARMA offre des paramètres de création pour la prise en charge de la règle (à la fois pour les antécédents et les conséquences), plutôt qu'une simple prise en charge d'antécédents. Cela signifie que les règles générées peuvent être utilisées dans un grand nombre d'applications, par exemple pour rechercher une liste des produits ou des services (antécédents) dont la conséquence correspond à l'élément que vous souhaitez promouvoir à l'occasion de cette période de congés. [Pour plus d'informations, reportez-vous à la section Noeud CARMA dans le chapitre 12 sur p. 421.](#)



Le noeud Séquence recherche des règles d'association dans des données dotées d'une dimension temporelle. Une séquence est une liste de jeux d'éléments ayant tendance à survenir dans un ordre prévisible. Par exemple, un client qui achète un rasoir et une lotion après-rasage achètera vraisemblablement de la crème à raser. Le noeud Séquence est basé sur l'algorithme de règles d'association CARMA, qui utilise une méthode efficace de double lecture pour rechercher des séquences. [Pour plus d'informations, reportez-vous à la section Noeud Séquence dans le chapitre 12 sur p. 443.](#)

Modèles de segmentation

Les *modèles de segmentation* divisent les données en segments, ou classes, d'enregistrement ayant des profils similaires de champs d'entrée. Comme ils ne s'occupent que des champs d'entrée, les modèles de segmentation n'ont aucun concept de champs de sortie ou cible. Parmi les exemples de modèles de segmentation, on trouve les réseaux Kohonen, la classification K-Means, la classification en deux étapes et la détection d'anomalies.

Les modèles de Segmentation (appelés aussi « modèles de classification ») sont utiles dans les cas où le résultat précis est inconnu (par exemple, lorsque vous identifiez de nouveaux types de fraude ou lorsque vous identifiez des groupes d'intérêt dans votre clientèle). Les modèles de classification se chargent essentiellement d'identifier des groupes d'enregistrements similaires et de répertorier les enregistrements en fonction du groupe auquel ils appartiennent. Cette opération peut s'effectuer sans la connaissance préalable des groupes et de leurs caractéristiques, et distingue les modèles de classification non supervisée des autres techniques de modélisation par le fait qu'aucun champ de sortie ni champ cible n'est prédéfini pour le modèle à prévoir. Il n'y a pas de réponse vraie ou fausse pour ces modèles. Leur valeur est déterminée par leur capacité à capturer des groupements intéressants dans les données et ils fournissent des descriptions utiles de ces mêmes groupements. Les modèles de classification non supervisée sont souvent utilisés pour créer des classes ou des segments qui sont ensuite utilisés en tant qu'entrées dans les analyses suivantes (par exemple, par la segmentation de clients potentiels dans des sous-groupes homogènes).

Noeuds Segmentation



Le noeud Classification automatique évalue et compare les modèles de classification identifiant des groupes d'enregistrements ayant des caractéristiques similaires. Le noeud fonctionne de la même manière que les autres noeuds de modélisation automatiques, vous permettant de tester plusieurs combinaisons d'options en une seule modélisation. Les modèles peuvent être comparés à l'aide de mesures de bases permettant d'essayer de filtrer et de classer l'utilité des modèles de classification et de fournir une mesure en fonction de l'importance de champs particuliers. [Pour plus d'informations, reportez-vous à la section Noeud Classification automatique dans le chapitre 5 sur p. 117.](#)



Le noeud K-Means classe l'ensemble de données dans différents groupes (ou classes). La méthode définit un nombre de classes fixe, affecte à plusieurs reprises des enregistrements à des classes et ajuste les centres de classe, jusqu'à ce que le modèle ne puisse plus être amélioré. Au lieu de tenter de prédire un résultat, le modèle *k*-means utilise un processus connu sous le nom d'apprentissage non supervisé pour découvrir des tendances dans l'ensemble de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud K-Means dans le chapitre 11 sur p. 390.](#)



Le noeud Kohonen génère un type de réseau de neurones qui peut être utilisé pour classer les données en groupes distincts. Lorsque l'apprentissage du réseau est terminé, les enregistrements similaires doivent être regroupés dans la connexion de sortie, tandis que les enregistrements différents sont à l'opposé. Vous pouvez étudier le nombre d'observations capturées par chaque unité du nœud de modèle afin d'identifier les unités fortes. Vous pouvez ainsi vous faire une idée du nombre de classes approprié. [Pour plus d'informations, reportez-vous à la section Noeud Kohonen dans le chapitre 11 sur p. 383.](#)



Le noeud TwoStep utilise une méthode de classification non supervisée en deux étapes. La première étape consiste en une exploration des données visant à compresser les données d'entrée brutes en sous-classes plus faciles à manipuler. Au cours de la seconde étape, l'utilisation d'une méthode de classification hiérarchique permet de fusionner progressivement les sous-classes en classes de plus en plus importantes. La technique TwoStep a l'avantage d'évaluer automatiquement le nombre de classes optimal pour les données d'apprentissage. Il peut prendre en charge de manière efficace des types de champ mixtes et des ensembles de données volumineux. [Pour plus d'informations, reportez-vous à la section Noeud Classification TwoStep dans le chapitre 11 sur p. 395.](#)



Le noeud Détection des anomalies identifie les observations inhabituelles, ou valeurs éloignées, qui ne se conforment pas aux patrons de données « normales ». Il vous permet d'identifier les valeurs éloignées même si celles-ci ne correspondent pas aux patrons connus précédemment et même si vous ne savez pas exactement ce que vous recherchez. [Pour plus d'informations, reportez-vous à la section Noeud Détection des anomalies dans le chapitre 4 sur p. 86.](#)

Exploration des bases de données des modèles

SPSS Modeler prend en charge l'intégration des outils de Data mining et de modélisation disponibles auprès des fournisseurs de base de données, notamment Oracle Data Miner, IBM DB2 InfoSphere Warehouse 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 SPSS Modeler. Pour plus d'informations, consultez le *Guide d'exploration de base de données SPSS Modeler*, disponible sur le DVD du produit.

Modèles IBM SPSS Statistics

Si vous avez installé une copie sous licence de IBM® SPSS® Statistics sur votre ordinateur, vous pouvez accéder et exécuter certains programmes SPSS Statistics depuis SPSS Modeler pour créer et évaluer des modèles. [Pour plus d'informations, reportez-vous à la section Noeuds IBM SPSS Statistics - Présentation dans le chapitre 8 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

Informations supplémentaires

Une documentation détaillée sur les algorithmes de modélisation est également disponible. Pour plus d'informations, consultez le *Guides des algorithmes de SPSS Modeler*, disponible sur le DVD du produit.

Création de modèles découpés

La modélisation découpée vous permet d'utiliser un flux unique pour créer des modèles séparés pour chaque valeur possible d'un champ d'entrée nominal, continu ou booléen, les modèles obtenus étant tous accessibles à partir d'un nugget de modèle unique. Les valeurs possibles de ces champs d'entrée peuvent avoir des effets très différents sur le modèle. Avec la modélisation découpée, vous pouvez créer facilement le modèle le mieux adapté pour chaque valeur de champ possible dans une seule exécution du flux.

Notez que les sessions de modélisation interactives ne peuvent pas utiliser le découpage. Avec la modélisation interactive, vous spécifiez chaque modèle individuellement. Par conséquent le découpage, qui consiste à créer plusieurs modèles automatiquement, ne présente pas d'intérêt.

La modélisation découpée repose sur la désignation d'un champ d'entrée particulier comme champ de découpage. Pour ce faire, vous pouvez définir le rôle du champ sur Découpage dans la spécification du Type :

Figure 3-1

Désignation d'un champ d'entrée comme champ de découpage

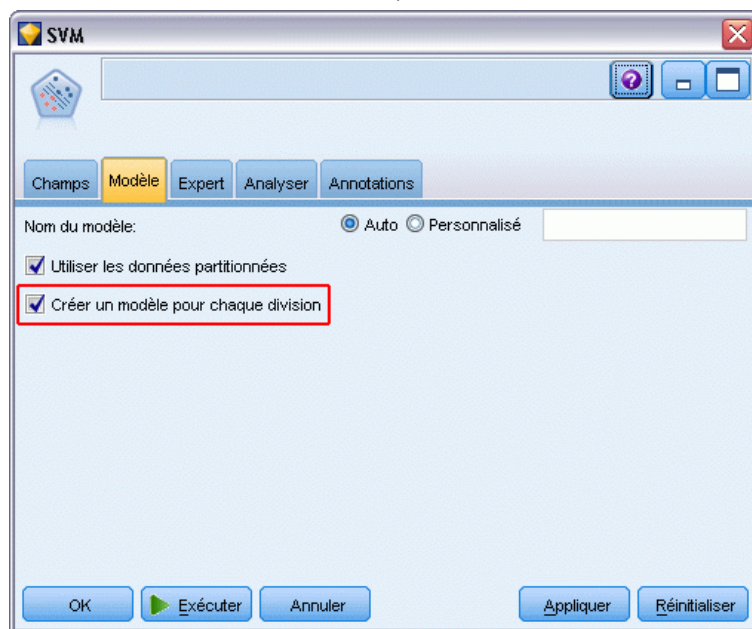


Vous ne pouvez désigner que des champs ayant un niveau de mesure Booléen, Nominal, Ordinal ou Continu comme champs de découpage.

Vous pouvez affecter plusieurs champs d'entrée comme champ de découpage. Dans ce cas, toutefois, le nombre de modèles créés peut être considérablement augmenté. Un modèle est créé pour chaque combinaison de valeurs possible pour le champ de découpage sélectionné. Par exemple, si trois champs d'entrée, comportant chacun trois valeurs possibles, sont désignés comme champs de découpage, cela entraînera la création de 27 modèles différents.

Même après avoir affecté un ou plusieurs champs comme champs de découpage, vous pouvez toujours choisir de créer des modèles découpés ou un modèle unique, en sélectionnant ou non la case de la boîte de dialogue du noeud de modélisation :

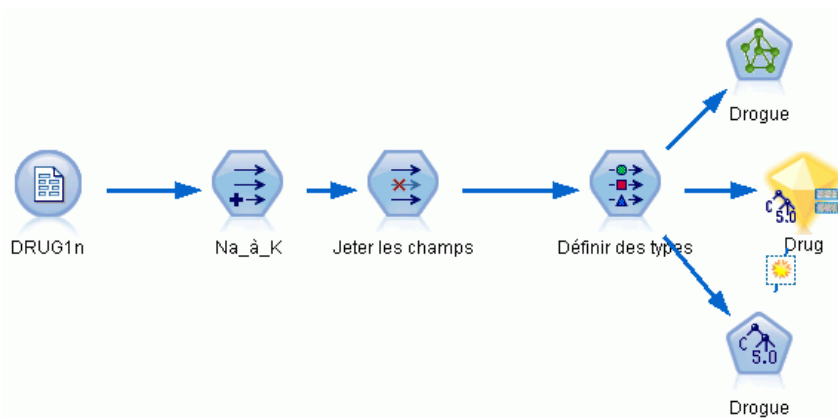
Figure 3-2
Choisir de créer des modèles découpés



Si des champs de découpage sont définis mais que la case n'est pas cochée, seul un modèle unique est généré. De la même manière, si la case est cochée mais qu'aucun champ de découpage n'est défini, le découpage est ignoré et un modèle unique est généré.

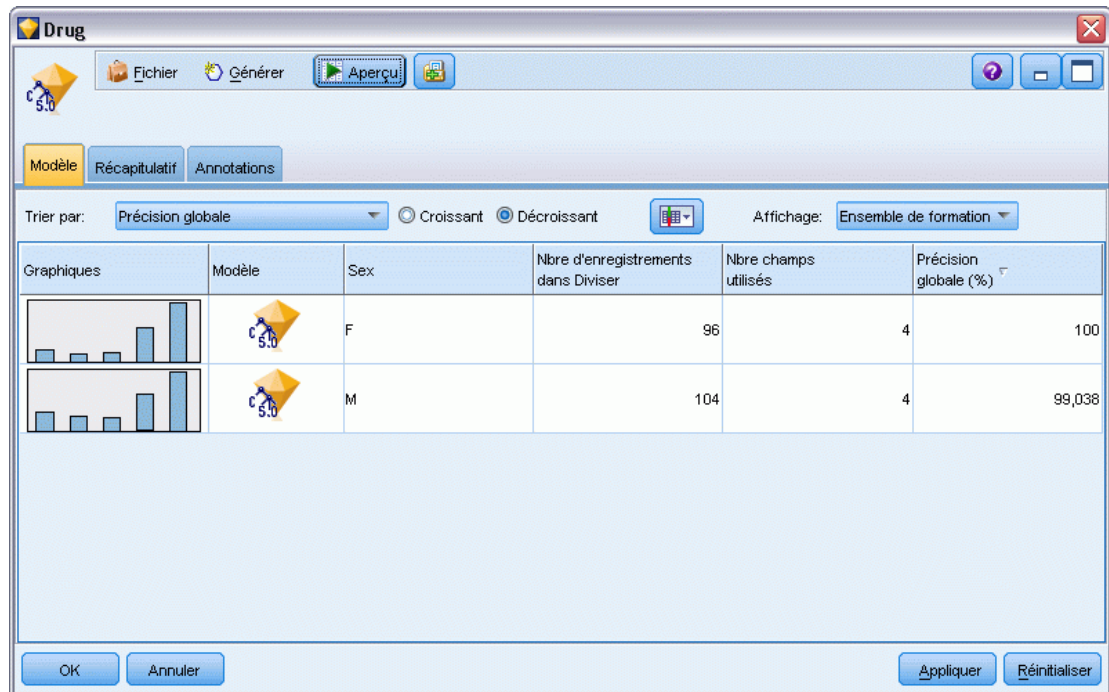
Lorsque vous exécutez le flux, des modèles séparés sont créés en arrière-plan pour chaque valeur possible du ou des champs de découpage, mais seul un nugget de modèle unique est placé dans la palette de modèles et dans l'espace du flux de travail. Un nugget de modèle découpé est signalé par le symbole de découpage :

Figure 3-3
Nugget de modèle découpé dans un flux



Lorsque vous accédez au nugget de modèle découpé, une liste de tous les modèles séparés qui ont été créés s'affiche :

Figure 3-4
Afficheur de modèle découpé



Vous pouvez examiner un modèle individuel d'une liste en double-cliquant sur l'icône de son nugget dans l'afficheur. Ceci ouvre une fenêtre de navigateur standard pour le modèle individuel. Lorsque le nugget est sur l'espace de travail, double-cliquer sur un graphique en miniature ouvre le graphique grandeur nature. [Pour plus d'informations, reportez-vous à la section Afficheur de modèle découpé sur p. 69.](#)

Une fois qu'un modèle a été créé en tant que modèle découpé, vous ne pouvez pas supprimer le processus de découpage inhérent à ce modèle, ni annuler le découpage en aval d'un noeud ou d'un nugget de modélisation découpé.

Exemple : Un détaillant national souhaite estimer ses ventes par catégorie de produit dans chacun de ses magasins à l'échelle du pays. Au moyen de la modélisation découpée, il désigne le champ Magasin de ses données d'entrée comme champ de découpage, ce qui lui permet de créer des modèles séparés pour chaque catégorie de chaque magasin dans une seule opération. Il peut ensuite utiliser les informations obtenues pour contrôler les niveaux de stock de manière beaucoup plus précise qu'il ne le pourrait avec un modèle unique.

Découpage et partitionnement

Le découpage possède quelques fonctions en commun avec le partitionnement, mais les deux sont utilisés de manières très différentes.

Le **Partitionnement** divise l'ensemble de données au hasard en deux ou trois parties : formation, test et (en option) validation, et il sert à tester la performance d'un modèle unique.

Le **Découpage** divise l'ensemble de données en autant de parties qu'il existe de valeurs possibles pour un champ de découpage, et sert à créer des modèles multiples.

Le partitionnement et le découpage fonctionnent de manière complètement indépendante l'un de l'autre. Dans un modèle de modélisation vous pouvez choisir l'un ou l'autre, les deux, ou aucun des deux.

Noeuds de modélisation prenant en charge les modèles découpés

Plusieurs noeuds de modélisation peuvent créer des modèles découpés. Les exceptions sont les noeuds Classification automatique, Séries temporelles, Analyse factorielle/ACP, Sélection de fonction, MRAA, les modèles d'association (Apriori, Carma et Séquence), les modèles de classification (K-Means, Kohonen, Two Step et Anomaly), le modèle Statistics et les noeuds utilisés pour la modélisation de base de données.

Les noeuds de modélisation qui prennent en charge la modélisation découpée sont les suivants :



Arbre C&RT



Bayes Net



QUEST



Modèles linéaires généralisés



CHAID



KNN



C5.0



Cox



R. neurones



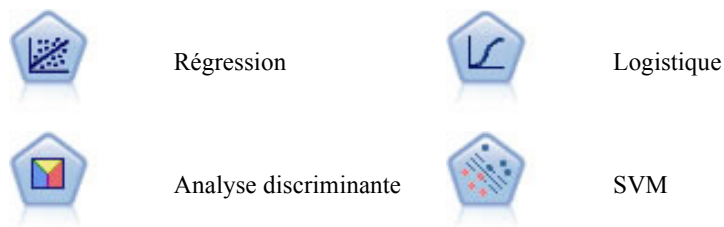
Classificateur automatique



Liste de décision



Numérisation automatique



Fonctions affectées par le découpage

L'utilisation de modèles découpés affecte plusieurs fonctions de IBM® SPSS® Modeler de différentes manières. Cette section fournit des directives sur l'utilisation de modèles découpés en association avec d'autres noeuds dans un flux.

Noeuds Ops sur lignes

Lors de l'utilisation de modèles découpés dans un flux contenant un noeud **Echantillon**, stratifiez les enregistrements par le champ de découpage pour obtenir un échantillonnage apparié d'enregistrements. Cette option est disponible lorsque vous choisissez la méthode d'échantillonnage Complexe. [Pour plus d'informations, reportez-vous à la section Paramètres de classification et de stratification dans le chapitre 3 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Si le flux contient un noeud **Equilibrer**, notez que l'équilibrage s'applique à l'ensemble global d'enregistrements d'entrée, et non au sous-ensemble d'enregistrements dans un découpage.

Lors de l'agrégation d'enregistrements au moyen du noeud **Agréger**, définissez les champs de découpage comme les champs clés si vous souhaitez calculer des agrégats pour chaque découpage. [Pour plus d'informations, reportez-vous à la section Noeud Agréger dans le chapitre 3 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Noeuds Ops sur champs

Le noeud **Typer** vous permet de spécifier le ou les champs à utiliser comme champs de découpage. [Pour plus d'informations, reportez-vous à la section Noeud Typer dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Remarque : si le noeud **Ensemble** est utilisé pour combiner deux nuggets de modèle ou plus, il ne peut pas être utilisé pour annuler l'action de découpage, car les modèles découpés sont contenus dans un nugget de modèle unique.

noeuds de modélisation

Les modèles découpés ne prennent pas en charge le calcul de l'importance de la variable indépendante (l'importance relative des champs d'entrées de la valeur indépendante dans l'estimation du modèle). Les paramètres de l'importance de la variable indépendante sont ignorés lors de la création de modèles découpés.

Le noeud **KNN** (agrégation suivant le saut minimum) prend en charge les modèles découpés uniquement s'il est défini pour prédire un champ cible. L'autre paramètre (uniquement identifier les agrégations suivant le saut minimum) ne crée pas de modèle. Si l'option "Sélectionner k automatiquement" est choisie, chacun des modèles découpés peut avoir un nombre différent d'agrégations suivant le saut minimum. Le modèle global aura ainsi un nombre de colonnes générées égal au plus grand nombre d'agrégations suivant le saut minimum trouvées dans l'ensemble des modèles découpés. Pour les modèles découpés où le nombre d'agrégations suivant le saut minimum est inférieur à ce maximum, il existera un nombre correspondant de colonnes remplies avec les valeurs \$null. [Pour plus d'informations, reportez-vous à la section Noeud KNN dans le chapitre 16 sur p. 507.](#)

noeuds de modélisation de base de données

Les noeuds de modélisation à l'intérieur des bases de données ne prennent pas en charge les modèles découpés.

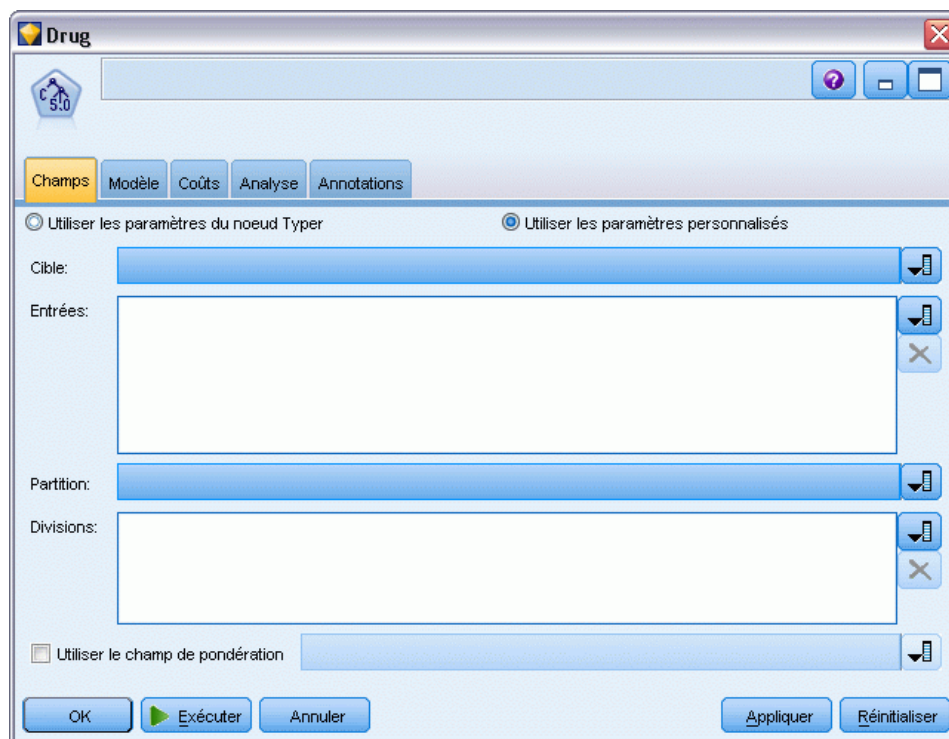
Nuggets de modèle

L'**Exportation au format PMML** à partir d'un nugget de modèle découpé n'est pas possible, car le nugget contient plusieurs modèles et PMML ne prend pas en charge un tel regroupement. Toutefois, l'exportation au format texte ou HTML est possible.

Options de champs des noeuds de modélisation

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

Figure 3-5
Exemple d'un onglet Champs de noeuds de modélisation



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. Si vous utilisez un nœud Typer pour sélectionner les champs d'entrée et les champs cible, vous n'avez aucune modification à apporter dans cet onglet. (Parmi ces exceptions figurent les noeuds Séquence et Extraction de texte, pour lesquels des paramètres de champ doivent être spécifiés dans le nœud de modélisation.)

Utiliser les paramètres du noeud Typer. Cette option indique au nœud d'utiliser les informations du champ à partir d'un nœud Typer 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 Typer en amont. Une fois cette option sélectionnée, renseignez les champs ci-dessous.

Remarque : Tous les champs ne sont pas affichés pour tous les noeuds.

- **Utilisez le format transactionnel (noeuds Apriori, CARMA, Règles d'association MS, et Oracle Apriori uniquement).** 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 sur p. 416.](#)

- **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).
- **Les ID sont contigus.** (Pour les noeuds Apriori et CARMA seulement) Si vos données sont prétriées de façon à ce que tous les enregistrements avec le même ID soient regroupés dans le flux de données, sélectionnez cette option pour accélérer le traitement. Si vos données ne sont pas pré-triées (ou si vous n'en êtes pas certain), ne sélectionnez pas cette option ; le noeud triera automatiquement les données.

Remarque : Si vos données ne sont pas triées et si vous sélectionnez cette option, vous risquez d'obtenir des résultats incorrects dans votre modèle.

- **Contenu.** Permet d'ajouter des informations aux champs d'analyse du modèle. Ces champs contiennent les éléments d'intérêt concernant la modélisation des associations. Vous pouvez définir plusieurs champs booléens (si les données sont au format tabulaire) ou un champ nominal unique (si les données sont au format transactionnel).
- **Cible.** Pour les modèles qui nécessitent un ou plusieurs champs cible, sélectionnez le ou les champs cible. Cela revient à définir le rôle du champ sur la valeur *Cible* dans un noeud Typer.
- **Evaluation.** (Uniquement pour des modèles de classification automatique.) Aucune cible n'est spécifiée pour des modèles de classe. Cependant, vous pouvez sélectionner un champ d'évaluation pour identifier son niveau d'importance. En outre, vous pouvez évaluer la façon dont les classes différencient les valeurs de ce champ, ce qui à son tour indique si les classes peuvent être utilisées pour prédire ce champ.
- **Entrées.** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un noeud 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 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.)
- **Découpages.** Pour des modèles découpés, sélectionnez le ou les champs de découpage. Cela revient à définir le rôle du champ sur la valeur *Division* dans un noeud Typer. Vous ne pouvez désigner que des champs ayant un niveau de mesure Booléen, Nominal, Ordinal ou Continu comme champs de découpage. Les champs sélectionnés en tant que le champ de découpage

ne peuvent pas être utilisés comme champs de cible, d'entrée, de partition, de fréquence ou de pondération. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés sur p. 33.](#)

- **Utiliser un champ d'effectif.** Cette option vous permet de sélectionner un champ qui servira de pondérateur d'effectif. Utilisez cette option lorsque les enregistrements de vos données d'apprentissage représentent chacun plusieurs unités (c'est le cas si vos données sont agrégées). Les valeurs des champs doivent être égales au nombre d'unités représentées par chaque enregistrement. [Pour plus d'informations, reportez-vous à la section Utilisation des champs d'effectif et de pondération sur p. 42.](#)

Remarque : Si vous voyez le message d'erreur Métadonnées non valides (dans les champs entrée/sortie), vérifiez que vous avez spécifié tous les champs requis, comme par exemple le champ d'effectif.

- **Utiliser un champ de pondération.** Cette option vous permet de sélectionner un champ qui servira de pondérateur d'observation. Les pondérations d'observation sont utilisées pour représenter les différences de variance dans les niveaux du champ de résultat. [Pour plus d'informations, reportez-vous à la section Utilisation des champs d'effectif et de pondération sur p. 42.](#)
- **Conséquences.** Pour les noeuds d'induction de règles (Apriori), sélectionnez les champs à utiliser en tant que conséquences dans l'ensemble de règles généré, (il s'agit des champs dont le rôle est *Cible* ou *Les deux* dans un nœud Typer).
- **Antécédents.** Pour les noeuds d'induction de règles (Apriori), sélectionnez les champs à utiliser en tant que conséquences dans l'ensemble de règles généré, (il s'agit des champs dont le rôle est *Entrée* ou *Les deux* dans un nœud Typer).

Certains modèles disposent d'un onglet Champs différent de ceux décrits dans cette section.

- [Pour plus d'informations, reportez-vous à la section Noeud Séquence - Options de champs dans le chapitre 12 sur p. 444.](#)
- [Pour plus d'informations, reportez-vous à la section Noeud CARMA - Options des champs dans le chapitre 12 sur p. 422.](#)

Utilisation des champs d'effectif et de pondération

Les champs d'effectif et de pondération servent à donner une importance supplémentaire à certains enregistrements, par exemple, lorsque vous savez qu'une section de la population est sous représentée dans les données d'apprentissage (pondération) ou parce qu'un enregistrement représente plusieurs cas identiques (effectif).

- Les valeurs d'un champ d'effectif doivent être des entiers positifs. Les enregistrements dont la pondération d'effectif est nulle ou négative sont exclus du processus d'analyse. Si besoin est, les pondérations d'effectif sont arrondies de façon à atteindre des valeurs entières.
- Les pondérations d'observation doivent prendre des valeurs positives. Les enregistrements présentant une pondération d'observation nulle ou négative sont exclus du processus d'analyse.

Champs d'effectif et de pondération de scoring

Les champs d'effectif et de pondération sont utilisés dans les modèles d'apprentissage mais ne servent pas au scoring, car le score de chaque enregistrement est basé sur ses caractéristiques quel que soit le nombre de cas qu'il représente. Supposons, par exemple, que vous ayez les données suivantes :

Marié	Réponse
Oui	Oui
Oui	Oui
Oui	Oui
Oui	Non
Non	Oui
Non	Non
Non	Non

En fonction de cela, vous concluez que trois personnes mariées sur quatre répondent à la promotion, et que deux personnes non mariées sur trois n'ont pas répondu. Vous allez donc déterminer le score des nouveaux enregistrements en conséquence :

Marié	\$-Réponse	\$RP-Réponse
Oui	Oui	0,75 (trois/quatre)
Non	Non	0,67 (deux/trois)

Vous pouvez également stocker vos données d'apprentissage de façon plus compacte en utilisant un champ d'effectif :

Marié	Réponse	Effectif
Oui	Oui	3
Oui	Non	1
Non	Oui	1
Non	Non	2

Comme cela représente exactement le même ensemble de données, vous allez construire le même modèle et prédire des réponses uniquement en fonction du statut marital. Si vous avez dix personnes mariées dans vos données de scoring, vous prédirez *Oui* pour chacune, qu'elles soient présentées comme dix enregistrements séparés ou comme un enregistrement avec une valeur d'effectif de 10. La pondération, bien qu'elle ne soit généralement pas un entier, peut être considérée comme indiquant de manière similaire l'importance d'un enregistrement. C'est pourquoi les champs d'effectif et de pondération ne sont pas utilisés lors du scoring d'enregistrements.

Evaluation et comparaison de modèles

Certains types de modèles prennent en charge les champs d'effectif, certains prennent en charge des champs de pondération et certains les deux. Mais dans tous les cas où ils s'appliquent, ils servent uniquement à la construction de modèles et ne sont pas considérés comme des modèles

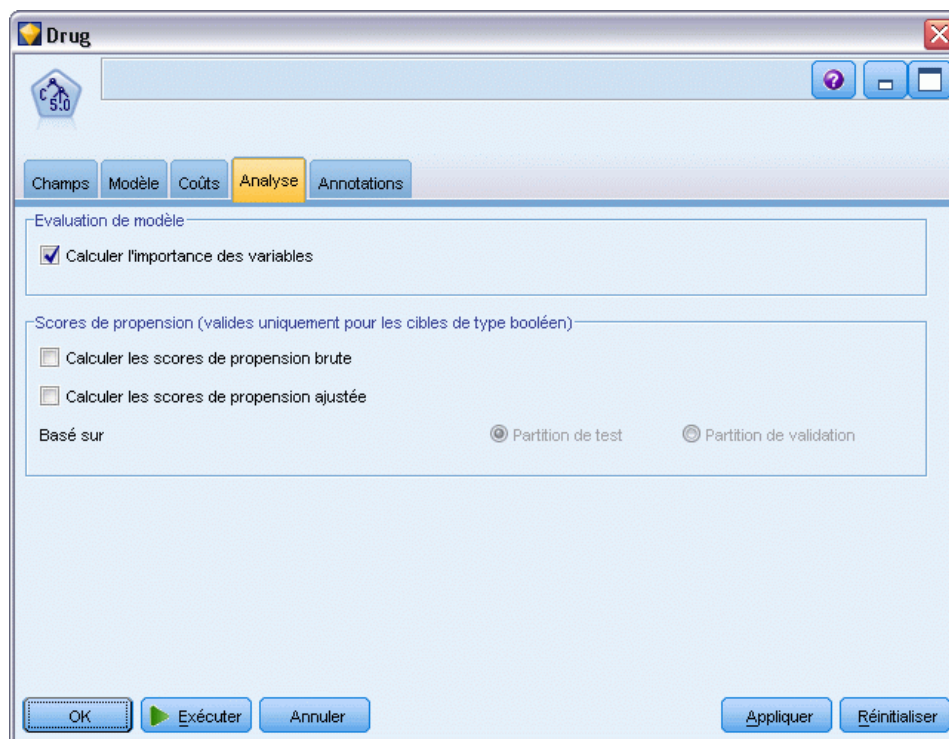
d'évaluation utilisant un noeud Evaluation ou un noeud Analyse, ou lors du classement de modèles utilisant la plupart des méthodes prises en charge par les noeuds Classificateur automatique et Numérisation automatique.

- Lors de la comparaison de modèles (utilisant des graphiques Evaluation, par exemple) les valeurs d'effectif et de pondération sont ignorées. Cela permet une comparaison de niveau entre des modèles qui utilisent ces champs et des modèles qui ne les utilisent pas, mais cela signifie que pour une évaluation précise, un ensemble de données qui représente précisément la population sans compter sur un champ d'effectif ou de pondération doit être utilisé. En termes pratiques, vous pouvez le faire en vous assurant que des modèles sont évalués en utilisant un échantillon de test dans lequel la valeur du champ d'effectif ou de pondération est toujours nulle ou vaut 1 (Cette restriction ne s'applique que lors de l'évaluation de modèles ; si les valeurs d'effectif ou de pondération valent toujours 1 à la fois pour les échantillons d'apprentissage et de test, il n'y a aucune raison d'utiliser ces champs en premier lieu.)
- Si vous utilisez un Classificateur automatique, l'effectif peut être pris en compte si des modèles de classement sont basés sur le Profit, de sorte que cette méthode est recommandée dans ce cas.
- Si nécessaire, vous pouvez diviser les données en échantillons d'apprentissage et de test en utilisant un noeud Partitionner. [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*.](#)

Options d'analyse des noeuds de modélisation

De nombreux noeuds de modélisation comprennent un onglet Analyser qui vous permet d'obtenir des informations concernant l'importance de la variable indépendante ainsi que des scores de propension brute et ajustée.

Figure 3-6
Onglet Analyser dans le noeud de modélisation



Evaluation de modèle

Calculer l'importance des valeurs prédites. Pour des modèles qui produisent une mesure appropriée d'importance, vous pouvez afficher un graphique qui indique l'importance relative de chaque valeur prédite dans l'estimation du modèle. En général, vous souhaitez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Notez que l'importance des valeurs prédites peut prendre davantage de temps à être calculée pour certains modèles, en particulier si vous travaillez sur de plus grands ensembles de données, et qu'elle est par conséquent désactivée par défaut pour certains modèles. L'importance des valeurs prédites n'est pas disponible pour des modèles de liste de décision. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites sur p. 57.](#)

Scores de propension

Des scores de propension peuvent être établis dans le noeud de modélisation et dans l'onglet Paramètres du nugget de modèle. Cette fonctionnalité n'est disponible que lorsque la cible sélectionnée est un champ booléen. [Pour plus d'informations, reportez-vous à la section Scores de propension sur p. 47.](#)

Calculer les scores de propension brute. Les scores de propension brute sont calculés à partir du modèle uniquement en fonction des données d'apprentissage. Si le modèle prédit la valeur *true* (*vrai*) (que cette valeur va être la réponse), alors la propension est identique à P, où P est

la probabilité de la prédiction. Si le modèle prédit la valeur false (faux), alors la propension est calculée comme $(1 - P)$.

- Si vous choisissez cette option lors de la construction du modèle, des scores de propension sont activés par défaut dans le nugget du modèle. Cependant, vous pouvez toujours choisir d'activer des scores de propension brute dans le nugget du modèle que vous les sélectionnez ou non dans le noeud de modélisation.
- Lors du scoring du modèle, des scores de propension brute seront ajoutés dans un champ avec les lettres *RP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RRP-churn*.

Calculer les scores de propension ajustée. Les propensions brutes sont uniquement basées sur des estimations données par le modèle, qui peut être surajusté, ce qui entraîne des estimations trop optimistes de la propension. Les propensions ajustées tentent de compenser en vérifiant comment le modèle se comporte sur des partitions de test ou de validation et en ajustant les propensions pour donner une meilleure estimation en conséquence.

- Ce paramètre nécessite qu'un champ de partition valide soit présent dans le flux. [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*.](#)
- A la différence des scores de confiance brute, les scores de propension ajustée doivent être calculés lors de la construction du modèle ; sinon, ils ne seront pas disponibles lors du scoring du nugget du modèle.
- Lors du scoring du modèle, des scores de propension ajustée seront ajoutés dans un champ avec les lettres *AP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RAP-churn*. Les scores de propension ajustée ne sont pas disponibles pour des modèles de régression logistique.
- Lors du calcul des scores de propension ajustée, la partition de test ou de validation utilisée pour le calcul ne doit pas avoir été équilibrée. Pour éviter cela, assurez-vous que l'option Equilibrer uniquement les données d'apprentissage est sélectionnée dans tous les noeuds Equilibrer en amont. [Pour plus d'informations, reportez-vous à la section Définition des options du noeud Equilibrer dans le chapitre 3 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#) De plus, si un échantillon complexe a été pris en amont, cela invalide les scores de propension ajustée.
- Les scores de propension ajustée ne sont pas disponibles pour les modèles d'arbre "amélioré" et les modèles d'ensemble de règles. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés dans le chapitre 6 sur p. 193.](#)

En fonction de. Pour que les scores de propension ajustée soient calculés, un champ de partition doit être présent dans le flux. Vous pouvez spécifier s'il faut utiliser la partition de test ou de validation pour ce calcul. Pour de meilleurs résultats, la partition de test ou de validation doit comprendre au moins autant d'enregistrements que la partition utilisée pour l'apprentissage du modèle original. [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*.](#)

Scores de propension

Pour des modèles renvoyant une prédiction *oui* ou *non*, vous pouvez demander des scores de propension en plus des valeurs de prédiction et de confiance standard. Les scores de propension indiquent la probabilité d'une prévision ou d'une réponse particulière. Par exemple :

Table 3-1
Scores de propension

Client	Propension de réponse
Joe Smith	35%
Jane Smith	15%

Les scores de propension sont disponibles uniquement pour des modèles avec des cibles de type booléen, et indiquent la vraisemblance de la valeur *true* (*vraie*) définie pour le champ, comme spécifié dans un noeud source ou Typer. [Pour plus d'informations, reportez-vous à la section Spécification des valeurs d'un champ booléen dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Scores de propension / scores de confiance

Les scores de propension diffèrent des scores de confiance, qui s'appliquent à la prédiction actuelle, qu'elle soit *oui* ou *non*. Dans les cas où la prédiction est *non*, par exemple, une confiance élevée signifie en réalité une forte vraisemblance *de non* réponse. Les scores de propension contournent cette limite en autorisant une comparaison plus facile avec tous les enregistrements. Par exemple, une prédiction *non* avec une confiance de *0,85* se traduit par une propension brute de *0,15* (soit *1 moins 0,85*).

Table 3-2
Scores de confiance

Client	Prévision	Confiance
Joe Smith	Va répondre	.35
Jane Smith	Ne va pas répondre	.85

Obtention de score de propension

- Des scores de propension peuvent être établis dans l'onglet Analyser du noeud de modélisation ou dans l'onglet Paramètres du nugget de modèle. Cette fonctionnalité n'est disponible que lorsque la cible sélectionnée est un champ booléen. [Pour plus d'informations, reportez-vous à la section Options d'analyse des noeuds de modélisation sur p. 44.](#)
- Des scores de propension peuvent aussi être calculés par le noeud Ensemble, en fonction de la méthode d'ensemble utilisée. [Pour plus d'informations, reportez-vous à la section Noeud Ensemble dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Calcul de scores de propension ajustée

Les scores de propension ajustée sont calculés en tant que partie du processus de construction du modèle, et ne sont pas disponibles autrement. Une fois le modèle construit, son score est ensuite établi grâce à des données de la partition de test ou de validation, et un nouveau modèle est

construit pour livrer des scores de propension ajustée en analysant les performances du modèle original sur cette partition. En fonction du type de modèle, une des deux méthodes peut être utilisée pour calculer les scores de propension ajustée.

- Pour les modèles de règles et d'arbres, les scores de propension ajustée sont générés en recalculant l'effectif de chaque catégorie à chaque noeud arbre (pour les modèles d'arbres) ou la prise en charge et la confiance de chaque règle (pour les modèles à règles). Cela entraîne un nouvel ensemble de règles ou modèle d'arbre qui est stocké dans le modèle original, pour être utilisé à chaque fois que des scores de propension ajustée sont nécessaires. A chaque fois que le modèle original est appliqué aux nouvelles données, le nouveau modèle peut être appliqué ultérieurement aux scores de propension brute pour générer les scores ajustés.
- Pour d'autres modèles, des enregistrements produits par le scoring du modèle original sur la partition de test ou de validation sont alors discrétisés par leur score de propension brute. Ensuite, un modèle de réseau de neurones est formé et définit une fonction non linéaire qui effectue un mappage de la propension brute moyenne dans chaque intervalle à la propension moyenne observée dans le même intervalle. Comme indiqué précédemment pour les modèles d'arbres, le modèle de réseau de neurones résultant est stocké avec le modèle original et peut être appliqué aux scores de propension brute à chaque fois que des scores de propension ajustée sont nécessaires.

Précautions concernant les valeurs manquantes dans la partition de test. Le traitement des valeurs manquantes dans la partition de test/de validation varie selon le modèle (reportez-vous aux algorithmes d'évaluation de modèles individuels pour plus de détails). Le modèle C5 ne peut pas calculer la propension ajustée lorsqu'il existe des valeurs manquantes.

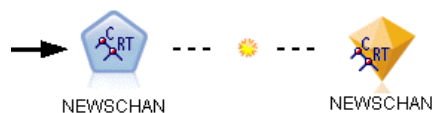
Nuggets de modèle

Figure 3-7
Nugget de modèle



Un modèle de nugget est un conteneur destiné à un modèle, c'est-à-dire l'ensemble des règles, des formules et des équations qui représentent les résultats des opérations de construction de votre modèle dans IBM® SPSS® Modeler. 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. L'ouverture d'un nugget de modèle sur l'écran vous permet de consulter plusieurs détails concernant le modèle, tels que l'importance relative dans le champ d'entrée de la création du modèle. Pour afficher les prévisions, vous devez attacher et exécuter un noeud d'exécution ou de sortie. [Pour plus d'informations, reportez-vous à la section Utilisation de nuggets de modèle dans les flux sur p. 70.](#)

Figure 3-8
Lien de modèle entre un noeud de modélisation et un nugget de modèle



Lorsque vous êtes parvenu à exécuter un noeud de modélisation, un nugget de modèle correspondant est placé dans l'espace de travail du flux, où il est représenté par une icône en forme de diamant doré (d'où le nom « nugget » (pépite en français)). Dans l'espace de travail du flux, le nugget est affiché avec une connexion (ligne pleine) au noeud approprié le plus proche précédant le noeud de modélisation et un lien (ligne en pointillé) au noeud de modélisation même.

Le nugget est aussi placé dans la palette Modèles dans le coin supérieur droit de la fenêtre de SPSS Modeler. Il est possible de sélectionner et de parcourir les nuggets à partir de ces emplacements afin d'afficher les détails du modèle.

Les nuggets sont toujours placés dans la palette Modèles lorsqu'un de modélisation est exécuté avec succès. Vous pouvez définir une option utilisateur afin de vérifier si le nugget est en outre placé dans l'espace du travail du flux. [Pour plus d'informations, reportez-vous à la section Définition d'options de notification dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Les rubriques suivantes fournissent des informations sur l'utilisation des nuggets de modèle dans SPSS Modeler. Pour mieux comprendre les algorithmes utilisés, consultez le guide *Guide des algorithmes SPSS Modeler*, disponible dans le dossier \Documentation de DVD pour IBM® SPSS® Modeler.

Liens de modèle

Par défaut, un nugget est affiché dans l'espace de travail avec un lien pointant vers le noeud de modélisation qui l'a créé. Ceci est particulièrement utile avec des flux complexes comportant plusieurs nuggets, et vous permet d'identifier le nugget qui sera mis à jour par chaque noeud de modélisation. Chaque lien contient un symbole qui indique si le modèle est remplacé lorsque le noeud de modélisation est exécuté. [Pour plus d'informations, reportez-vous à la section Remplacement d'un modèle sur p. 51.](#)

Définition et suppression de liens de modèle

Vous pouvez définir et supprimer manuellement des liens dans l'espace de travail. Lorsque vous définissez un nouveau lien, le curseur change et devient un curseur de lien.

Figure 3-9
Curseur de lien



Définition d'un nouveau lien (menu contextuel)

- ▶ Cliquez avec le bouton droit de la souris sur le noeud de modélisation à partir duquel vous souhaitez que le lien commence.
- ▶ Sélectionnez Définir un lien de modèle dans le menu contextuel.
- ▶ Cliquez sur le nugget où vous souhaitez que le lien se termine.

Définition d'un nouveau lien (menu principal)

- ▶ Cliquez sur le noeud de modélisation à partir duquel vous souhaitez que le lien commence.
- ▶ Dans le menu principal, sélectionnez :
Edition > Nœud > Définir un lien de modèle
- ▶ Cliquez sur le nugget où vous souhaitez que le lien se termine.

Suppression d'un lien existant (menu contextuel)

- ▶ Cliquez avec le bouton droit sur le nugget à l'extrémité du lien.
- ▶ Sélectionnez Supprimer un lien de modèle dans le menu contextuel.
Sinon :
- ▶ Cliquez avec le bouton droit sur le symbole situé au milieu du lien.
- ▶ Sélectionnez Supprimer un lien dans le menu contextuel.

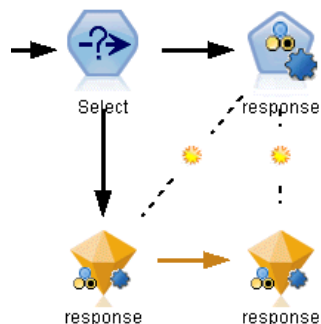
Suppression d'un lien existant (menu principal)

- ▶ Cliquez sur le noeud ou sur le nugget de modélisation à partir duquel vous souhaitez supprimer le lien.
- ▶ Dans le menu principal, sélectionnez :
Edition > Nœud > Supprimer un lien de modèle

Copier et coller des liens de modèle

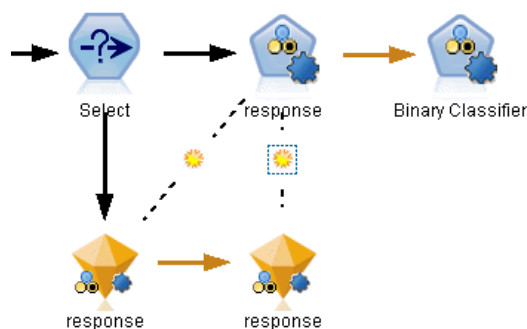
Si vous copiez un nugget qui possède des liens sans son noeud de modélisation, et que vous le collez dans le même flux, le nugget est collé avec un lien pointant sur le noeud de modélisation. Le nouveau lien possède le même état de remplacement de modèle (reportez-vous à Remplacement d'un modèle sur p. 51) que son lien d'origine :

Figure 3-10
Copier et coller un nugget qui possède un lien



Si vous copiez et collez un nugget avec son noeud de modélisation qui possède un lien, le lien est conservé, que les objets soient collés dans le même flux ou dans un nouveau flux :

Figure 3-11
Copier et coller un nugget qui possède un lien



Remarque : Si vous copiez un nugget qui possède un lien sans son noeud de modélisation, et que vous collez le nugget dans un nouveau flux (ou dans un super noeud qui ne contient pas le noeud de modélisation), le lien est brisé et seul le nugget est collé.

Liens de modèle et super noeuds

Si vous définissez un super noeud pour inclure soit le noeud de modélisation soit le nugget de modèle d'un modèle qui possède un lien (mais pas les deux), le lien est brisé. Le développement du super noeud ne restaure pas le lien ; vous ne pouvez réaliser ceci qu'en annulant la création du super noeud.

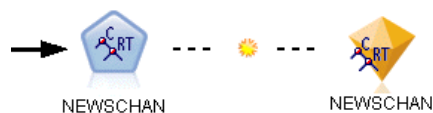
Remplacement d'un modèle

Vous pouvez choisir de remplacer (c'est-à-dire de mettre à jour) un nugget existant lors de la réexécution du noeud de modélisation qui a créé le nugget. Si vous désactivez l'option de remplacement, un nouveau nugget est créé lorsque vous réexécutez le noeud de modélisation.

Remarque : Le remplacement d'un modèle est différent du rafraîchissement d'un modèle, lequel se rapporte à la mise à jour d'un modèle dans un scénario. [Pour plus d'informations, reportez-vous à la section Rafraîchissement de modèle dans le chapitre 9 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Chaque lien à partir d'un noeud de modélisation vers un nugget contient un symbole qui indique si le modèle est remplacé lorsque le noeud de modélisation est réexécuté.

Figure 3-12
Lien de modèle lorsque le remplacement de modèle est activé



À l'origine, le lien est affiché avec le remplacement de modèle activé, indiqué par le petit symbole de soleil levant dans le lien. Dans cet état, la réexécution du noeud de modélisation à une extrémité du lien met simplement à jour le nugget à l'autre extrémité.

Figure 3-13

Lien de modèle lorsque le remplacement de modèle est désactivé



Si le remplacement de modèle est désactivé, le symbole du lien est remplacé par un point gris. Dans cet état, la réexécution du noeud de modélisation à une extrémité du lien ajoute une nouvelle version mise à jour du nugget dans l'espace de travail.

Dans les deux cas, dans la palette Modèles, le nugget existant est mis à jour ou un nouveau nugget est ajouté, selon la configuration de l'option du système Remplacer le modèle précédent. [Pour plus d'informations, reportez-vous à la section Définition d'options de notification dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15.*](#)

Ordre d'exécution

Lorsque vous exécutez un flux ayant plusieurs branches contenant des nuggets de modèle, le flux est d'abord évalué afin de vérifier qu'une branche dont le remplacement de modèle est activé est exécutée avant tout autre branche qui utilise le nugget de modèle résultant.

Si vos exigences sont plus complexes, vous pouvez définir l'ordre d'exécution manuellement via la fonction de génération de scripts.

Modification des paramètres de remplacement de modèle

Pour modifier le paramètre de remplacement de modèle :

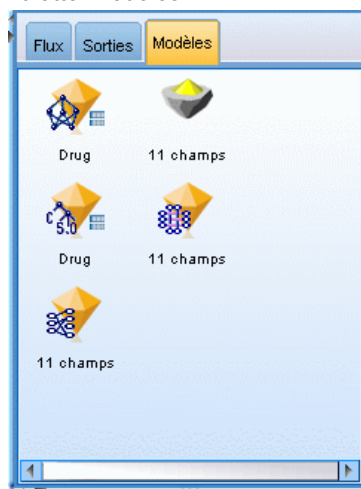
- ▶ Cliquez avec le bouton droit de la souris sur le symbole du lien.
- ▶ Sélectionnez (Dés)activer le remplacement de modèle selon votre choix.

Remarque : Le paramètre de remplacement de modèle sur un lien de modèle remplace le paramètre de l'onglet Notifications de la boîte de dialogue Options utilisateur (Outils > Options > Options utilisateur).

Palette Modèles

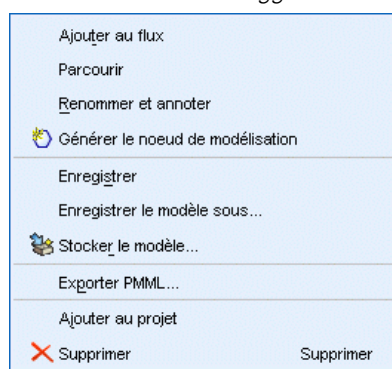
La palette de modèles (dans l'onglet Modèles de la fenêtre des gestionnaires) met à votre disposition différents moyens pour utiliser, examiner et modifier des nuggets de modèle.

Figure 3-14
Palette Modèles



Cliquer avec le bouton droit de la souris sur un nugget de modèle dans la palette de modèles permet d'ouvrir un menu contextuel avec les options suivantes :

Figure 3-15
Menu contextuel de nugget de modèle

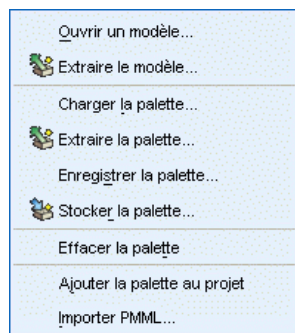


- **Ajouter au flux.** Ajoute le nugget de modèle généré au flux actif. Si un noeud est sélectionné dans le flux, le nugget de modèle est relié au noeud sélectionné si une connexion est possible, sinon au noeud possible le plus proche. Le nugget est affiché avec un lien au noeud de modélisation qui a créé le modèle, si ce noeud se trouve encore dans le flux.
- **Parcourir...** Ouvre le navigateur de modèle du nugget.
- **Renommer et annoter.** Permet de renommer le nugget de modèle et/ou de modifier l'annotation du nugget.
- **Générer un noeud de modélisation.** Si vous souhaitez modifier ou mettre à jour l'un de vos nuggets de modèle et que le flux utilisé pour créer le modèle n'est pas disponible, vous pouvez utiliser cette option pour recréer un noeud de modélisation avec les options qui avaient servi à créer le modèle d'origine.
- **Enregistrer le modèle, Enregistrer le modèle sous.** Enregistre le nugget de modèle dans un fichier binaire de modèle généré (.gm) externe.

- **Stocker le modèle.** Stocke le nugget de modèle dans le IBM® SPSS® Collaboration and Deployment Services Repository. [Pour plus d'informations, reportez-vous à la section A propos du IBM SPSS Collaboration and Deployment Services Repository dans le chapitre 9 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)
- **Exporter PMML.** Permet d'exporter le nugget de modèle sous la forme d'un fichier en langage de balisage de modèle de prévision (PMML), qui peut ensuite être utilisé pour évaluer de nouvelles données en dehors de IBM® SPSS® Modeler. Exporter PMML est disponible pour tous les noeuds de modèle générés. *Remarque* : Une licence pour IBM® SPSS® Modeler Server est nécessaire pour utiliser cette fonctionnalité. [Pour plus d'informations, reportez-vous à la section Définition des options d'exportation PMML dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)
- **Ajouter au projet.** Enregistre le nugget de modèle et l'ajoute au projet en cours. Dans l'onglet Classes, le nugget est ajouté au dossier Modèles générés. Dans l'onglet CRISP-DM, il est ajouté à la phase du projet par défaut. (Reportez-vous à [Définition de la phase par défaut du projet](#) pour plus d'informations sur le mode de changement de la phase du projet par défaut.)
- **Supprimer.** Supprime le nugget de modèle de la palette.

Figure 3-16

Menu contextuel de la palette de modèles



Cliquer avec le bouton droit de la souris sur une zone libre de la palette de modèles permet d'ouvrir un menu contextuel contenant les options suivantes :

- **Ouvrir un modèle.** Charge un nugget de modèle créé dans SPSS Modeler.
- **Extraire le modèle.** Extrait un modèle stocké d'un référentiel IBM SPSS Collaboration and Deployment Services.
- **Charger la palette.** Charge une palette de modèles enregistrée depuis un fichier externe.
- **Extraire une palette.** Extrait une palette de modèles stockée d'un référentiel IBM SPSS Collaboration and Deployment Services.
- **Enregistrer la palette.** Enregistre le contenu complet de la palette de modèles dans un fichier de palette de modèles générée (.gen) externe.
- **Stocker une palette.** Stocke l'intégralité du contenu de la palette de modèles dans un référentiel IBM SPSS Collaboration and Deployment Services.
- **Effacer la palette.** Supprime tous les nuggets de la palette.

- **Ajouter une palette au projet.** Enregistre la palette de modèles et l'ajoute au projet en cours. Dans l'onglet Classes, le nugget est ajouté au dossier Modèles générés. Dans l'onglet CRISP-DM, il est ajouté à la phase du projet par défaut.
- **Importer PMML.** Charge un modèle depuis un fichier externe. Vous pouvez ouvrir, parcourir et déterminer le score de modèles PMML créés par IBM® SPSS® Statistics ou d'autres applications qui prennent en charge ce format. [Pour plus d'informations, reportez-vous à la section Importation et exportation de modèles au format PMML dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Navigation dans les nuggets de modèle

Les navigateurs de nuggets de modèle vous permettent d'examiner et d'utiliser les résultats de vos modèles. Depuis le navigateur, vous pouvez enregistrer, imprimer ou exporter le modèle généré, examiner le récapitulatif du modèle, et afficher ou modifier les annotations du modèle. Pour certains types de nuggets de modèle, vous pouvez également générer de nouveaux noeuds, comme les noeuds Filtrer ou Ensemble de règles. Pour certains modèles, vous pouvez également afficher les paramètres du modèle, comme les règles ou les centres de classes. Pour certains types de modèle (modèles en arborescence et modèles de classes), il est possible d'afficher une représentation graphique de la structure du modèle. Les commandes permettant d'utiliser les navigateurs de nugget de modèle sont décrites ci-dessous.

Menus

Menu Fichier. Tous les nuggets de modèle disposent d'un menu Fichier contenant un sous-ensemble des options suivantes :

- **Enregistrer le noeud.** Enregistre le nugget de modèle dans un fichier de noeud (.nod).
- **Stocker un noeud.** Stocke le nugget de modèle dans un référentiel IBM SPSS Collaboration and Deployment Services.
- **En-tête et pied de page.** Permet d'éditer l'en-tête et le pied de page d'impression à partir du nugget.
- **Mise en page.** Permet de changer la mise en page pour l'impression à partir du nugget.
- **Aperçu avant impression.** Affiche un aperçu de l'aspect du nugget une fois imprimé. Sélectionnez les informations à afficher à partir du sous-menu.
- **Imprimer.** Imprime le contenu du nugget. Sélectionnez les informations à imprimer à partir du sous-menu.
- **Imprimer la vue** Imprime la vue actuelle ou toutes les vues.
- **Exporter texte.** Exporte le contenu du nugget dans un fichier texte. Sélectionnez les informations à exporter à partir du sous-menu.
- **Exporter HTML.** Exporte le contenu du nugget dans un fichier HTML. Sélectionnez les informations à exporter à partir du sous-menu.
- **Exporter PMML.** Exporte le modèle au format PMML (langage de balisage de modèle de prévision) 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 *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#) *Remarque* : Une licence

pour IBM® SPSS® Modeler Server est nécessaire pour utiliser cette fonctionnalité. [Pour plus d'informations, reportez-vous à la section Définition des options d'exportation PMML dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

- **Exporter SQL.** Exporte le modèle au format SQL (Structured Query Language), que vous pouvez éditer et utiliser dans d'autres bases de données.

Remarque : La fonctionnalité d'export SQL n'est disponible qu'avec les modèles suivants : C5, C&RT, CHAID, QUEST, Régression linéaire, Régression logistique, R. neurones, ACP/Facteur et modèles de Liste de décision. [Pour plus d'informations, reportez-vous à la section Noeuds prenant en charge la génération SQL dans le chapitre 6 dans *Guide d'administration et des performances de IBM SPSS Modeler Server 15*.](#)

- **Publier pour l'adaptateur de scoring de serveur.** Publie le modèle dans une base de données qui contient un adaptateur de scoring ce qui permet d'exécuter le scoring du modèle dans la base de données. [Pour plus d'informations, reportez-vous à la section Publication des modèles pour un adaptateur de scoring sur p. 76.](#)

Menu Générer. La plupart des nuggets de modèle disposent également d'un menu Générer. Il vous permet de générer de nouveaux noeuds sur la base du nugget de modèle. Les options disponibles dans ce menu dépendent du type de modèle que vous explorez. Reportez-vous à la section consacrée au type de nugget de modèle qui vous intéresse pour savoir ce que ce modèle vous permet de générer.

Menu Affichage. Dans l'onglet Modèle d'un nugget, ce menu vous permet d'afficher ou de masquer les diverses barres d'outils de visualisation disponibles dans le mode en cours. Pour accéder à l'ensemble complet des barres d'outils, sélectionnez Mode d'édition (icône du pinceau) dans la barre d'outils Général.

Bouton Aperçu. Certains nuggets de modèles disposent d'un bouton Aperçu, qui vous permet d'afficher un échantillon des données de modèle, y compris les champs supplémentaires créés par le processus de modélisation. Le nombre de lignes affichées par défaut est de 10, mais vous pouvez le modifier dans les propriétés du flux. [Pour plus d'informations, reportez-vous à la section Définition d'options générales pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Bouton Ajouter au projet en cours. Enregistre le nugget de modèle et l'ajoute au projet en cours. Dans l'onglet Classes, le nugget est ajouté au dossier Modèles générés. Dans l'onglet CRISP-DM, il est ajouté à la phase du projet par défaut. (Reportez-vous à [Définition de la phase par défaut du projet](#) pour plus d'informations sur le mode de changement de la phase du projet par défaut.)

Récapitulatif /Informations sur les nuggets de modèle

L'onglet Récapitulatif ou la vue Informations d'un nugget de modèle affiche des informations sur les champs, les paramètres de création et le processus d'estimation du modèle. Les résultats sont présentés dans un arbre que vous pouvez développer ou réduire en cliquant sur des éléments précis.

Analyse. Affiche des informations sur le modèle. Les détails varient par type de modèle (reportez-vous aux sections correspondant à chaque nugget de modèle). En outre, si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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. Pour les modèles découpés, répertorie également les champs qui ont déterminé les découpages.

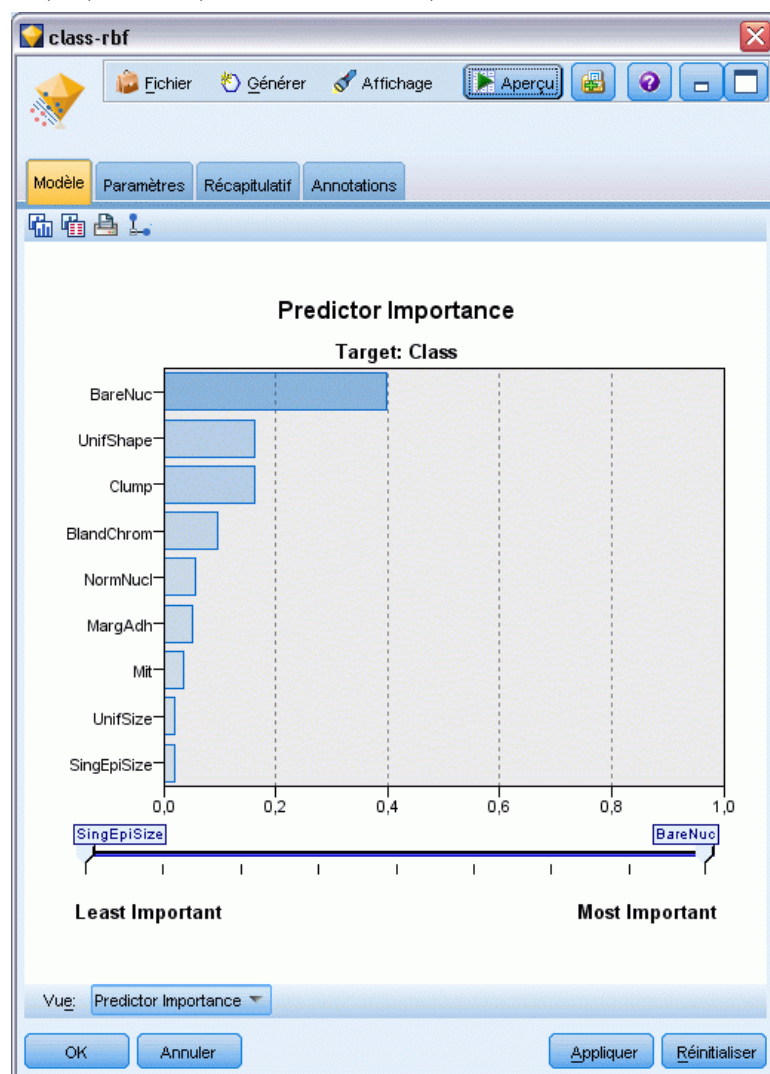
Paramètres / options de création. Contient des informations sur les paramètres utilisés lors de la création du modèle.

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.

Importance des valeurs prédites

En général, vous préférerez concentrer vos efforts de modélisation sur les champs prédictifs les plus importants et abandonner ou ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Étant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des valeurs prédites affichée est 1,0. L'importance des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Figure 3-17
Graphique de l'importance des valeurs prédites



L'importance des valeurs prédites est disponible pour des modèles qui produisent une mesure statistique appropriée de l'importance, y compris des réseaux de neurones, arbres de décision (C&R Tree, C5.0, CHAID, et QUEST), réseaux Bayésiens, discriminant, SVM, et modèles MRAA, régressions linéaire et logistique, modèles linaires généralisés et modèles d'agrégation suivant le saut minimum (KNN). Pour la plupart de ces modèles, l'importance des variables prédites peut être activée dans l'onglet Analyser du noeud de modélisation. [Pour plus d'informations, reportez-vous à la section Options d'analyse des noeuds de modélisation sur p. 44.](#) Pour les modèles KNN, reportez-vous à Voisins sur p. 512.

Remarque : L'importance des valeurs prédites n'est pas prise en charge par les modèles découpés. Les paramètres de l'importance de la variable indépendante sont ignorés lors de la création de modèles découpés. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés sur p. 33.](#)

Le calcul de l'importance des valeurs prédites peut prendre un temps considérablement plus long que la construction de modèle, en particulier lors de l'utilisation de grands ensembles de données. Ce calcul est plus long pour les modèles SVM et de régression logistique que pour d'autres modèles et il est désactivé par défaut pour ces modèles. Si vous utilisez un ensemble de données possédant un grand nombre de valeurs prédites, un filtrage initial utilisant un noeud Sélection de fonction peut donner des résultats plus rapides (reportez-vous à la section ci-dessous).

- L'importance de la variable indépendante est calculée à partir de la partition de test, si elle est disponible. Sinon les données d'apprentissage sont utilisées. [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*.](#)
- Pour les modèles MRAA, l'importance de la variable indépendante est disponible mais elle est calculée par l'algorithme MRAA. [Pour plus d'informations, reportez-vous à la section Nuggets de modèle MRAA dans le chapitre 14 sur p. 493.](#)
- Vous pouvez utiliser les outils de graphiques de IBM® SPSS® Modeler pour interagir, modifier et enregistrer le graphique.
- Vous pouvez éventuellement générer un noeud Filtrer en fonction des informations du graphique d'importance de la variable indépendante. [Pour plus d'informations, reportez-vous à la section Filtrage de variables en fonction de l'importance sur p. 59.](#)

Importance de la variable indépendante et sélection de fonction

Le graphique d'importance de la variable indépendante affiché dans un nugget de modèle peut sembler donner des résultats similaires au noeud Sélection de fonction dans certains cas. Alors que la sélection de fonction classe chaque champ d'entrée en fonction de l'intensité de sa relation avec la cible spécifiée, indépendamment des autres entrées, le graphique d'importance de la variable indépendante indique l'importance relative de chaque entrée pour *ce* modèle particulier. Ainsi la sélection de fonction devient plus conservatrice pour le filtrage des entrées. Par exemple, si *l'intitulé de poste* et *la catégorie de poste* sont tous deux fortement liés au salaire, alors la sélection de fonction indique que les deux sont importants. Mais dans la modélisation, les interactions et les corrélations sont aussi prises en considération. Ainsi vous pouvez trouver qu'une seule entrée parmi deux entrées est utilisée si toutes deux contiennent en grande partie des informations identiques. Dans la pratique, la sélection de fonction est plus utile pour le filtrage préliminaire, particulièrement avec de grands ensembles de données comportant un grand nombre de variables, et l'importance de la variable indépendante est plus utile pour affiner le modèle.

Filtrage de variables en fonction de l'importance

Vous pouvez éventuellement générer un noeud Filtrer en fonction des informations du graphique d'importance de la variable indépendante.

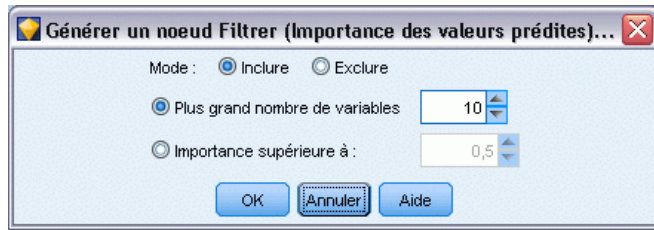
Marquez les variables indépendantes que vous souhaitez inclure sur le graphique, s'il y a lieu, et, dans les menus, choisissez :

Générer > Noeud Filtrer (Importance de la variable indépendante)

OU

> Sélection des champs (Importance des variables indépendantes)

Figure 3-18
 Filtrage de variables indépendantes en fonction de l'importance



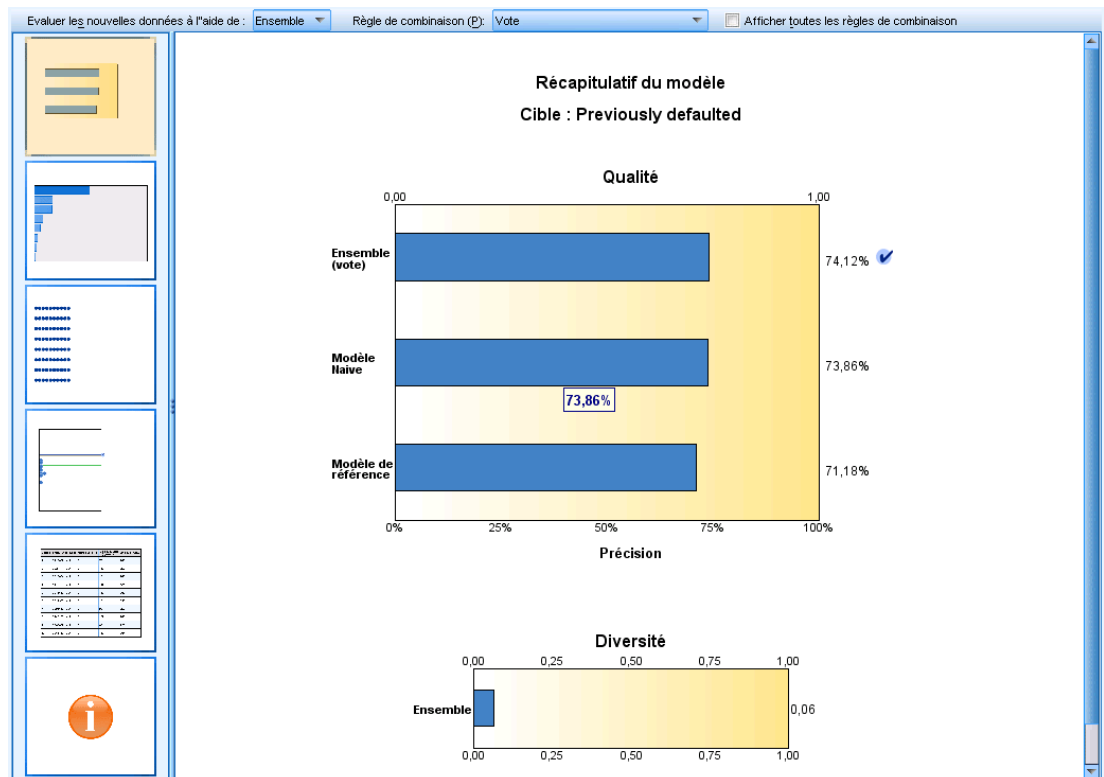
Plus grand nombre de variables. Inclut ou exclut les variables indépendantes les plus importantes jusqu'au nombre spécifié.

Importance supérieure à. Inclut ou exclut toutes les variables indépendantes dont l'importance relative est supérieure à la valeur spécifiée.

Modèles pour des ensembles

Le modèle d'un ensemble fournit des informations sur les modèles de composants de l'ensemble et les performances de l'ensemble en général.

Figure 3-19
 Vue récapitulative du modèle



La barre d'outils principale (indépendante de la vue) vous permet de choisir d'évaluer l'ensemble ou un modèle de référence. Si l'ensemble est utilisé pour l'évaluation, vous pouvez aussi sélectionner la règle de combinaison. Ces changements ne nécessitent pas de nouvelle exécution du modèle ; cependant, ces choix sont sauvegardés dans le (nugget) de modèle pour une évaluation et/ou pour une évaluation du modèle en aval. Ils ont aussi une influence sur le PMML exporté à partir du viewer d'ensemble.

Règle de combinaison. Lors de l'évaluation d'un ensemble, il s'agit de la règle utilisée pour combiner les valeurs prédites à partir des modèles de base pour calculer la valeur de score d'un ensemble.

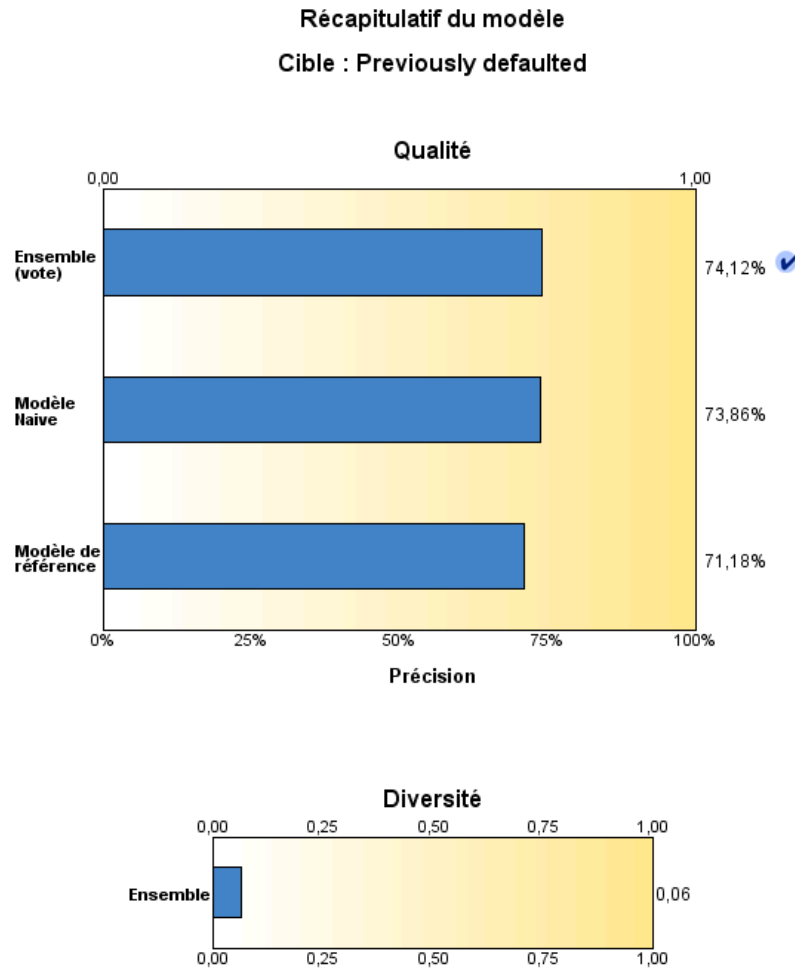
- Les valeurs prédites d'ensemble pour des cibles **qualitatives** peuvent être combinées au moyen du vote, de la probabilité la plus élevée ou de la probabilité de moyenne la plus élevée. Le **vote** sélectionne la catégorie ayant le plus souvent la probabilité la plus élevée parmi les modèles de base. La **probabilité la plus élevée** sélectionne la catégorie ayant la probabilité unique la plus élevée parmi tous les modèles de base. La **probabilité de moyenne la plus élevée** sélectionne la catégorie ayant la valeur la plus élevée lorsqu'on réalise la moyenne des probabilités de catégorie parmi les modèles de base.
- Des valeurs prédites d'ensemble pour des cibles **continues** peuvent être combinées à l'aide de la moyenne ou de la médiane des valeurs prédites à partir des modèles de base.

La valeur par défaut provient des spécifications réalisées lors de la création du modèle. Une modification de la règle de combinaisons recalcule la précision du modèle et met à jour toutes les vues de la précision du modèle. Le diagramme d'importance de la valeur prédite est aussi mis à jour. Cette commande est désactivée si le modèle de référence est sélectionné pour l'évaluation.

Afficher toutes les règles de combinaison. Lorsque cette option est sélectionnée, les résultats de toutes les règles de combinaisons disponibles sont affichés dans le diagramme de qualité du modèle. Le diagramme de précision du modèle de composants est aussi mis à jour afin d'afficher les lignes de référence pour chaque méthode de vote.

Récapitulatif de modèle

Figure 3-20
Vue récapitulative du modèle



La vue Récapitulatif du modèle est un instantané, permettant de consulter en un coup d'œil la qualité et la diversité de l'ensemble.

Qualité. Le diagramme affiche la précision du modèle final, comparé à un modèle de référence et à un modèle Naive. La précision est présentée dans un format plus grand et meilleur ; le « meilleur » modèle dispose de la précision la plus élevée. Pour une cible qualitative, la précision est tout simplement le pourcentage des enregistrements pour lesquels la valeur de prédiction correspond à la valeur observée. Pour une cible continue, la précision est égale à 1 moins le ratio de l'erreur moyenne absolue en prédiction (la moyenne des valeurs absolues des valeurs prédites moins les valeurs observées) de la plage des valeurs prédites (la valeur prédite maximale moins la valeur prédite minimale).

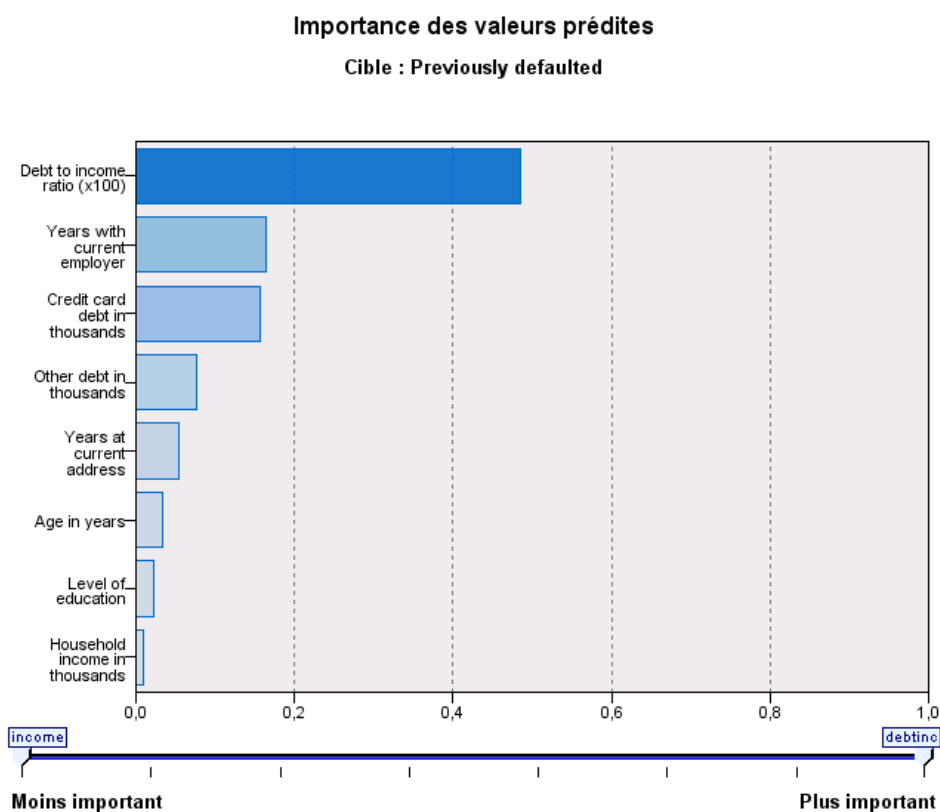
Pour les ensembles de bagging, le modèle de référence est un modèle standard construit sur l'intégralité de la partition de formation. Pour les ensembles boostés, le modèle de référence est le premier modèle de composant.

Le modèle Naive représente la précision si aucun modèle n'était construit, et attribue tous les enregistrements à la catégorie modale. Le modèle Naive n'est pas calculé pour des cibles continues.

Diversité. Le graphique affiche la « diversité d'opinions » parmi les modèles de composants utilisés pour construire l'ensemble, présenté dans un format plus grand et plus varié. Il s'agit d'une mesure indiquant la variation des prédictions parmi les modèles de base. La diversité n'est pas disponible pour les modèles d'ensemble boostés et elle n'est pas non plus affichée pour les cibles continues.

Importance des valeurs prédites

Figure 3-21
Vue Importance des valeurs prédites



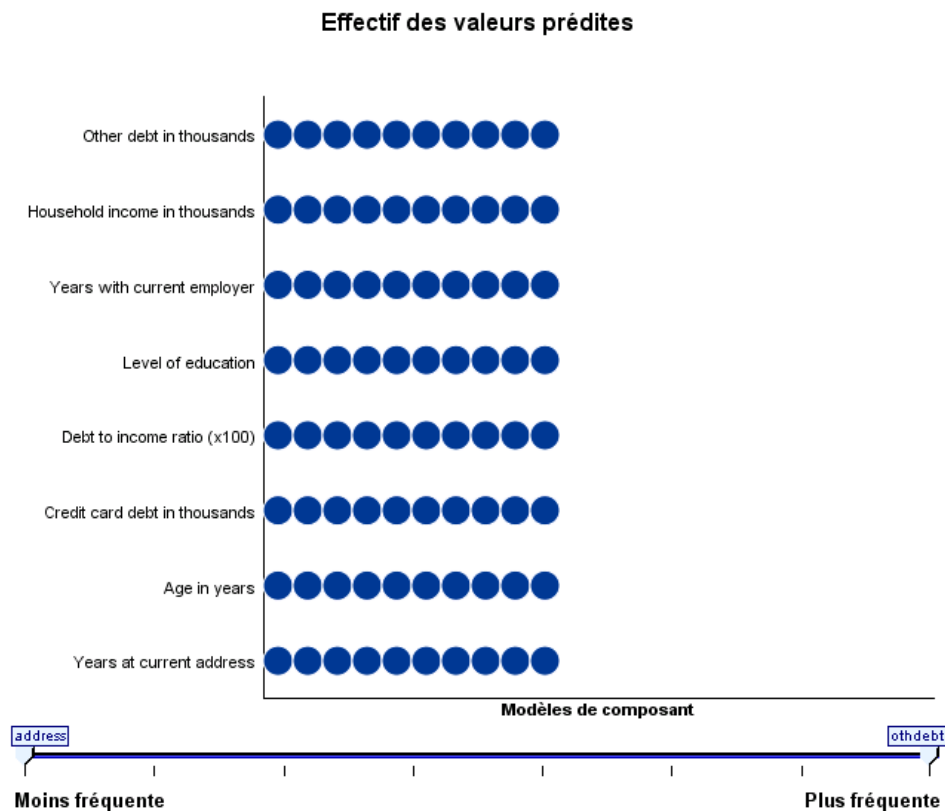
En général, vous préférerez concentrer vos efforts de modélisation sur les champs prédictifs les plus importants et abandonner ou ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Étant donné que les valeurs sont relatives, la somme des valeurs pour

L'ensemble des valeurs prédites affichée est 1,0. L'importance des des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

L'importance des valeurs prédites n'est pas disponible pour tous les modèles d'ensemble. L'ensemble des valeurs prédites peuvent varier parmi les modèles de composants, mais l'importance doit être calculée pour les valeurs prédites utilisées dans au moins un modèle de composant.

Fréquence des valeurs prédites

Figure 3-22
Vue Fréquence des valeurs prédites



L'ensemble des valeurs prédites peut varier parmi les modèles de composants selon le choix de la méthode de modélisation ou de la sélection des valeurs prédites. Le diagramme de la fréquence des valeurs prédites est un graphique en points qui affiche la distribution des valeurs prédites parmi les modèles de composants dans l'ensemble. Chaque point représente un ou plusieurs modèles de composants qui contient la valeur prédite. Les valeurs prédites sont représentées sur l'axe y et sont triées en ordre décroissant de fréquence ; par conséquent, la valeur prédite la plus haute est celle utilisée dans le plus grand nombre de modèles de composants, et la valeur prédite

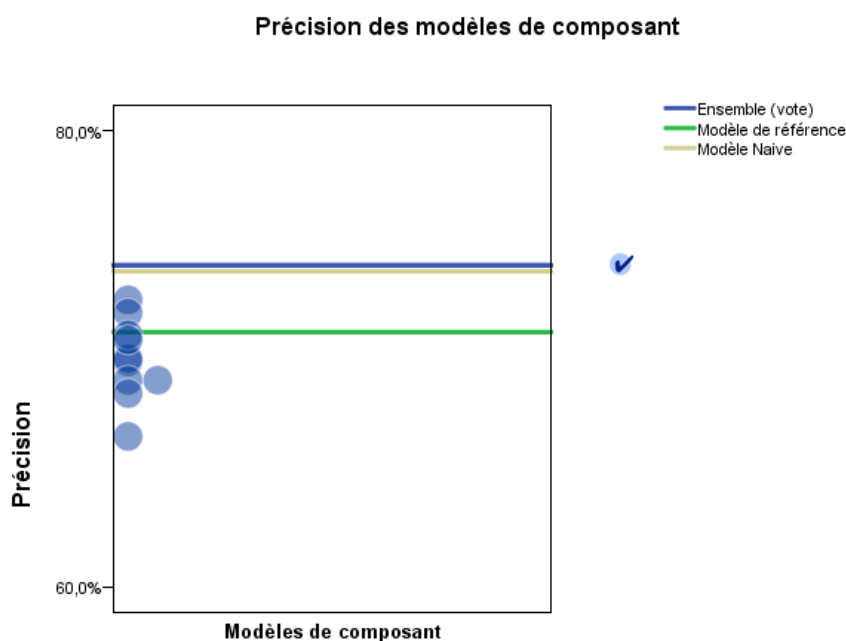
la plus basse est celle utilisée dans le moins grand nombre de modèles de composants. Les 10 valeurs prédites les plus hautes sont affichées.

Les valeurs prédites qui apparaissent le plus fréquemment sont généralement les plus importantes. Ce diagramme n'est pas pratique pour les méthodes dans lesquelles l'ensemble des valeurs prédites ne peut pas varier parmi les modèles de composants.

Précision d'un modèle de composant

Figure 3-23

Vue Précision d'un modèle de composant



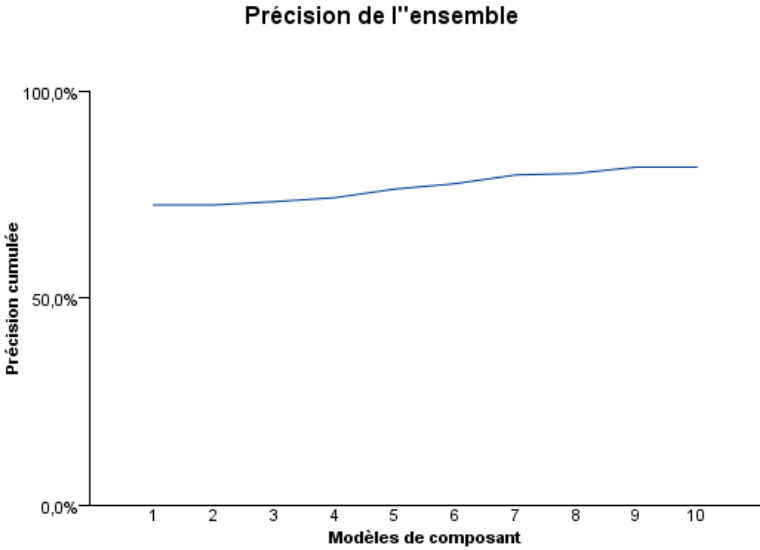
Le diagramme est un graphique en points du pouvoir prédictif des modèles de composants. Chaque point représente un ou plusieurs modèles de composants avec le niveau de précision indiqué sur l'axe y. Passez la souris sur un point pour obtenir des informations sur le modèle de composant individuel correspondant.

Lignes de référence. Le diagramme affiche des lignes de code de couleurs pour l'ensemble ainsi que le modèle de référence et les modèles Naive. Une coche s'affiche en regard de la ligne correspondant au modèle qui sera utilisé pour l'évaluation.

Interactivité. Le graphique est mis à jour si vous modifiez la règle de combinaison.

Ensembles boostés. Un diagramme curviligne s'affiche pour les ensembles boostés.

Figure 3-24
Vue Précision d'un ensemble, ensemble boosté



Détails d'un modèle de composant

Figure 3-25
Vue Détails d'un modèle de composant

Modèle	Précision	Méthode	Valeurs prédites	Taille du modèle (synapses)	Enregistrements
1	72,6%		8	77	700
2	72,0%		8	107	692
3	69,9%		8	92	708
4	70,0%		8	107	685
5	71,1%		8	107	706
6	69,1%		8	92	690
7	69,1%		8	92	696
8	70,8%		8	122	703
9	66,6%		8	62	726
10	68,5%		8	107	701

Le tableau affiche des informations sur les modèles de composants, répertoriés par ligne. Par défaut, les modèles de composants sont triés dans l'ordre croissant des numéros de modèle. Vous pouvez trier les lignes par ordre croissant ou décroissant des valeurs de n'importe quelle colonne.

Modèle : Un nombre représentant l'ordre séquentiel dans lequel le modèle de composants a été créé.

Précision. Précision générale sous la forme d'un pourcentage.

Méthode. La méthode de modélisation.

Valeurs prédites. Le nombre de valeurs prédites utilisées dans le modèle de composant.

Taille du modèle. La taille du modèle dépend de la méthode de modélisation : pour les arborescences, il s'agit du nombre de noeuds de l'arborescence ; pour les modèles linéaires, il s'agit du nombre de coefficients ; pour les réseaux neuronaux, il s'agit du nombre des synapses.

Enregistrements. Le nombre pondéré d'enregistrements d'entrée dans l'échantillon d'apprentissage.

Préparation automatique des données

Figure 3-26
Vue Préparation automatique des données

Préparation automatique des données		
Cible : Total sales		
Champ	Rôle	Actions entreprises
Age category	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible
Primary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible
Promotion	Valeur prédite	Modifier le niveau de mesure de continu en ordinal
Secondary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible

Si le nom du champ d'origine est X, le nom du champ transformé est X transformé. Le champ d'origine est exclu de l'analyse et le champ transformé est inclus à sa place.

Cette vue affiche des informations concernant les champs qui ont été exclus et la façon dont les champs transformés ont été dérivés dans l'étape de préparation automatique des données (ADP). Pour chaque champ transformé ou exclu, le tableau répertorie le nom du champ, son rôle au sein de l'analyse et l'action entreprise par l'étape ADP. Les champs sont triés selon l'ordre alphabétique croissant des noms de champ.

L'action Enlever les valeurs éloignées, lorsqu'elle est affichée, indique que les valeurs des prédicteurs continus situés au-delà d'une valeur de césure (écart-type de 3 par rapport à la moyenne) sont définies sur la valeur de césure.

Nuggets de modèle pour les modèles découpés

Le nugget de modèle pour un modèle découpé permet d'accéder à tous les modèles séparés créés par les découpages.

Un nugget de modèle découpé contient :

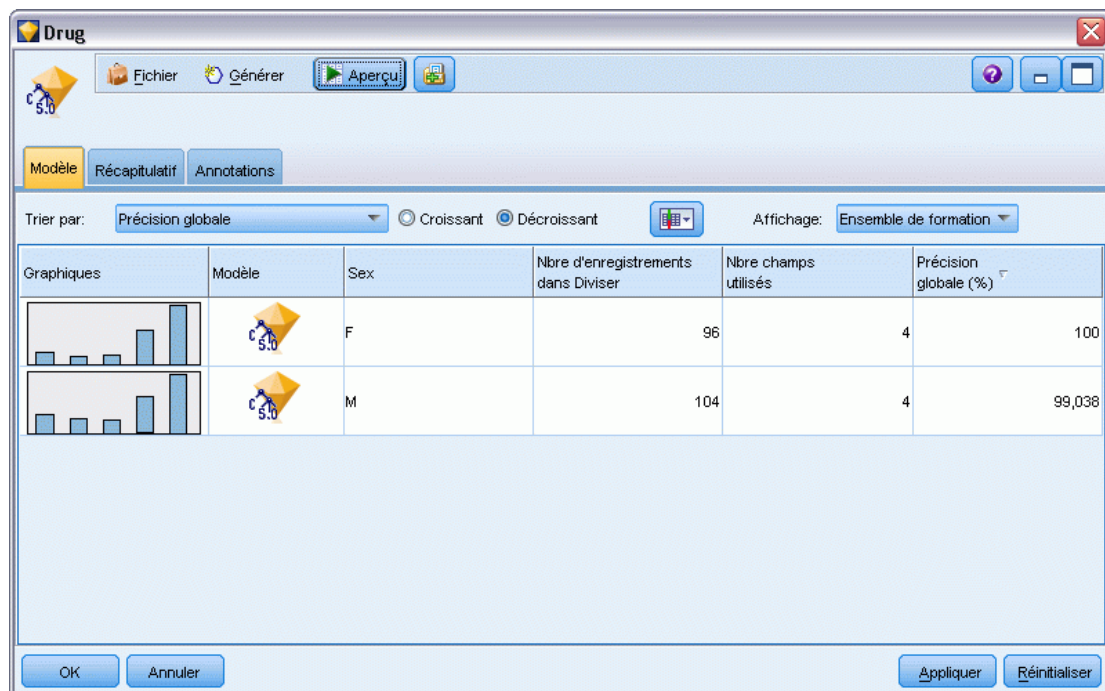
- une liste de tous les modèles découpés créés, ainsi qu'un ensemble de statistiques sur chaque modèle
- informations sur le modèle global

Dans la liste des modèles découpés, vous pouvez ouvrir des modèles individuels pour les examiner de manière plus approfondie.

Afficheur de modèle découpé

L'onglet Modèle répertorie tous les modèles contenus dans le nugget et fournit des statistiques sous différentes formes sur les modèles découpés. Il peut avoir deux formes générales, en fonction du noeud de modélisation, comme suit.

Figure 3-27
Afficheur de modèle découpé



Trier par. Utilisez cette liste pour choisir l'ordre d'apparition des modèles. Vous pouvez trier la liste en fonction des valeurs de l'une des colonnes de l'affichage, dans l'ordre croissant ou décroissant. Vous pouvez également cliquer sur l'en-tête d'une colonne en fonction de laquelle réaliser le tri. La présentation par défaut est l'ordre décroissant de la précision globale.

Menu Afficher/masquer les colonnes. Cliquez sur ce bouton pour afficher un menu dans lequel vous pouvez choisir des colonnes individuelles à afficher ou à masquer.

Affichage. Si vous utilisez le partitionnement, vous pouvez afficher les résultats des données d'apprentissage ou de test. [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.](#)

Pour chaque découpage, les informations présentées sont les suivantes :

Graphique. Miniature indiquant la distribution des données pour ce modèle. Lorsque le nugget est sur l'espace de travail, double-cliquez sur la miniature pour ouvrir le graphique grandeur nature.

Modèle. Icône du type de modèle. Double-cliquez sur l'icône pour ouvrir le nugget de modèle pour ce découpage spécifique.

Champs de découpage. Champs désignés dans le noeud de modélisation comme champs de découpage, avec leurs différentes valeurs possibles.

Nbre d'enregistrements dans le découpage. Le nombre d'enregistrements impliqués dans ce découpage spécifique.

Nbre de champs utilisés. Classe les modèles découpés en fonction du nombre de champs d'entrée utilisés.

Précision globale (%). Pourcentage d'enregistrements correctement prévus par le modèle découpé par rapport au nombre total d'enregistrements dans ce découpage.

Figure 3-28

Viewer de modèles scindés

Groupes de division			
ed	Précision	Taille du modèle (synapses)	Enregistrements
Did not complete high school	68,3%	62	372
High school degree	66,8%	42	198
Some college	72,3%	42	87
College degree	59,2%	12	38
Post-undergraduate degree	.	.	.

Les modèles ne peuvent pas être créés pour un ou plusieurs groupes de division.

Scission. L'en-tête de la colonne indique le ou les champs utilisés pour créer des scissions, et les cellules correspondent aux valeurs de scission. Double-cliquez sur une scission afin d'ouvrir un viewer de modèle pour le modèle construit pour cette scission.

Précision. Précision générale sous la forme d'un pourcentage.

Taille du modèle. La taille du modèle dépend de la méthode de modélisation : pour les arborescences, il s'agit du nombre de noeuds de l'arborescence ; pour les modèles linéaires, il s'agit du nombre de coefficients ; pour les réseaux neuronaux, il s'agit du nombre des synapses.

Enregistrements. Le nombre pondéré d'enregistrements d'entrée dans l'échantillon d'apprentissage.

Utilisation de nuggets de modèle dans les flux

Les nuggets de modèles sont placés dans des flux afin de vous permettre de déterminer les scores des nouvelles données et de générer de nouveaux noeuds. La détermination des **scores** des données vous permet d'utiliser les informations que vous avez pu rassembler lors de la création de modèles pour générer des prévisions pour de nouveaux enregistrements. Pour consulter les

résultats de scoring, vous devez attacher un noeud terminal (c'est-à-dire à un noeud d'exécution ou de sortie) au nugget et exécuter le noeud terminal.

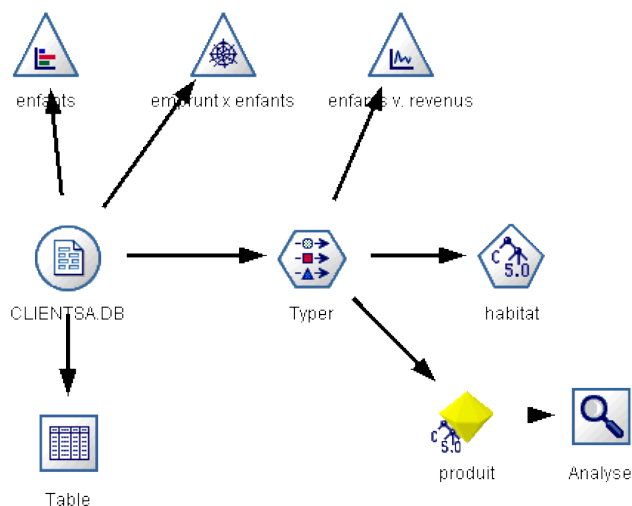
Pour certains modèles, les nuggets de modèle peuvent également vous fournir des informations sur la qualité de la prévision, telles que les degrés de confiance et les distances par rapport au centre des classes. En générant de nouveaux noeuds, vous pouvez facilement créer des noeuds basés sur la structure du modèle généré. Par exemple, la plupart des modèles qui effectuent une sélection des champs d'entrée vous permettent de générer des noeuds Filtrer qui transmettront uniquement les champs d'entrée que le modèle a identifié comme importants.

Pour utiliser un nugget de modèle destiné au scoring des données

- Connectez le nugget de modèle à une source de données ou à un flux qui lui transmettra des données.

Figure 3-29

Utilisation d'un nugget de modèle pour le scoring



- Ajoutez ou connectez un ou plusieurs noeuds de traitement ou de sortie (un noeud Table ou Analyse, par exemple) au nugget de modèle.
- Exécutez l'un des noeuds situés en aval du nugget de modèle.

Remarque : Vous ne pouvez pas utiliser le noeud Règle non affinée pour déterminer les scores des données. Pour déterminer les scores des données en fonction d'un modèle de règle d'association, utilisez le noeud Règle non affinée pour générer un nugget Ensemble de règles, et utilisez ce dernier pour le scoring. [Pour plus d'informations, reportez-vous à la section Génération d'un ensemble de règles à partir d'un nugget de modèle d'association dans le chapitre 12 sur p. 437.](#)

Pour utiliser un nugget de modèle destiné à générer des noeuds de traitement

- Dans la palette, parcourez le modèle, ou éditez ce dernier dans l'espace de travail de flux.
- Sélectionnez le type de noeud voulu dans le menu Générer du navigateur de nugget de modèle. Les options disponibles varient en fonction du type de nugget de modèle. Reportez-vous à la

section consacrée au type de nugget de modèle qui vous intéresse pour savoir ce que ce modèle vous permet de générer.

Régénération d'un noeud de modélisation

Si vous souhaitez modifier ou mettre à jour l'un de vos nuggets de modèle et que le flux utilisé pour créer le modèle n'est pas disponible, vous pouvez régénérer un noeud de modélisation avec les options qui avaient servi à créer le modèle d'origine.

- ▶ Pour recréer un modèle, cliquez avec le bouton droit de la souris sur le modèle avec le bouton droit de la souris dans la palette des modèles, puis sélectionnez Générer le noeud de modélisation.
- ▶ Sinon, lorsque vous parcourez un modèle, choisissez Générer un noeud de modélisation dans le menu Générer.

Le noeud de modélisation régénéré est fonctionnellement identique à celui que vous aviez utilisé pour créer le modèle d'origine, dans la plupart des cas.

- Dans le cas des modèles Arbre décision, des paramètres supplémentaires définis lors de la session interactive peuvent également être stockés avec le noeud. L'option Utiliser les directives d'arbre est activée dans le noeud de modélisation régénéré.
- Dans le cas des modèles Liste de décision, l'option Utiliser les informations de session interactive enregistrées est activée. [Pour plus d'informations, reportez-vous à la section Options du modèle Liste de décision dans le chapitre 9 sur p. 235.](#)
- Dans le cas des modèles de séries temporelles, l'option Poursuivre l'estimation à l'aide du ou des modèles existants est activée, ce qui vous permet de régénérer le modèle précédent avec les données actuelles. [Pour plus d'informations, reportez-vous à la section Options du modèle de séries temporelles dans le chapitre 13 sur p. 469.](#)

Importation et exportation de modèles au format PMML

PMML (Predictive Model Markup Language - Langage de balisage de modèle de prévision) est un format XML permettant de décrire les modèles de Data mining et de statistiques, notamment les entrées des modèles, les transformations utilisées avant la préparation des données pour le Data mining et les paramètres définissant les modèles eux-mêmes. IBM® SPSS® Modeler peut importer et exporter le format PMML, permettant ainsi de partager des modèles avec d'autres applications qui prennent en charge ce format, tel que IBM® SPSS® Statistics.

Remarque : Une licence pour IBM® SPSS® Modeler Server est nécessaire pour exporter au format PMML.

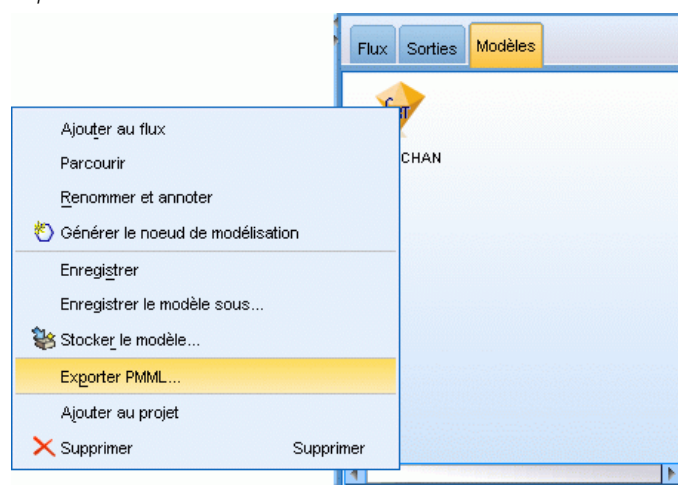
Pour plus d'informations sur le langage PMML, reportez-vous au site Web du Data mining group (<http://www.dmg.org>).

Pour exporter un modèle

L'exportation PMML est prise en charge pour la plupart des types de modèle générés dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Types de modèle prenant en charge le format PMML dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

- ▶ Cliquez avec le bouton droit de la souris sur un nugget de modèle dans la palette des modèles. (Vous pouvez aussi double-cliquer sur un nugget de modèles dans l'espace de travail et sélectionner le menu Fichier.)
- ▶ Dans le menu, cliquez sur Exporter PMML.

Figure 3-30
Exportation d'un modèle au format PMML



- ▶ Dans la boîte de dialogue Exporter (ou Enregistrer), choisissez un répertoire cible et un nom unique pour le modèle.

Remarque : Vous pouvez modifier les options d'exportation PMML dans la boîte de dialogue Options utilisateur. Dans le menu principal, cliquez sur :
Outils > Options > Options utilisateur

et cliquez sur l'onglet PMML.

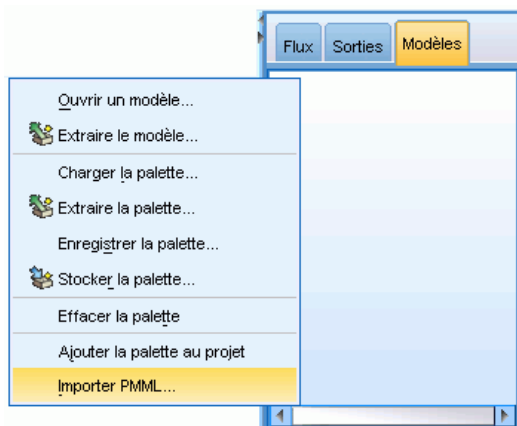
[Pour plus d'informations, reportez-vous à la section Définition des options d'exportation PMML dans le chapitre 12 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Pour importer un modèle enregistré au format PMML

Les modèles exportés au format PMML à partir de SPSS Modeler ou d'une autre application peuvent être importés dans la palette de modèles. [Pour plus d'informations, reportez-vous à la section Types de modèle prenant en charge le format PMML dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

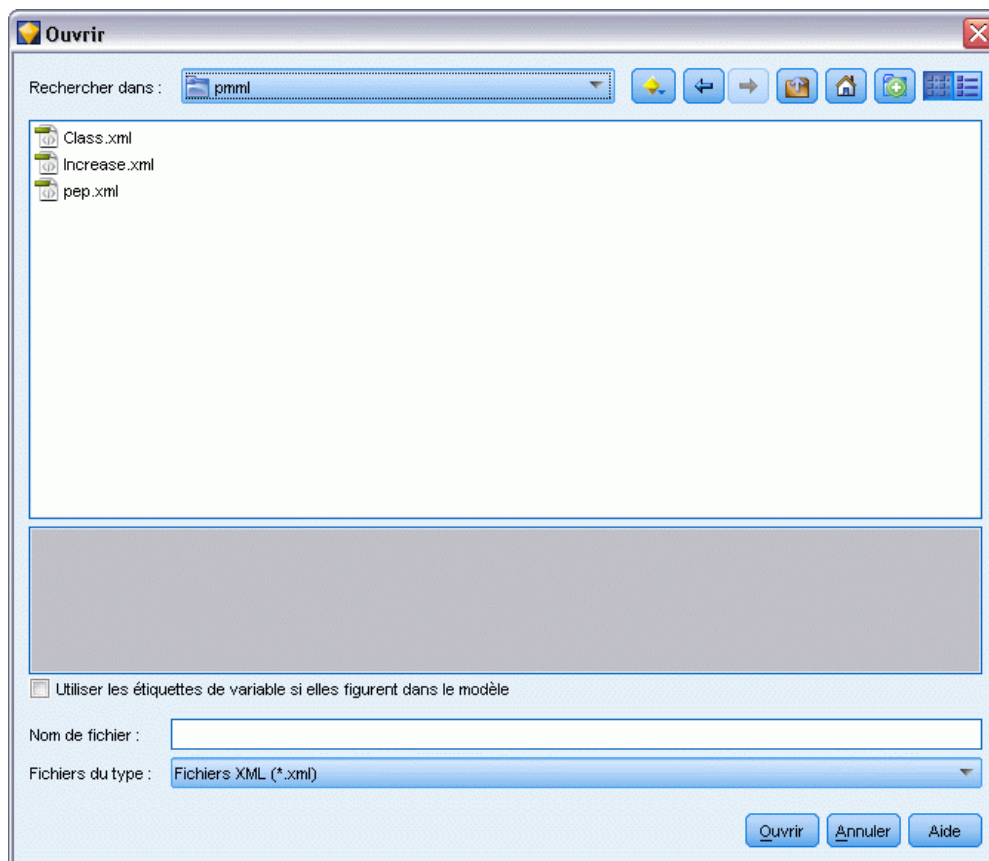
- ▶ Cliquez avec le bouton droit de la souris sur cette palette, puis sélectionnez Importer PMML dans le menu.

Figure 3-31
Importation d'un modèle au format PMML



- ▶ Sélectionnez le fichier à importer et indiquez les options requises pour les étiquettes de variables.
- ▶ Cliquez sur Ouvrir.

Figure 3-32
Sélection du fichier XML pour un modèle enregistré au format PMML



Utilisez des étiquettes de variables si elles sont présentes dans le modèle. Le langage PMML (Predictive Model Markup Language) permet de spécifier les noms et les étiquettes (par exemple, ID Référence pour *IDRéf*) des variables du dictionnaire de données. Sélectionnez cette option pour utiliser des étiquettes de variables si elles apparaissent dans le langage PMML initialement exporté.

Si vous avez sélectionné l'option d'étiquette de variable alors que le langage PMML ne contient pas d'étiquette de variable, le nom des variables est utilisé.

Types de modèle prenant en charge le format PMML

Exportation en PMML

SPSS Modeler modèles. Les modèles suivants créés dans IBM® SPSS® Modeler peuvent être exportés au format PMML 4.0 :

- Arbre C&RT
- QUEST
- CHAID
- Régression linéaire
- R. neurones
- C5.0
- Régression logistique
- Genlin
- SVM
- Bayes Net
- A priori
- Carma
- K-means
- Kohonen
- TwoStep
- KNN
- Modèle Statistics

Le modèle suivant créé dans SPSS Modeler peut être exporté au format PMML 3.2 :

- Liste de décision

Modèles natifs de base de données. Pour les modèles générés à l'aide d'algorithmes natifs de base de données, l'exportation PMML est disponible pour les modèles IBM InfoSphere Warehouse uniquement. Les modèles créés avec Analysis Services de Microsoft ou Oracle Data Miner ne peuvent pas être exportés. Notez également que les modèles IBM exportés au format PMML ne peuvent pas être réimportés dans SPSS Modeler. [Pour plus d'informations, reportez-vous à la section Modélisation de base de données - Présentation dans le chapitre 2 dans *Guide d'exploration de base de données de IBM SPSS Modeler 15*.](#)

Importation de PMML

SPSS Modeler peut importer et évaluer les modèles PMML générés par les versions actuelles de tous les produits IBM® SPSS® Statistics, y compris les modèles exportés depuis SPSS Modeler, et les modèles ou transformations PMML générés par SPSS Statistics 17.0 ou une version ultérieure. En d'autres termes, il s'agit de tout PMML que le moteur d'évaluation peut évaluer, à l'exception des éléments suivants :

- Les modèles Apriori, CARMA, Détection des anomalies et Séquence ne peuvent pas être importés.
- Les modèles PMML ne seront peut-être pas parcourus après importation dans SPSS Modeler, même s'ils peuvent être utilisés pour le scoring. (Notez que cela inclut les modèles qui ont été exportés à l'origine de SPSS Modeler. Pour contourner cet obstacle, exportez le modèle au format de fichier de modèle généré [*.gm] plutôt qu'au format PMML.)
- Les modèles IBM InfoSphere Warehouse exportés au format PMML ne peuvent pas être importés.
- La validation limitée se produit lors de l'importation, mais la validation complète s'effectue lors de l'évaluation du modèle. Par conséquent, il est possible que l'importation réussisse mais que l'évaluation échoue ou produise des résultats incorrects.

Publication des modèles pour un adaptateur de scoring

Vous pouvez publier des modèles dans un serveur de base de données contenant un adaptateur de scoring. Un adaptateur de scoring permet d'exécuter le scoring du modèle dans la base de données en utilisant les fonctionnalités de fonction définie par l'utilisateur (UDF) de la base de données. Exécuter le scoring dans la base de données permet de ne pas avoir besoin d'extraire les données avant le scoring. La publication d'un adaptateur de scoring génère également un exemple de SQL pour exécuter l'UDF.

Pour publier dans un adaptateur de scoring

- ▶ Faites un double clic sur le nugget de modèle pour l'ouvrir.
- ▶ Depuis le menu du nugget de modèle, choisissez :
Fichier > Publier pour l'adaptateur de scoring de serveur
- ▶ Remplissez les champs appropriées de la boîte de dialogue et cliquez sur OK.

Connexion à la base de données. Les informations de connexion de la base de données à utiliser pour le modèle.

ID de publication. (DB2 pour bases de données z/OS uniquement) Un identifiant du modèle. Si vous recréez le même modèle et que vous utilisez le même ID de publication, le SQL généré reste le même, par conséquent, il est possible de recréer un modèle sans avoir à modifier l'application qui utilise le SQL généré précédemment. (Pour les autres bases de données, le SQL généré est unique au modèle).

Générer un exemple de SQL. Si cette option est sélectionnée, l'exemple de SQL est généré dans le fichier spécifié dans le champ Fichier.

Modèles bruts

Un modèle brut comporte des informations extraites des données, mais n'est pas conçu pour générer des prévisions de manière directe. Cela signifie qu'il ne peut pas être ajouté à des flux. Les modèles bruts apparaissent sous forme de "diamants à l'état brut" dans la palette des modèles générés.

Figure 3-33
Icône de modèle brut



Pour obtenir des informations sur le modèle de règle non affinée, cliquez avec le bouton droit de la souris sur le modèle et choisissez *Parcourir* dans le menu contextuel. Comme les autres modèles générés dans IBM® SPSS® Modeler, les divers onglets fournissent des informations récapitulatives et de règle sur le modèle créé.

Génération de noeuds. Le menu *Générer* vous permet de créer des noeuds sur la base de règles.

- **Noeud Sélectionner.** Génère un noeud *Sélectionner* pour sélectionner les enregistrements auxquels la règle sélectionnée s'applique. Lorsqu'aucune règle n'est sélectionnée, cette option est désactivée.
- **Ensemble de règles.** Génère un noeud *Ensemble de règles* pour prévoir les valeurs d'un seul champ cible. [Pour plus d'informations, reportez-vous à la section Génération d'un ensemble de règles à partir d'un nugget de modèle d'association dans le chapitre 12 sur p. 437.](#)

Modèles de filtrage

Filtrage des champs et des enregistrements

Plusieurs noeuds de modélisation peuvent être utilisés au cours des phases préliminaires d'une analyse ; ils permettent en effet de rechercher les champs et les enregistrements les plus intéressants pour une modélisation. Vous pouvez utiliser le noeud Sélection de fonction pour filtrer et classer les champs par ordre d'importance. Le noeud Détection des anomalies permet, quant à lui, de repérer les enregistrements inhabituels qui ne se conforment pas aux patrons de données « normales » connus.



Le noeud Sélection de fonction filtre les champs d'entrée en vue de leur suppression, en fonction d'un ensemble de critères donné (tel que le pourcentage de valeurs manquantes) ; il classe ensuite les entrées restantes selon leur importance par rapport à la cible indiquée. Si l'on prend, par exemple, un ensemble de données comportant des centaines d'entrées potentielles, quelles sont celles susceptibles d'être les plus utiles dans la modélisation des résultats de patients ? [Pour plus d'informations, reportez-vous à la section Noeud Sélection de fonction sur p. 78.](#)



Le noeud Détection des anomalies identifie les observations inhabituelles, ou valeurs éloignées, qui ne se conforment pas aux patrons de données « normales ». Il vous permet d'identifier les valeurs éloignées même si celles-ci ne correspondent pas aux patrons connus précédemment et même si vous ne savez pas exactement ce que vous recherchez. [Pour plus d'informations, reportez-vous à la section Noeud Détection des anomalies sur p. 86.](#)

La détection d'anomalies identifie les observations ou enregistrements inhabituels par le biais d'une analyse des classes appliquée à l'ensemble de champs sélectionné dans le modèle, et ce, sans prendre en compte aucun champ (dépendant) cible spécifique, que ces champs soient pertinents ou non pour le patron sur lequel porte la prévision. C'est pourquoi il peut paraître utile d'associer la détection d'anomalies à la sélection de fonction ou à toute autre technique permettant de filtrer et de classer les champs. Vous pouvez, par exemple, utiliser la sélection de fonction pour identifier les champs les plus importants pour une cible donnée, puis exécuter la détection d'anomalies afin de repérer les enregistrements les plus inhabituels concernant ces champs. (Une autre approche consiste à créer un modèle d'arbre décision, puis à examiner tous les enregistrements qui n'ont pas été correctement classés afin de détecter des anomalies potentielles. Toutefois, cette méthode est plus difficile à répliquer et à automatiser à grande échelle.)

Noeud Sélection de fonction

L'un des problèmes du Data mining réside dans le fait que des centaines, voire des milliers de champs peuvent servir de champs d'entrée. Vous pouvez ainsi passer beaucoup de temps à déterminer les champs ou variables à inclure dans le modèle. Il est possible de réduire le choix grâce à l'algorithme Sélection de fonction qui permet d'identifier les champs les plus importants pour une analyse donnée. Par exemple, si vous essayez de prévoir les résultats relatifs à des

patients en fonction d'un certain nombre de facteurs, il convient de répondre à une question : quels facteurs semblent être les plus importants ?

La sélection de fonction se déroule en trois étapes :

- **Filtrage.** Supprime les entrées (et les enregistrements ou observations) non significatives et problématiques, telles que les champs d'entrée comportant un trop grand nombre de valeurs manquantes ou présentant une variation trop ou pas assez importante pour être utiles.
- **Rang :** Trie les entrées restantes et leur affecte un rang en fonction de leur importance.
- **Sélection.** Identifie le sous-ensemble de fonctions à utiliser dans les modèles suivants (en ne conservant, par exemple, que les entrées les plus importantes, et en filtrant ou en excluant toutes les autres).

Alors que bon nombre d'entreprises croulent aujourd'hui sous une quantité excessive de données, la sélection de fonction, en simplifiant et en accélérant le processus de modélisation, peut s'avérer très bénéfique. Concentrez-vous rapidement sur les champs présentant le plus grand intérêt, et réduisez ainsi le nombre de calculs requis ; repérez plus facilement des relations ténues mais importantes qui auraient sinon pu être ignorées et, enfin, obtenez des modèles plus simples, plus précis et plus facilement explicables. En restreignant le nombre de champs utilisés dans le modèle, vous pouvez réduire le nombre de scorings, ainsi que la quantité de données collectées dans les futures itérations.

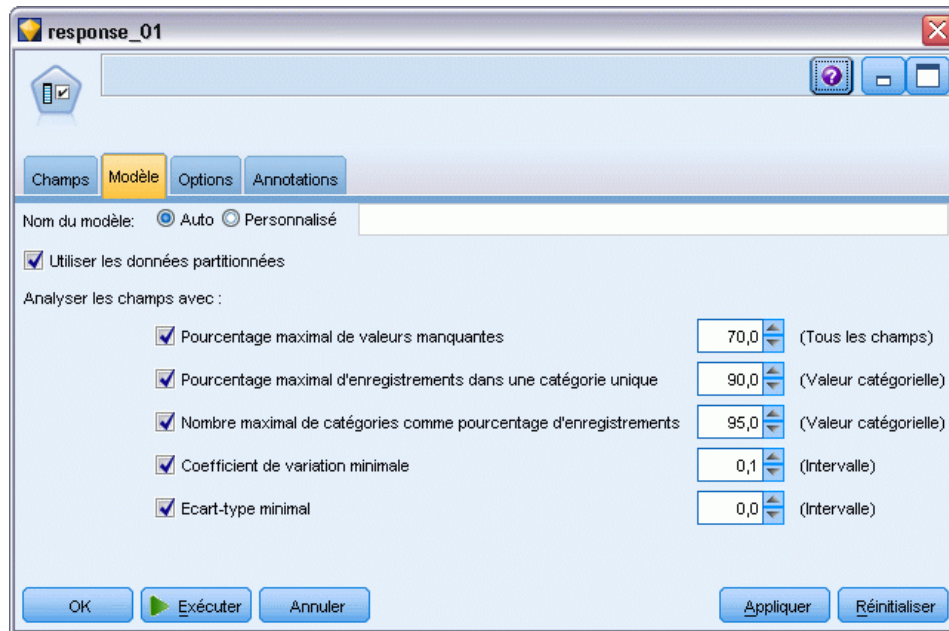
Exemple : Un opérateur de téléphonie possède un entrepôt de données contenant des informations concernant les réponses données par 5 000 clients de l'entreprise à une promotion spéciale. Ces données incluent un grand nombre de champs comprenant l'âge, la profession et les revenus des clients, ainsi que les statistiques d'utilisation de leur téléphone. Trois champs cible indiquent si le client a répondu à chacune des trois offres. L'opérateur souhaite utiliser ces données pour connaître les clients les plus susceptibles de répondre à des offres similaires à l'avenir. [Pour plus d'informations, reportez-vous à la section Filtrage des variables indépendantes \(sélection de fonction\) dans le chapitre 9 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Un champ de cible unique (dont le rôle est configuré sur *Cible*), avec plusieurs champs d'entrée que vous souhaitez filtrer ou classer par rapport à la cible. Les champs cible et d'entrée ont un niveau de mesure *Continu* (intervalle numérique) ou *Catégoriel*.

Paramètres des modèles Sélection de fonction

Les paramètres affichés dans l'onglet *Modèle* incluent des options de modèle standard, ainsi que des paramètres qui vous permettent d'affiner les critères de filtrage des champs d'entrée.

Figure 4-1
Onglet *Modèle* Sélection de fonction



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.

Filtrage des champs d'entrée

Le filtrage procède à la suppression des entrées ou des observations qui n'ajoutent aucune information utile quant à la relation entrée/cible. Les options de filtrage reposent sur les attributs du champ concerné et ne tiennent pas compte de la puissance de prévision relative au champ cible sélectionné. Les champs filtrés sont exclus des calculs utilisés pour classer les entrées ; ils peuvent également être filtrés ou supprimés dans les données servant à la modélisation.

Le filtrage des champs peut reposer sur les critères suivants :

- **Pourcentage maximal de valeurs manquantes.** Filtre les champs comportant trop de valeurs manquantes, exprimées sous forme d'un pourcentage du nombre total d'enregistrements. Les champs présentant un fort pourcentage de valeurs manquantes n'offrent que peu d'informations prévisionnelles.
- **Pourcentage maximal d'enregistrements dans une catégorie unique.** Filtre les champs qui comportent, par rapport au nombre total d'enregistrements, trop d'enregistrements relatifs à une même catégorie. Supposons, par exemple, que 95 % des clients de la base de données conduisent le même type de voiture ; cette information n'est pas pertinente pour différencier un client d'un autre. Tous les champs qui dépassent le nombre maximal indiqué sont filtrés. Cette option s'applique aux champs catégoriels uniquement.
- **Nombre maximal de catégories comme pourcentage d'enregistrements.** Filtre les champs comportant trop de catégories par rapport au nombre total d'enregistrements. Si un pourcentage élevé des catégories ne contient qu'une seule observation, le champ peut

présenter un intérêt limité. Supposons, par exemple, que chaque client porte un chapeau différent ; cette information a peu de chance d'être utile pour la modélisation de patrons de comportements. Cette option s'applique aux champs catégoriels uniquement.

- **Coefficient de variation minimal.** Filtre les champs dont le coefficient de variance est inférieur ou égal à la valeur minimale indiquée. Cette mesure correspond au rapport entre l'écart type et la moyenne du champ d'entrée. Une valeur proche de zéro indique une faible variabilité entre les valeurs de la variable. Cette option s'applique aux champs continus (intervalle numérique) uniquement.
- **Ecart-type minimal.** Filtre les champs dont l'écart-type est inférieur ou égal à la valeur minimale indiquée. Cette option s'applique aux champs continus (intervalle numérique) uniquement.

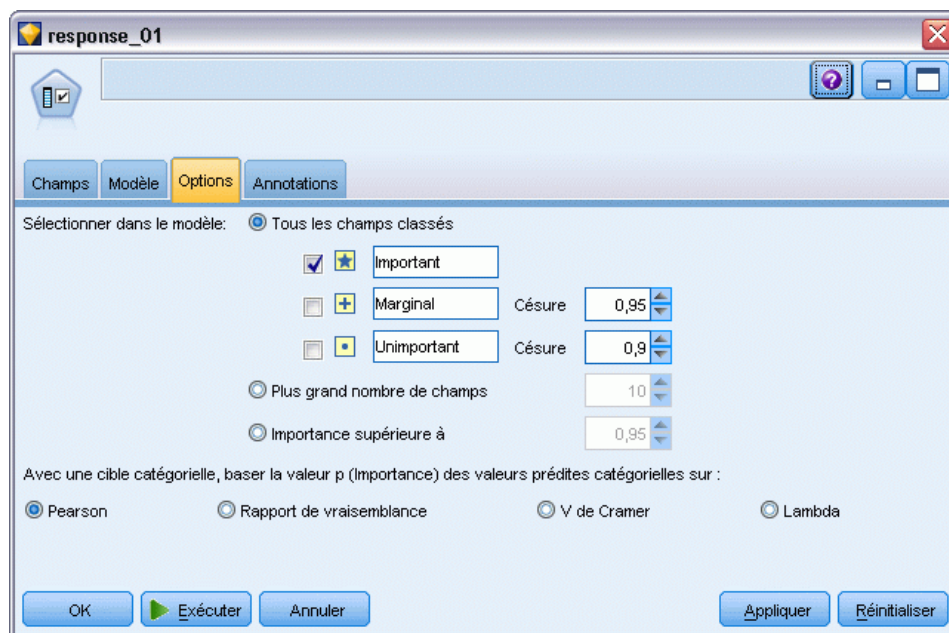
Enregistrements contenant des données manquantes. Les enregistrements ou observations présentant des valeurs manquantes pour le champ cible ou pour l'ensemble des entrées sont exclus automatiquement de tous les calculs utilisés dans les classements.

Options de sélection des caractéristiques

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.

Pour plus d'informations, reportez-vous à la section Résultats du modèle Sélection de fonction sur p. 83.

Figure 4-2
Onglet Options de la sélection de fonction



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.

Options de classement en fonction de l'importance

Toutes de type catégoriel. Lorsque toutes les entrées et la cible sont de type catégoriel, l'importance peut être déterminée en fonction de l'une des quatre mesures suivantes :

- **Pearson du Chi-deux.** Tests d'indépendance de la cible et de l'entrée, sans indication de la force et de la direction des relations existantes.
- **Chi-deux de rapport de vraisemblance.** Semblable aux statistiques Pearson du Chi-deux, mais teste également l'indépendance cible/entrée.

- **V de Cramer.** Mesure d'association basée sur les statistiques Pearson du Chi-deux. Les valeurs sont comprises entre 0 (ce qui signifie qu'aucune association n'est présente) et 1 (ce qui désigne une association parfaite).
- **Lambda.** Mesure d'association reflétant la réduction proportionnelle d'erreurs lorsque la variable est utilisée pour la prévision de la valeur cible. Une valeur de 1 indique que le champ d'entrée offre une prévision parfaite de la cible ; la valeur 0 signifie, quant à elle, que l'entrée ne fournit aucune information utile sur la cible.

Certaines de type catégoriel. Lorsque certaines entrées, mais pas toutes, sont de type catégoriel et que la cible est du même type, l'importance peut être déterminée par le biais du Pearson du Chi-deux ou du Chi-deux de rapport de vraisemblance. (Les mesures V de Cramer et lambda ne sont disponibles que si les toutes les entrées sont de type catégoriel.)

Type catégoriel/Type continu. Lors du classement d'une entrée de type catégoriel par rapport à une cible continue, ou inversement (l'une ou l'autre est catégorielle mais pas les deux), les statistiques F sont utilisées.

Type continu uniquement. Lors du classement d'une entrée continue par rapport à une cible continue, les statistiques T basées sur le coefficient de corrélation sont utilisées.

Nuggets de modèles Sélection de fonction

Les nuggets de modèles Sélection de fonction affichent l'importance de chaque entrée pour une cible sélectionnée, en fonction du classement du noeud Sélection de fonction. Les champs ayant été filtrés avant la réalisation du classement sont également répertoriés. [Pour plus d'informations, reportez-vous à la section Noeud Sélection de fonction sur p. 78.](#)

Lorsque vous exécutez un flux contenant un nugget de modèles Sélection de fonction, ce dernier agit comme un filtre qui ne conserve que les entrées sélectionnées, en fonction de la sélection effectuée dans l'onglet Modèle. Par exemple, vous pouvez sélectionner tous les champs classés comme importants (l'une des options par défaut) ou sélectionner manuellement un sous-ensemble de champs dans l'onglet Modèle. Le champ cible est également conservé quelle que soit la sélection. Tous les autres champs sont exclus.

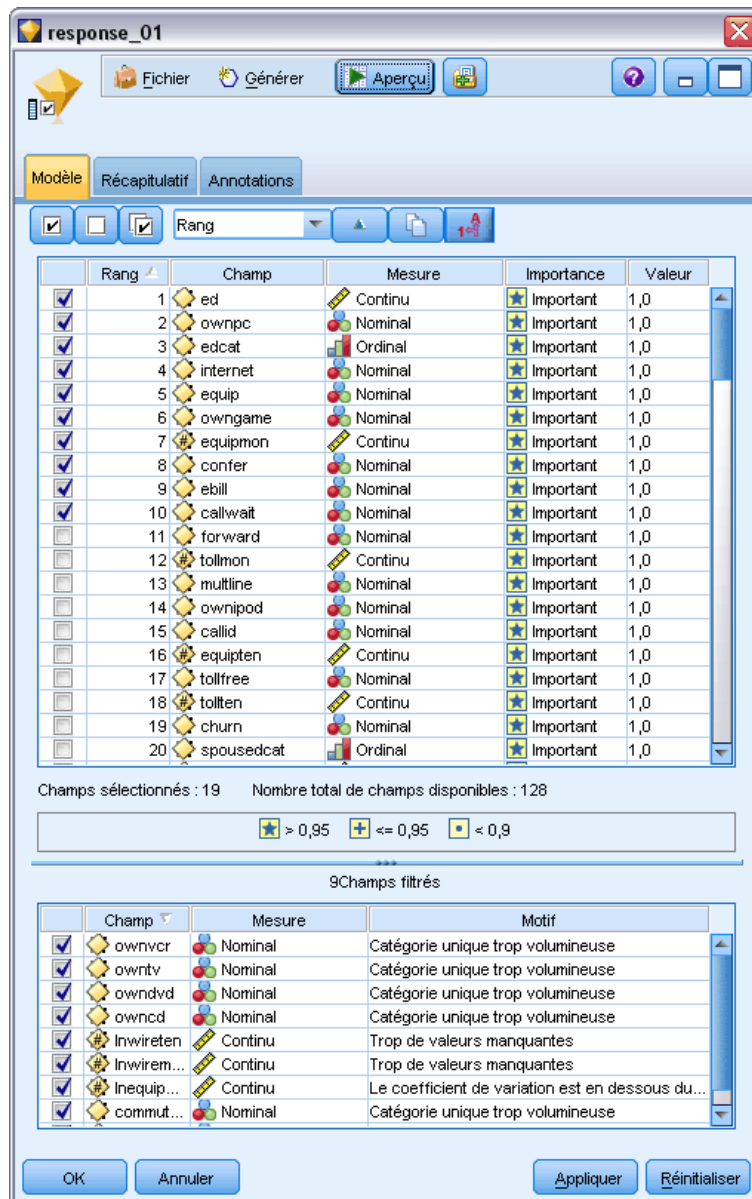
Le filtrage est basé sur les noms de champs uniquement ; par exemple, si vous sélectionnez *âge* et *revenu*, tout champ portant l'un de ces deux noms est conservé. Le modèle ne met pas à jour les classements de champs en fonction des nouvelles données ; il se contente de filtrer les champs sur la base des noms sélectionnés. Par conséquent, soyez vigilant lorsque vous appliquez le modèle à des données nouvelles ou mises à jour. Si vous avez des doutes, il est recommandé de régénérer le modèle.

Résultats du modèle Sélection de fonction

L'onglet Modèle correspondant à un nugget de modèles Sélection de fonction affiche le classement et l'importance de toutes les entrées dans le panneau supérieur ; il 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. Les autres champs sont ignorés. Les sélections par défaut dépendent des options indiquées dans le noeud de création de modèle, mais vous pouvez sélectionner ou désélectionner d'autres champs, comme souhaité.

Le panneau inférieur répertorie les entrées qui ont été exclues des classements en fonction du pourcentage de valeurs manquantes ou d'autres critères définis dans le noeud de modélisation. Tout comme dans le cas des champs classés, vous pouvez décider d'inclure ou d'exclure ces champs en cochant les cases de la colonne de gauche. [Pour plus d'informations, reportez-vous à la section Paramètres des modèles Sélection de fonction sur p. 79.](#)

Figure 4-3
Résultats du modèle Sélection de fonction



- 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 utiliser la barre d'outils pour sélectionner l'élément souhaité dans la liste 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 étendre la sélection en maintenant les touches Maj et Ctrl enfoncées et en cliquant sur les champs souhaités, et utiliser la barre d'espacement pour activer ou désactiver un groupe de fichiers sélectionnés. [Pour plus d'informations, reportez-vous à la section Sélection des champs en fonction de leur importance sur p. 85.](#)
- 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. [Pour plus d'informations, reportez-vous à la section Options de sélection des caractéristiques sur p. 81.](#)

Sélection des champs en fonction de leur importance

Lors du scoring de données à l'aide d'un nugget de modèles Sélection de fonction, tous les champs sélectionnés dans la liste des champs classés ou filtrés (c'est-à-dire comportant une coche dans la colonne de gauche) sont conservés. Les autres champs sont ignorés. Pour modifier la sélection, utilisez la barre d'outils pour accéder à la boîte de dialogue Sélectionner les champs, et sélectionnez les champs par rang ou importance.

Figure 4-4
Boîte de dialogue Sélectionner les champs



Tous les champs marqués. Sélectionne tous les champs indiqués comme étant importants, marginaux ou non significatifs.

Plus grand nombre de champs. Permet de sélectionner les n premiers champs en termes d'importance.

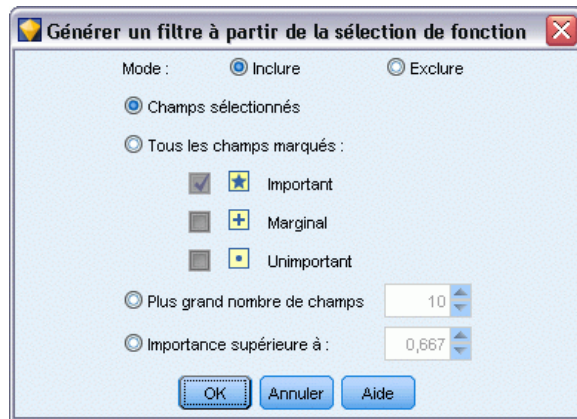
Importance supérieure à. Sélectionne tous les champs présentant une importance supérieure au seuil indiqué.

Génération d'un filtre à partir d'un modèle Sélection de fonction

A partir des résultats d'un modèle Sélection de fonction, vous pouvez générer un ou plusieurs noeuds Filtrer qui incluent ou excluent des sous-ensembles de champs en fonction de leur importance par rapport à la cible indiquée. Alors que le nugget de modèle peut également servir de filtre, cette opération vous permet de bénéficier de la souplesse nécessaire pour tester différents

sous-ensembles de champs sans avoir à copier ni modifier le modèle. Le champ cible reste conservé par le filtre, que vous ayez choisi l'opération d'inclusion ou d'exclusion.

Figure 4-5
Génération d'un noeud Filtrer



Inclusion/Exclusion. Vous pouvez choisir d'inclure ou d'exclure des champs ; il est ainsi possible d'inclure les 10 premiers champs ou d'exclure tous les champs indiqués comme étant non significatifs.

Champs sélectionnés. Inclut ou exclut tous les champs actuellement sélectionnés dans le tableau.

Tous les champs marqués. Sélectionne tous les champs indiqués comme étant importants, marginaux ou non significatifs.

Plus grand nombre de champs. Permet de sélectionner les n premiers champs en termes d'importance.

Importance supérieure à. Sélectionne tous les champs présentant une importance supérieure au seuil indiqué.

Noeud Détection des anomalies

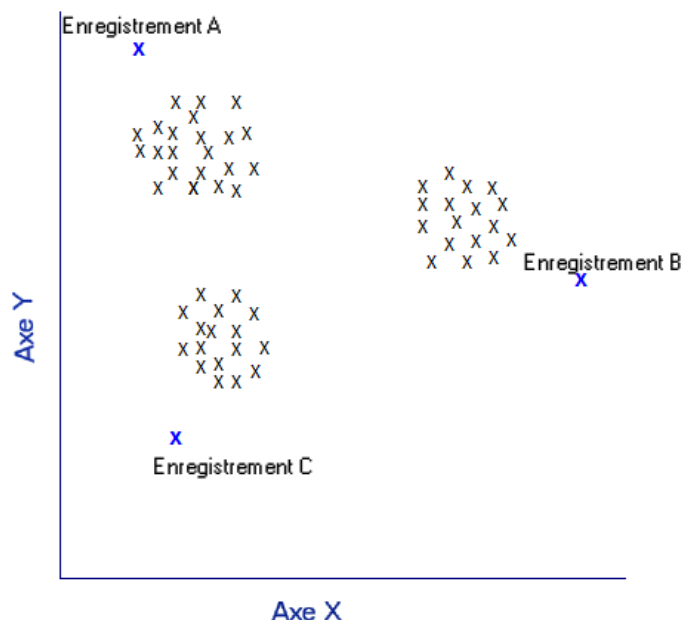
Les modèles de détection des anomalies permettent d'identifier les valeurs éloignées (ou les cas inhabituels) au sein des données. Contrairement aux méthodes de modélisation qui contiennent des règles portant sur les observations inhabituelles, les modèles de détection des anomalies stockent des informations relatives à ce que doit être un comportement normal. Ils permettent ainsi d'identifier les valeurs éloignées, et ce, même si ces dernières ne correspondent pas à un patron connu ; ils trouvent toute leur utilité dans la détection de fraudes où de nouveaux patrons apparaissent sans cesse. La détection d'anomalies est une méthode non supervisée, ce qui signifie qu'elle ne requiert comme point de départ aucun ensemble de données d'apprentissage contenant des cas de fraude connus.

Si les méthodes traditionnelles d'identification des valeurs éloignées prennent en compte une ou deux variables à la fois, la détection d'anomalies peut analyser un grand nombre de champs afin d'identifier les classes ou les groupes d'homologues contenant des enregistrements semblables. Chaque enregistrement peut ensuite être comparé aux autres enregistrements du groupe d'homologues en vue de déceler des anomalies possibles. Plus une observation s'éloigne

du point central défini comme étant normal, plus il y a de chances qu'elle soit inhabituelle. Par exemple, l'algorithme peut répartir les enregistrements dans trois classes distinctes et signaler ceux qui se trouvent les plus éloignés du centre de chaque classe.

Figure 4-6

Utilisation des classifications non supervisées pour identifier des anomalies potentielles



Chaque enregistrement se voit affecter un index d'anomalies, qui correspond au rapport entre l'index d'écart du groupe et la moyenne de la classe à laquelle l'observation appartient. Plus la valeur de l'index est élevée, plus l'observation s'écarte de la moyenne. En règle générale, les observations comportant des valeurs d'index d'anomalies inférieures à 1, voire à 1,5, ne sont pas considérées comme des anomalies car l'écart est quasi identique (ou légèrement supérieur) à la moyenne. Toutefois, les observations dont la valeur d'index est supérieure à 2 ont de fortes chances d'être anormales puisque l'écart correspond à plus du double de la moyenne.

La détection d'anomalies est une méthode exploratoire conçue pour détecter rapidement les observations ou enregistrements inhabituels qui doivent faire l'objet d'une analyse plus poussée. Il convient de parler ici d'anomalies *suspectées* ; en effet, à la suite d'un examen approfondi, elles peuvent s'avérer réelles ou non. Il se peut qu'un enregistrement soit parfaitement valide mais que vous décidiez de le filtrer à partir des données, à des fins de création de modèle. En outre, si l'algorithme renvoie à plusieurs reprises de fausses anomalies, cela peut être le résultat d'une erreur ou d'un artefact dans le processus de collecte des données.

La détection d'anomalies identifie les observations ou enregistrements inhabituels par le biais d'une analyse des classes appliquée à l'ensemble de champs sélectionné dans le modèle, et ce, sans prendre en compte aucun champ (dépendant) cible spécifique, que ces champs soient pertinents ou non pour le patron sur lequel porte la prévision. C'est pourquoi il peut paraître utile d'associer la détection d'anomalies à la sélection de fonction ou à toute autre technique permettant de filtrer et de classer les champs. Vous pouvez, par exemple, utiliser la sélection de fonction pour identifier les champs les plus importants pour une cible donnée, puis exécuter la détection d'anomalies afin de repérer les enregistrements les plus inhabituels concernant ces champs. (Une autre approche

consiste à créer un modèle d'arbre décision, puis à examiner tous les enregistrements qui n'ont pas été correctement classés afin de détecter des anomalies potentielles. Toutefois, cette méthode est plus difficile à répliquer et à automatiser à grande échelle.)

Exemple : En filtrant les subventions pour le développement agricole dans le but d'y détecter d'éventuels cas de fraude, la détection des anomalies peut être utilisée pour découvrir les écarts par rapport à la norme, en mettant en relief les enregistrements non conformes et qui méritent des recherches supplémentaires. Vous vous intéressez en particulier aux demandes de subvention qui semblent requérir une somme trop élevée (ou trop faible) pour le type et la taille de la ferme concernée.

Conditions requises. Au moins un champ d'entrée. Seuls les champs dont le rôle est paramétré sur Entrée et utilisant un noeud source ou Typer peuvent être utilisés comme entrées. Les champs cible (dont le rôle se voit affecter la valeur Cible ou Les deux) sont ignorés.

Puissance. En signalant les observations *non* conformes à un ensemble de règles plutôt que celles conformes, les modèles Détection des anomalies permettent d'identifier les observations inhabituelles, et ce même si celles-ci n'obéissent pas à des patrons précédents connus. Associée à la sélection de fonction, la détection d'anomalies permet de filtrer de grandes quantités de données afin d'identifier relativement rapidement les enregistrements présentant le plus grand intérêt.

Options des modèles Détection des anomalies

Figure 4-7
Onglet du modèle Détection des anomalies

The screenshot shows the 'Anomaly' configuration window. The 'Modèle' tab is active. The model name is 'Anomaly'. The 'Personnalisé' radio button is selected. The checkbox 'Utiliser les données partitionnées' is checked. The 'Déterminer la valeur de césure de l'anomalie en fonction de:' section has three options: 'Niveau d'index minimal des anomalies' (2,0), 'Pourcentage des enregistrements les plus irréguliers des données d'apprentissage' (5,0), and 'Nombre des enregistrements les plus irréguliers des données d'apprentissage' (10). The 'Nombre des enregistrements les plus irréguliers des données d'apprentissage' option is selected. The 'Nombre de champs d'anomalies à reporter:' field is set to 3. The bottom buttons are 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser'.

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.

Déterminer la valeur de césure de l'anomalie en fonction de. Indique la méthode utilisée pour déterminer la valeur de césure associée au signalement des anomalies. Les options suivantes sont disponibles :

- **Niveau d'index minimal des anomalies.** Indique la valeur de césure limite pour le signalement des anomalies. Les enregistrements qui atteignent ou dépassent ce seuil sont signalés.
- **Pourcentage des enregistrements les plus irréguliers des données d'apprentissage.** Définit automatiquement le seuil sur un niveau permettant de signaler le pourcentage d'enregistrements indiqué dans les données d'apprentissage. La césure ainsi déterminée est incluse en tant que paramètre dans le modèle. Cette option permet de définir la valeur de césure et *non* le pourcentage réel d'enregistrements à signaler au cours du scoring. Les résultats de scoring réels peuvent varier en fonction des données.
- **Nombre des enregistrements les plus irréguliers des données d'apprentissage.** Définit automatiquement le seuil sur un niveau permettant de signaler le nombre d'enregistrements indiqué dans les données d'apprentissage. Le seuil ainsi déterminé est inclus en tant que paramètre dans le modèle. Cette option permet de définir la valeur de césure et *non* le nombre d'enregistrements donné à signaler au cours du scoring. Les résultats de scoring réels peuvent varier en fonction des données.

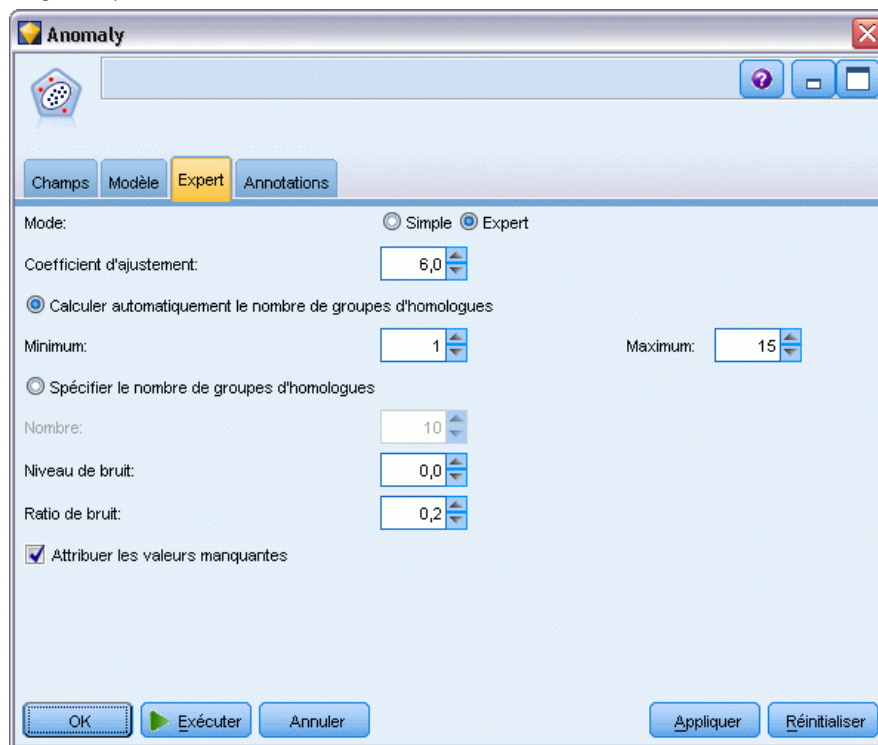
Remarque : Quel que soit son mode de détermination, la valeur de césure n'a aucune incidence sur la valeur d'index d'anomalies sous-jacente indiquée pour chaque enregistrement. Elle définit simplement le seuil de signalement des enregistrements jugés comme irréguliers au cours de l'estimation ou du scoring du modèle. Si, par la suite, vous souhaitez étudier un nombre plus grand ou plus petit d'enregistrements, utilisez un noeud Sélectionner pour définir un sous-ensemble d'enregistrements sur la base de la valeur de l'index des anomalies ($\$O\text{-AnomalyIndex} > X$).

Nombre de champs d'anomalies à reporter. Définit le nombre de champs à signaler, afin d'indiquer la raison pour laquelle un enregistrement donné est marqué comme étant anormal. Les champs les plus irréguliers sont indiqués et définis comme étant ceux présentant le plus grand écart par rapport à la valeur « normale » de champ de la classe à laquelle est affectée l'enregistrement.

Options expert de détection d'anomalies

Afin d'indiquer des options pour les valeurs manquantes, ainsi que pour d'autres paramètres, définissez le mode Expert dans l'onglet Expert.

Figure 4-8
Onglet Expert de la détection des anomalies



Coefficient d'ajustement. Valeur utilisée pour équilibrer la pondération relative attribuée aux champs continus (intervalle numérique) et aux champs catégoriels pour les calculs de distance. Les valeurs élevées augmentent l'influence des champs continus. Il doit s'agir d'une valeur différente de zéro.

Calculer automatiquement le nombre de groupes d'homologues. La détection d'anomalies peut analyser rapidement un grand nombre de solutions possibles afin de déterminer le nombre optimal de groupes d'homologues pour les données d'apprentissage. Vous pouvez élargir ou réduire l'intervalle en définissant un nombre minimal et maximal de groupes d'homologues. Les valeurs élevées permettent au système d'explorer un plus grand intervalle de solutions possibles, mais induisent un temps de traitement plus long.

Spécifier le nombre de groupes d'homologues. Si vous connaissez le nombre de classes à inclure dans votre modèle, sélectionnez cette option et entrez le nombre de groupes d'homologues. La sélection de cette option améliore généralement les performances du système.

Niveau de bruit et Ratio de bruit. Ces paramètres déterminent la façon dont les valeurs éloignées sont traitées lors d'une classification non supervisée en deux étapes. Lors de la première étape, un arbre des fonctions de classe (CF) est utilisé pour condenser les données d'un très grand nombre d'enregistrements individuels en un nombre de classes plus facile à gérer. L'arbre est créé à partir de mesures de similarité et lorsqu'un noeud de l'arbre contient trop d'enregistrements, il se divise en noeuds enfant. Lors de la seconde étape, la classification non supervisée hiérarchique commence sur les noeuds terminaux de l'arbre CF. La gestion du bruit est activée lors du premier

passage de données et désactivée lors du second. Les observations figurant dans la classe de bruit issue du premier passage de données sont affectées aux classes standard lors du second passage.

- **Niveau de bruit.** Indiquez une valeur comprise entre 0 et 0,5. Ce paramètre n'est pertinent que si l'arbre CF se remplit lors de l'étape de développement ; cela signifie qu'il ne peut accepter aucune autre observation dans un noeud feuille et qu'aucun noeud feuille ne peut être divisé.

Si l'arbre CF se remplit et que le niveau de bruit est paramétré sur 0, le seuil est augmenté et l'arbre CF redéveloppé à partir de toutes les observations. Après la classification finale, les valeurs ne pouvant être affectées à une classe deviennent des valeurs éloignées étiquetées.

La classe des valeurs éloignées se voit attribuer le numéro d'identification -1. Elle n'est pas incluse dans le décompte du nombre de classes ; autrement dit, si vous définissez n classes et la gestion du bruit, l'algorithme génère les n classes, plus une classe de bruit. En pratique, vous pouvez augmenter cette valeur afin de donner à l'algorithme davantage de latitude pour intégrer des enregistrements inhabituels dans l'arbre plutôt que de les affecter à une classe de valeurs éloignées distincte.

Si l'arbre CF se remplit et que le niveau de bruit est supérieur à 0, l'arbre CF se redéveloppe après avoir déplacé les données des feuilles sporadiques vers leur propre feuille de bruit. Une feuille est considérée comme étant « sporadique » si le rapport entre le nombre d'observations contenues dans la feuille sporadique et le nombre d'observations contenues dans la plus grande feuille est inférieur au niveau de bruit. Une fois l'arbre redéveloppé, les valeurs éloignées sont placées dans l'arbre CF lorsque cela s'avère possible. Sinon, elles sont ignorées pour la seconde étape de classification non supervisée.

- **Ratio de bruit.** Indique la quantité de mémoire allouée au composant qui doit être utilisée pour la mise en mémoire tampon du bruit. Cette valeur est comprise entre 0,0 et 0,5. Si l'insertion d'une observation particulière dans une feuille de l'arbre entraîne un manque de place inférieur au seuil, la feuille n'est pas divisée. Si le manque de place dépasse le seuil, la feuille est divisée et une autre classe de petite taille est ajoutée à l'arbre CF. D'un point de vue pratique, si vous augmentez la valeur de ce paramètre, l'algorithme peut s'orienter plus rapidement vers un arbre plus simple.

Inclure les valeurs manquantes. Pour les champs continus, remplace toute valeur manquante par la valeur moyenne des champs. Pour les champs catégoriels, les catégories manquantes sont combinées et traitées comme une catégorie valide. Si cette option est désélectionnée, tous les enregistrements comportant des valeurs manquantes sont exclus de l'analyse.

Nuggets de modèles Détection des anomalies

Les nuggets de modèles Détection des anomalies contiennent toutes les informations rassemblées par le modèle Détection des anomalies, ainsi que des informations sur les données d'apprentissage et le processus d'estimation.

Lorsque vous exécutez un flux contenant un nugget de modèles Détection des anomalies, un certain nombre de nouveaux champs est ajouté au flux, en fonction de la sélection effectuée dans l'onglet Paramètres du nugget de modèle. [Pour plus d'informations, reportez-vous à la section](#)

Paramètres des modèles Détection des anomalies sur p. 94. Le nom des nouveaux champs est composé du nom du modèle, précédé de $\$O$, comme le récapitule le tableau suivant :

$\$O$ -Anomaly	Champ booléen indiquant si l'enregistrement est irrégulier.
$\$O$ -AnomalyIndex	Valeur d'index d'anomalies de l'enregistrement.
$\$O$ -PeerGroup	Indique le groupe d'homologues auquel l'enregistrement est affecté.
$\$O$ -Field- n	Nom des n champs les plus irréguliers en termes d'écart par rapport à la norme de la classe.
$\$O$ -FieldImpact- n	Index d'écart de variable du champ. Cette valeur mesure l'écart par rapport à la norme du champ pour la classe à laquelle l'enregistrement est affecté.

Vous pouvez éventuellement supprimer les scores des enregistrements réguliers afin de rendre les résultats plus lisibles.

Figure 4-9

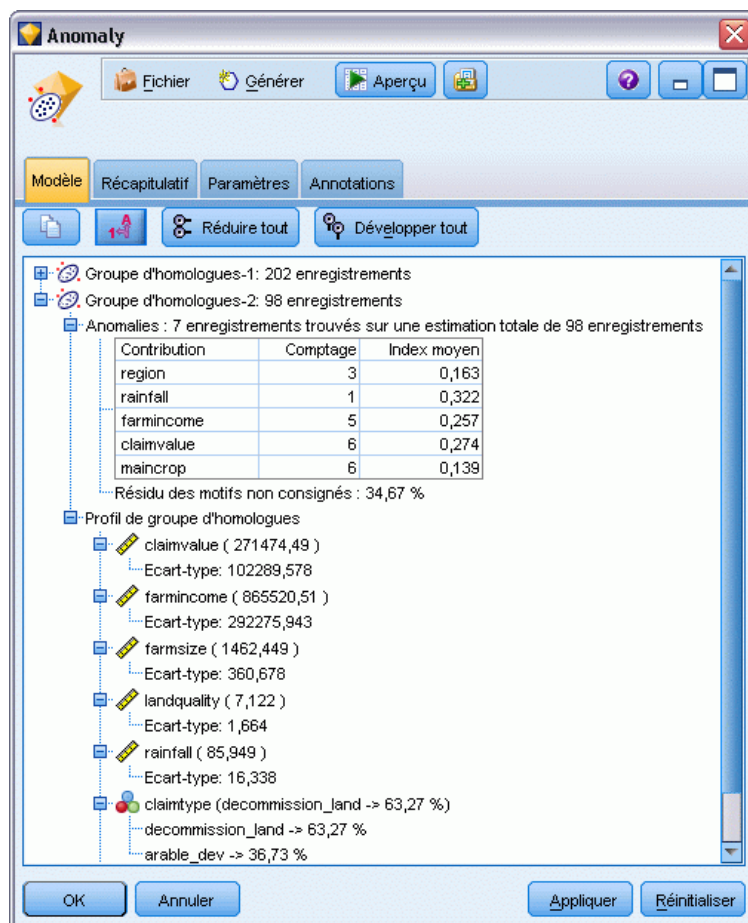
Résultats de scoring dont les enregistrements réguliers ont été supprimés

	id	$\$O$ -Anomaly	$\$O$ -AnomalyIndex	$\$O$ -PeerGroup	$\$O$ -Field-1	$\$O$ -FieldImpact-1	$\$O$ -Field-2	$\$O$ -FieldImpact-2
1	id633	T	1.600	2	claimvalue	0.358	farmincome	0.275
2	id647	T	1.403	2	farminco...	0.334	claimvalue	0.161
3	id654	T	1.495	2	rainfall	0.322	maincrop	0.181
4	id703	T	1.358	1	rainfall	0.230	region	0.219
5	id704	T	1.427	2	farminco...	0.287	maincrop	0.190
6	id739	T	1.684	2	claimvalue	0.404	farmincome	0.233
7	id752	T	1.770	2	claimvalue	0.391	farmincome	0.155
8	id791	T	1.386	1	maincrop	0.236	rainfall	0.163
9	id813	T	1.641	1	region	0.181	landquality	0.160
10	id883	T	1.350	2	region	0.187	maincrop	0.169

Détails relatifs aux modèles Détection des anomalies

L'onglet Modèle d'un modèle Détection des anomalies affiche des informations sur les groupes d'homologues de ce modèle.

Figure 4-10
Détails relatifs au nugget de modèle Détection des anomalies

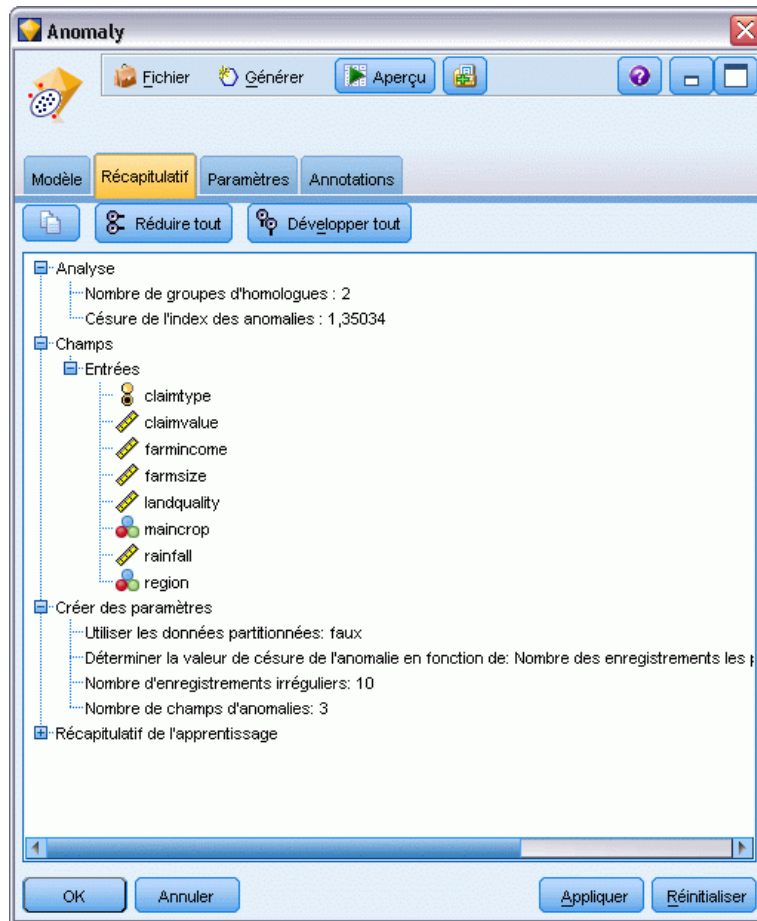


Notez que la taille et les statistiques signalées pour les groupes d'homologues sont évaluées sur la base des données d'apprentissage ; elles peuvent différer légèrement des résultats de scoring réels, et ce même si elles sont obtenues à partir des mêmes données.

Récapitulatif du modèle Détection des anomalies

L'onglet Récapitulatif relatif à un nugget de modèle Détection des anomalies affiche des informations sur les champs, les paramètres de création et le processus d'estimation. Le nombre de groupes d'homologues est également fourni, ainsi que la valeur de césure utilisée pour signaler les enregistrements comme étant anormaux.

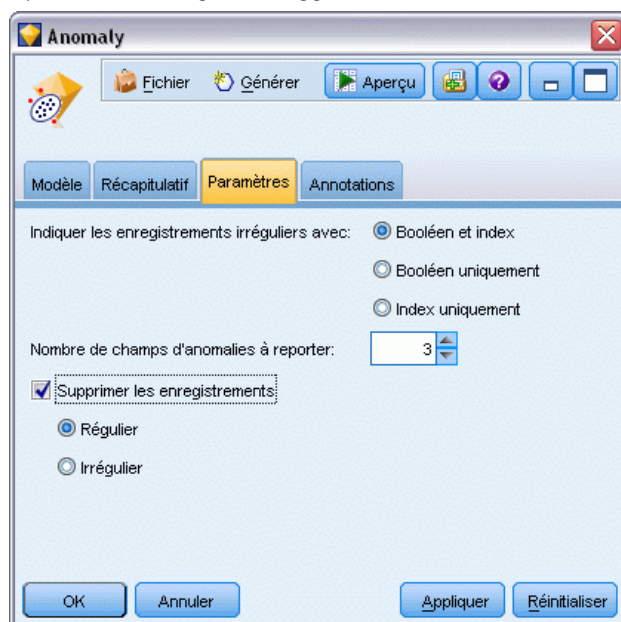
Figure 4-11
Récapitulatif du nugget de modèle Détection des anomalies



Paramètres des modèles Détection des anomalies

L'onglet Paramètres vous permet d'indiquer des options pour le scoring du nugget de modèle.

Figure 4-12
Options de scoring d'un nugget de modèle Détection des anomalies



Indiquer les enregistrements irréguliers avec. Indique le mode de traitement des enregistrements anormaux dans la sortie.

- **Booléen et index.** Crée un champ booléen paramétré sur *True(vrai)* pour tous les enregistrements qui dépassent la valeur de césure incluse dans le modèle. L'index d'anomalies est également fourni dans un champ distinct pour chaque enregistrement. [Pour plus d'informations, reportez-vous à la section Options des modèles Détection des anomalies sur p. 88.](#)
- **Booléen uniquement.** Crée un champ booléen mais ne fournit pas l'index d'anomalies pour chaque enregistrement.
- **Index uniquement.** Fournit l'index d'anomalies sans créer de champ booléen.

Nombre de champs d'anomalies à reporter. Définit le nombre de champs à signaler, afin d'indiquer la raison pour laquelle un enregistrement donné est marqué comme étant anormal. Les champs les plus irréguliers sont indiqués et définis comme étant ceux présentant le plus grand écart par rapport à la valeur « normale » de champ de la classe à laquelle est affectée l'enregistrement.

Supprimer les enregistrements. Sélectionnez cette option pour supprimer du flux tous les enregistrements ne présentant pas d'anomalies ; il est ainsi plus facile de se concentrer sur les anomalies potentielles survenues dans les noeuds en aval. Vous pouvez également ignorer tous les enregistrements anormaux afin de limiter l'analyse suivante aux enregistrements qui ne sont pas signalés comme présentant éventuellement des anomalies, en fonction du modèle défini.

Remarque : Compte tenu des légères différences d'arrondi, le nombre réel d'enregistrements signalés au cours du scoring risque d'être différent de celui trouvé lors de l'apprentissage du modèle, et ce même si les mêmes données sont exploitées.

Noeuds de modélisation automatisés

Les noeuds de modélisation automatisés évaluent et comparent de nombreuses méthodes de modélisation différentes, vous permettant de tester diverses approches en une seule modélisation. Vous pouvez sélectionner les algorithmes de modélisation à utiliser et les options spécifiques à chacun, y compris les combinaisons qui normalement s'exclueraient mutuellement. Par exemple, au lieu de choisir les méthodes Rapide, Dynamique ou Elagage d'un réseau de neurones, vous pouvez tester toutes ces méthodes. Le noeud explore toutes les combinaisons d'options possibles, classe chaque modèle candidat en fonction de la mesure que vous spécifiez et enregistre le meilleur à utiliser dans le scoring ou une analyse supplémentaire.

Vous pouvez choisir parmi trois noeuds de modélisation automatisés en fonction des besoins de votre analyse :



Le noeud Classificateur automatique crée et compare les résultats binaires de plusieurs modèles différents (oui ou non, avec ou sans attrition, etc.), ce qui vous permet de choisir la meilleure approche pour une analyse donnée. Plusieurs algorithmes de modélisation sont pris en charge. Vous pouvez alors sélectionner les méthodes que vous souhaitez utiliser, les options spécifiques pour chacune d'elles et le critère de comparaison des résultats. Le noeud génère un ensemble de modèles basé sur les options spécifiées et classe les meilleurs candidats en fonction des critères indiqués. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique sur p. 99.](#)



Le noeud Numérisation automatique évalue et compare des modèles pour des résultats d'intervalle numérique continus par le biais de différentes méthodes. Le noeud fonctionne de la même manière que le noeud Classificateur automatique, vous permettant ainsi de choisir les algorithmes à utiliser et à tester avec différentes combinaisons d'options en un seul passage de modélisation. Les algorithmes pris en charge comprennent les réseaux neuraux, l'arborescence C&R Tree, CHAID, la régression linéaire, la régression linéaire généralisée et support vector machines (SVM). Les modèles peuvent être comparés selon la corrélation, l'erreur relative ou le nombre de variables utilisées. [Pour plus d'informations, reportez-vous à la section Noeud Numérisation automatique sur p. 110.](#)



Le noeud Classification automatique évalue et compare les modèles de classification identifiant des groupes d'enregistrements ayant des caractéristiques similaires. Le noeud fonctionne de la même manière que les autres noeuds de modélisation automatisés, vous permettant de tester plusieurs combinaisons d'options en une seule modélisation. Les modèles peuvent être comparés à l'aide de mesures de bases permettant d'essayer de filtrer et de classer l'utilité des modèles de classification et de fournir une mesure en fonction de l'importance de champs particuliers. [Pour plus d'informations, reportez-vous à la section Noeud Classification automatique sur p. 117.](#)

Les meilleurs modèles sont enregistrés dans un seul nœud de modèle composite, vous permettant de les parcourir et de les comparer, et de choisir les modèles à utiliser dans le scoring.

- Pour les cibles binaires, nominales et numériques uniquement, vous pouvez sélectionner plusieurs modèles de scoring et combiner les scores dans un seul ensemble de modèles. En combinant des prévisions à partir de différents modèles, vous pouvez éviter les limites des modèles individuels, vous permettant ainsi d'obtenir une plus grande précision globale de chacun des modèles.
- Facultativement, vous pouvez choisir de faire défiler les résultats et générer des noeuds de modélisation ou des nuggets de modèle pour tous les modèles individuels que vous souhaitez utiliser ou analyser plus en détail.

Modèles et temps d'exécution

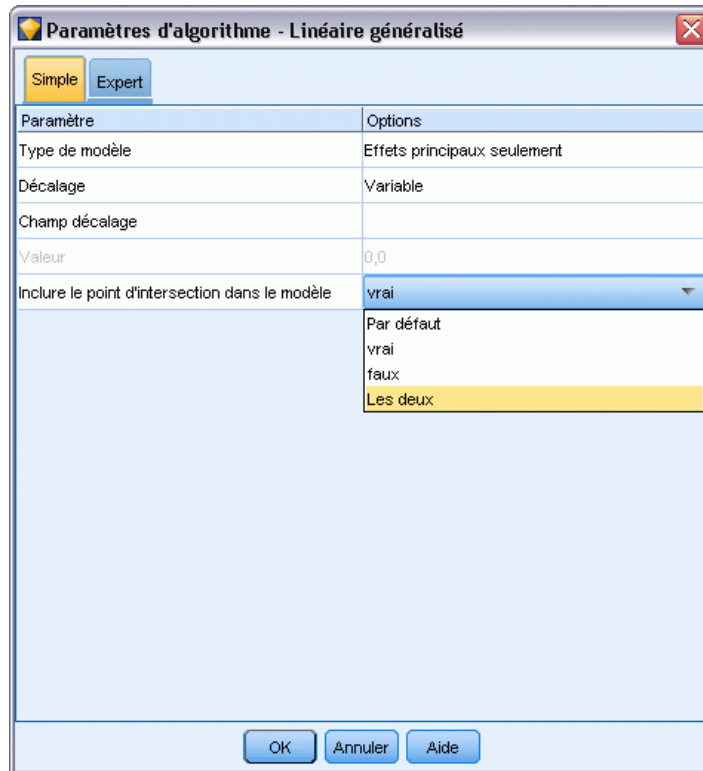
En fonction de l'ensemble de données et du nombre de modèles, l'exécution des noeuds de modélisation automatisés peut prendre des heures ou même davantage. Lorsque vous sélectionnez les options, faites attention au nombre de modèles produits. Pour plus de commodité, vous pouvez programmer l'exécution des modélisations pendant la nuit ou les week-ends, lorsque les ressources système sont soumises à une demande moindre.

- Si nécessaire, utilisez un noeud Partitionner ou Echantillonner pour réduire le nombre d'enregistrements inclus dans l'étape d'apprentissage initiale. Lorsque vous avez restreint les choix à quelques modèles candidats, vous pouvez restaurer l'ensemble de données complet. Pour plus d'informations, reportez-vous à [Noeud Echantillonner](#) ou à [Noeud Partitionner](#).
- Pour limiter le nombre de champs d'entrée, utilisez Sélection de fonction. [Pour plus d'informations, reportez-vous à la section Noeud Sélection de fonction dans le chapitre 4 sur p. 78.](#) Vous pouvez également utiliser les exécutions de modélisation initiales pour identifier les champs et les options à analyser plus en profondeur. Par exemple, s'il vous semble que vos modèles enregistrant les meilleurs résultats utilisent les trois mêmes champs, cela signifie que ces champs doivent être conservés.
- Vous pouvez également limiter le temps passé à estimer un modèle particulier et préciser les mesures d'évaluation utilisées pour filtrer et classer les modèles.

Noeud de modélisation automatisé - Paramètres d'algorithme

Vous pouvez utiliser les paramètres par défaut pour chaque type de modèle ou sélectionner des options pour chaque type de modèle. Les options sont semblables à celles proposées pour chaque noeud de modélisation, à la seule différence qu'au lieu de choisir un paramètre, vous pouvez en choisir autant que vous le souhaitez à appliquer à la grande majorité des cas. Ainsi, si vous comparez les modèles de réseau de neurones, vous pouvez choisir différentes méthodes d'apprentissage et essayez chaque méthode, avec ou sans graine aléatoire. Toutes les combinaisons possibles des options sélectionnées seront utilisées, facilitant ainsi la génération des différents modèles dans un seul passage. Faites preuve de prudence car choisir plusieurs paramètres peut entraîner une multiplication rapide du nombre de modèles.

Figure 5-1
Choisir les paramètres d'algorithme pour la modélisation automatisée



Pour choisir des options pour chaque type de modèle

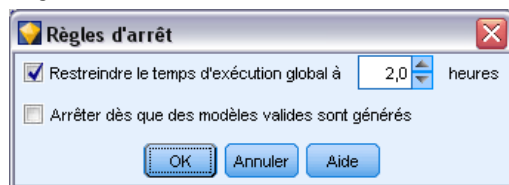
- ▶ Dans le noeud de modélisation automatique, sélectionnez l'onglet Expert.
- ▶ Cliquez dans la colonne Paramètres du modèle pour le type de modèle.
- ▶ Dans le menu déroulant, sélectionnez la commande Spécifier.
- ▶ Dans la boîte de dialogue Paramètres d'algorithme, sélectionnez des options dans la colonne Options.

Remarque : D'autres options sont disponibles dans l'onglet Expert de la boîte de dialogue Paramètres d'algorithme.

Noeuds de modélisation automatisés - Règles d'arrêt

Les règles d'arrêt définies pour les noeuds de modélisation automatisée s'appliquent à l'exécution du noeud entier et non aux différents modèles créés par le noeud.

Figure 5-2
Règles d'arrêt



Restreindre le temps d'exécution global à. Arrête l'exécution après un nombre d'heures spécifié (pour les modèles Réseau de neurones, K-Means, Kohonen, TwoStep, SVM, KNN, Réseau Bayes et Arbre C&RT uniquement). Tous les modèles générés jusqu'à ce moment-là sont inclus dans le nugget de modèle, mais aucun autre modèle n'est ensuite généré.

Arrêtez dès que des modèles valides sont générés. Arrête l'exécution lorsqu'un modèle transmet tous les critères définis dans l'onglet Supprimer (pour le noeud Classificateur automatique ou Classification automatique) ou dans l'onglet Modèle (pour le noeud Numérisation automatique). [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique - Options de suppression sur p. 107.](#) [Pour plus d'informations, reportez-vous à la section Noeud Classification - Options de suppression sur p. 121.](#)

Noeud Classificateur automatique

Le noeud Classificateur automatique évalue et compare des modèles pour des cibles nominales (ensemble) ou binaires (oui/non) à l'aide de plusieurs méthodes différentes, vous permettant de tester diverses approches dans une seule modélisation. Vous pouvez sélectionner les algorithmes à utiliser et tester diverses combinaisons d'options. Par exemple, au lieu de choisir les méthodes Rapide, Dynamique ou Elagage d'un réseau de neurones, vous pouvez tester toutes ces méthodes. Le noeud explore toutes les combinaisons d'options possibles, classe chaque modèle candidat en fonction de la mesure que vous spécifiez et enregistre les meilleurs modèles à utiliser dans le scoring ou une analyse supplémentaire. [Pour plus d'informations, reportez-vous à la section Noeuds de modélisation automatisés sur p. 96.](#)

Figure 5-3
Résultats de modélisation du Classificateur automatique

Utiliser ?	Graphiques	Modèle	Durée de création (min)	Profit max	Le profit max se produit dans	Augmentation(Pr...	Précision globale (%)	Nbre champs utilisés	Aire sous la courbe
<input checked="" type="checkbox"/>		CS 1	< 1	4 906,667	8	2,203	92,861	10	0,777
<input checked="" type="checkbox"/>		C&R T...	3	4 602,692	9	2,776	92,365	8	0,924
<input checked="" type="checkbox"/>		CHAI...	3	4 145,668	8	2,851	91,706	4	0,927

Exemple : Une société possède des données historiques sur les offres faites à des clients spécifiques lors des campagnes passées. Cette société souhaite réaliser des résultats plus rentables en adaptant l'offre à chaque client. [Démonstration](#)

Conditions requises. Un champ cible avec un niveau de mesure *Nominal* ou *Booléen* (dont le rôle est défini comme *Cible*), et au moins un champ d'entrée (dont le rôle est défini comme *Entrée*). Pour un champ booléen, la valeur *Vrai* définie pour le champ cible est considérée comme une correspondance lors du calcul des profits, du Lift et des statistiques connexes. Les champs d'entrée peuvent avoir un niveau de mesure *Continu* ou *Catégoriel*, avec les limites impliquant que certaines entrées peuvent ne pas être appropriées à certains types de modèles. Par exemple, les champs ordinaux utilisés comme entrées dans les modèles C&R Tree, CHAID et QUEST doivent disposer d'un stockage numérique (et non d'une chaîne), et seront ignorés par ces modèles si indication contraire. De la même manière, dans certains cas, les champs d'entrée continus peuvent être discrétisés. Les exigences sont les mêmes que pour les noeuds de modélisation individuels ; par exemple, un modèle du réseau Bayes fonctionne de la même façon s'il est généré à partir du noeud du réseau Bayes ou du noeud Classificateur automatique.

Les champs d'effectif et de pondération. L'effectif et la pondération sont utilisés pour donner plus d'importance à certains enregistrements ; par exemple, l'utilisateur sait que l'ensemble de données de création sous-représente une section de la population parent (Pondération) ou parce qu'un enregistrement représente un nombre d'observations identiques (Effectif). S'il cela est indiqué, un champs d'effectif peut être utilisé par les modèles de réseau C&R Tree, CHAID, QUEST, Liste de décision et Bayes. Un champs de pondération peut être utilisé par les modèles C&RT, CHAID et C5.0. Les autres types de modèles ignoreront ces champs et créeront les modèles de toute façon. Les champs d'effectif et de pondération sont utilisés uniquement pour la création de modèle et ne sont pas pris en compte lors de l'évaluation des modèles. [Pour plus d'informations, reportez-vous à la section Utilisation des champs d'effectif et de pondération dans le chapitre 3 sur p. 42.](#)

Types de modèle pris en charge

Les types de modèles pris en charge sont le réseau de neurones, C&R Tree, QUEST, CHAID, C5.0, Régression logistique, Liste de décision, Bayes Net, Discriminant, Agrégation suivant le saut minimum et SVM. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique - Options expert sur p. 103.](#)

Noeud Classificateur automatique - Options du modèle

L'onglet Modèle du noeud Classificateur automatique vous permet de préciser le nombre de modèles à créer, ainsi que les critères utilisés pour comparer les modèles.

Figure 5-4
Noeud Classificateur automatique : Onglet Modèle

The screenshot shows the 'response' dialog box with the 'Modèle' tab selected. The window title is 'response'. At the top, it says 'Nombre estimé de modèles à exécuter : 9'. Below this are several tabs: 'Champs', 'Modèle', 'Expert', 'Supprimer', 'Paramètres', and 'Annotations'. The 'Modèle' tab is active. The 'Nom du modèle:' section has two radio buttons: 'Auto' (selected) and 'Personnalisé'. Below this are two checked checkboxes: 'Utiliser les données partitionnées' and 'Créer un modèle pour chaque division'. The 'Classer les modèles par:' dropdown is set to 'Précision globale'. The 'Classer les modèles avec:' section has two radio buttons: 'Partition d'apprentissage' and 'Partition de test' (selected). The 'Nombre de modèles à utiliser:' is a spinner box set to '3'. There is a checked checkbox for 'Calculer l'importance des valeurs prédites'. Below this are two sections for criteria: 'Critères de profit (valides uniquement pour les cibles de type booléen)' and 'Critères d'augmentation (valides uniquement pour les cibles de type booléen)'. The profit section has three rows: 'Coûts' with 'Fixe' selected and value '5,0', 'Recette' with 'Fixe' selected and value '10,0', and 'Pondération' with 'Fixe' selected and value '1,0'. Each row has a 'Variable' option with a spinner. The augmentation section has a 'Centile à utiliser pour le calcul de l'augmentation:' spinner set to '30'. At the bottom are buttons for 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser'.

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*.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Classer les modèles par. Indique les critères utilisés pour comparer et classer des modèles. Les options comprennent : précision générale, zone sous la courbe ROC, profit, Lift et nombre de champs. Notez que toutes ces mesures seront disponibles dans le rapport récapitulatif, quelle que soit la sélection.

Remarque : Pour une cible nominale (ensemble), le classement est limité à **Précision globale** ou **Nombre de champs**.

Lors du calcul des profits, du Lift et des statistiques connexes, la valeur *Vrai* définie pour le champ cible est considérée comme une correspondance.

-
-
-
-
-

Classer les modèles avec. Lorsque vous utilisez une partition, vous pouvez indiquer si les rangs doivent être basés sur l'ensemble de données d'apprentissage ou sur l'ensemble de test. Lorsque les ensembles de données sont volumineux, utiliser une partition pour le filtrage préliminaire des modèles permet d'améliorer considérablement les performances.

Nombre de modèles à utiliser. Indique le nombre maximal de modèles à répertorier dans le nugget de modèle généré par le noeud. Les modèles situés en tête de classement seront répertoriés conformément au critère de classement indiqué. Les performances peuvent être moindres si vous augmentez cette limite. La valeur maximum autorisée est 100.

Calculer l'importance des valeurs prédites. Pour des modèles qui produisent une mesure appropriée d'importance, vous pouvez afficher un graphique qui indique l'importance relative de chaque valeur prédite dans l'estimation du modèle. En général, vous souhaitez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Notez que l'importance de la valeur prédite peut augmenter la durée nécessaire pour calculer certains modèles et ce n'est pas recommandé si vous souhaitez effectuer uniquement une large comparaison entre les différents modèles. Elle est plus utile une fois que vous avez réduit votre analyse à quelques modèles que vous souhaitez analyser plus en détail. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Critères de profit.*Remarque.* Uniquement pour les cibles de type booléen. Le profit est égal au revenu de chaque enregistrement moins le coût de l'enregistrement. Les profits d'un quantile correspondent à la somme des profits de tous ses enregistrements. Les profits sont supposés ne s'appliquer qu'aux correspondances, mais les coûts s'appliquent à tous les enregistrements.

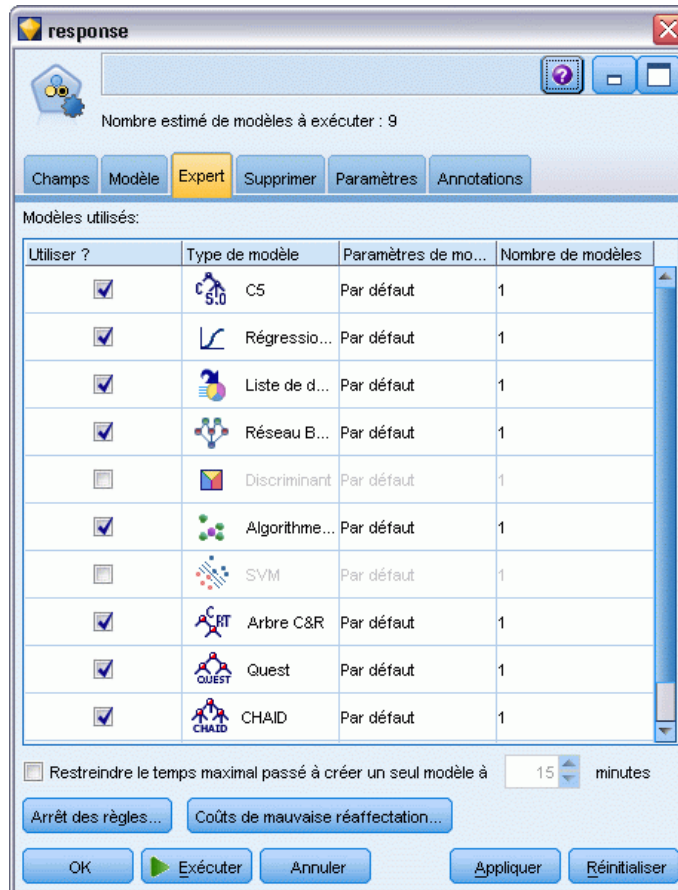
- **Coûts.** Précisez le coût associé à chaque enregistrement. Vous pouvez sélectionner Fixe ou Variable. Pour les coûts fixes, précisez la valeur du coût. Pour les coûts variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de coût.
- **Recette.** Précisez le revenu associé à chaque enregistrement représentant une correspondance. Vous pouvez sélectionner Fixe ou Variable. Pour les revenus fixes, précisez la valeur du revenu. Pour les revenus variables, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de revenu.
- **Pondération.** Si les enregistrements de vos données représentent plusieurs unités, vous pouvez utiliser les pondérations d'effectif pour ajuster les résultats. Indiquez la pondération associée à chaque enregistrement, à l'aide des options Fixe ou Variable. Pour une pondération fixe, précisez la valeur de la pondération (nombre d'unités par enregistrement). Pour une pondération variable, cliquez sur le sélecteur de champs pour sélectionner un champ comme champ de pondération.

Critères de Lift (augmentation)*Remarque.* Uniquement pour les cibles de type booléen. Indiquez l'utilisation du centile dans les calculs de Lift (augmentation). Vous pouvez également modifier cette valeur lorsque vous comparez les résultats. [Pour plus d'informations, reportez-vous à la section Nuggets de modèle automatisés sur p. 122.](#)

Noeud Classificateur automatique - Options expert

L'onglet Expert du noeud Classificateur automatique vous permet d'appliquer une partition (si elle existe), de sélectionner les algorithmes à utiliser et de définir les règles d'arrêt.

Figure 5-5
Noeud Classificateur automatique : Onglet Expert



Modèles utilisés. Dans la colonne de gauche, cochez les cases correspondant aux types de modèle (algorithmes) à inclure dans la comparaison. Plus vous sélectionnez de types, plus un nombre important de modèles sont créés et plus le traitement est long.

Type de modèle. Répertorie les algorithmes disponibles (voir ci-dessous).

Paramètres de modèle. Vous pouvez utiliser les paramètres par défaut pour chaque type de modèle ou sélectionner Spécifier pour choisir des options pour chaque type de modèle. Les options sont semblables à celles proposées pour chaque noeud de modélisation. Cependant, vous pouvez sélectionner ici plusieurs options ou combinaisons. Par exemple, si vous comparez des modèles R. neurones, vous pouvez choisir toutes les méthodes d'apprentissage pour apprendre six modèles en une seule étape au lieu de sélectionner chaque méthode indépendamment.

Nombre de modèles créés. Indique le nombre de modèles générés pour chaque algorithme en fonction des paramètres actuels. Lorsque vous combinez des options, le nombre de modèles peut augmenter rapidement. Il est donc fortement recommandé de surveiller ce nombre, en particulier si vous utilisez des ensembles de données volumineux.

Restreindre le temps maximal passé à créer un seul modèle. (Pour les modèles K-Means, Kohonen, TwoStep, SVM, KNN, Bayes Net et Liste de décision) Définit une limite de temps maximale pour un modèle donné. Par exemple, si l'apprentissage d'un modèle donné prenait un temps particulièrement long du fait d'une interaction complexe, vous ne voudriez pas qu'il ralentisse l'exécution de la modélisation complète.

Remarque : Si la cible est un champ nominal (ensemble), l'option Liste de décision n'est pas disponible.

Algorithmes pris en charge



Le noeud R. neurones est un modèle simplifié de la manière dont le cerveau humain traite les informations. Le fonctionnement de ce modèle repose sur la simulation d'un grand nombre d'unités de traitement simples interconnectées, qui sont en quelque sorte des versions abstraites de nos neurones. Les réseaux de neurones sont de puissants estimateurs de fonctions qui ne requièrent qu'une connaissance limitée en matière de statistiques ou de mathématiques.



Le noeud C5.0 crée un arbre décision ou un ensemble de règles. Le fonctionnement de ce modèle repose sur un découpage de l'échantillon basé sur le champ qui fournit le gain d'informations le plus important à chaque niveau. Le champ cible doit être catégoriel. Les divisions multiples en plus de deux sous-groupes sont autorisées. [Pour plus d'informations, reportez-vous à la section Noeud C5.0 dans le chapitre 6 sur p. 180.](#)



Le noeud Arbre Classification et Regression, (C&R) Tree, génère un arbre décision qui vous permet de prévoir ou de classifier les observations futures. La méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments en réduisant l'index d'impureté à chaque étape, un noeud de l'arbre étant considéré comme "pur" si 100 % de ses observations appartiennent à une catégorie spécifique du champ cible. Les champs cible et les champs d'entrée peuvent être des champs d'intervalle numériques ou des champs catégoriels numériques (nominal, ordinal ou booléen). Toutes les divisions sont binaires (deux sous-groupes uniquement). [Pour plus d'informations, reportez-vous à la section Noeud Arbre C&R dans le chapitre 6 sur p. 158.](#)



Le noeud QUEST est une méthode de classification supervisée binaire permettant de créer des arbres décision, développée pour réduire le temps de traitement nécessaire aux analyses C&R Tree importantes, tout en limitant la tendance, observée parmi les méthodes d'arbre de classification, à favoriser les entrées autorisant un nombre supérieur de divisions. Les champs d'entrée peuvent être des intervalles numériques (continues) mais les champs cible doivent être catégoriels. Toutes les divisions sont binaires. [Pour plus d'informations, reportez-vous à la section Noeud QUEST dans le chapitre 6 sur p. 160.](#)



Le noeud CHAID génère des arbres décision à l'aide des statistiques du Chi-deux pour identifier les séparations optimales. Contrairement aux noeuds C&R Tree et QUEST, CHAID peut générer des arbres non binaires, ce qui implique que certaines divisions possèdent plusieurs branches. Les champs cibles et les champs d'entrée peuvent être d'intervalle numérique (continu) ou catégoriels. La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en détail toutes les divisions possibles, mais dont les calculs sont plus longs. [Pour plus d'informations, reportez-vous à la section Noeud CHAID dans le chapitre 6 sur p. 159.](#)



La régression logistique est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est similaire à la régression linéaire. [Pour plus d'informations, reportez-vous à la section Noeud Logistique dans le chapitre 10 sur p. 288.](#)



Le noeud Liste de décision identifie les sous-groupes, ou les segments, qui présentent une probabilité plus élevée ou plus faible d'un résultat binaire donné par rapport à la population globale. Vous pouvez, par exemple, rechercher les clients qui ont une faible probabilité d'attrition ou ceux qui ont une plus forte probabilité de répondre favorablement à une campagne. Vous pouvez incorporer votre connaissance du marché dans le modèle en ajoutant vos propres segments personnalisés et en prévisualisant des modèles alternatifs côte à côte de façon à comparer les résultats. Les modèles Liste de décision se composent d'une liste de règles dans laquelle chaque règle présente une condition et un résultat. Les règles sont appliquées dans l'ordre et la première règle correspondante détermine le résultat. [Pour plus d'informations, reportez-vous à la section Liste de décision dans le chapitre 9 sur p. 229.](#)



Le noeud Réseau Bayésien permet de créer un modèle de probabilité en combinant les preuves observées et enregistrées avec les connaissances réelles pour établir la probabilité des occurrences. Le noeud est axé sur le Tree Augmented Naïve Bayes (TAN) et sur les réseaux Couverture de Markov qui servent principalement à la classification. [Pour plus d'informations, reportez-vous à la section Noeud Réseau Bayésien dans le chapitre 7 sur p. 203.](#)



L'analyse discriminante crée des hypothèses plus strictes que la régression logistique mais peut constituer une alternative ou un complément précieux à une analyse de régression logistique lorsque ces hypothèses sont réunies. [Pour plus d'informations, reportez-vous à la section Noeud discriminant dans le chapitre 10 sur p. 318.](#)



Le noeud k -Voisin le plus proche (KNN) associe une nouvelle observation à la catégorie ou à la valeur des objets k les plus proches dans l'espace du prédicteur, où k est un entier. Les observations semblables sont proches l'une de l'autre et les observations dissemblables sont éloignées l'une de l'autre. [Pour plus d'informations, reportez-vous à la section Noeud KNN dans le chapitre 16 sur p. 507.](#)



Le noeud Support Vector Machine (SVM) vous permet de classer les données dans l'un de deux groupes sans sur-ajustement. SVM fonctionne bien avec les grands ensembles de données, comme ceux qui disposent d'un très grand nombre de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud SVM dans le chapitre 15 sur p. 501.](#)

Coûts de mauvaise réaffectation

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*.

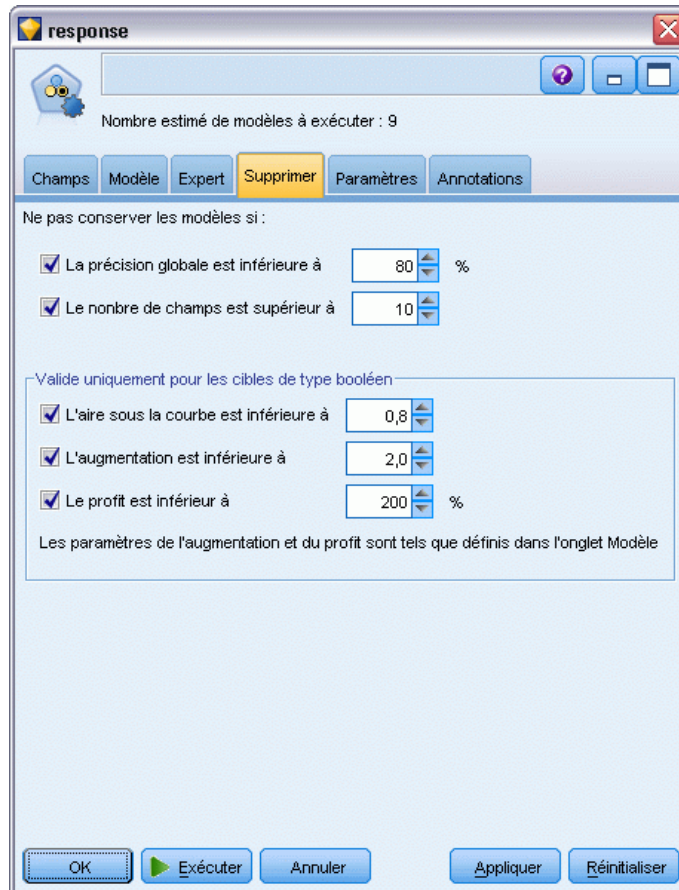
La matrice de mauvaise réaffectation des coûts affiche le coût de chaque combinaison possible de catégories prédites et de catégories réelles. Par défaut, tous les coûts de mauvaise réaffectation sont paramétrés sur 1. Pour entrer des valeurs de coût personnalisées, sélectionnez Utiliser les coûts de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

Noeud Classificateur automatique - Options de suppression

L'onglet Supprimer du noeud Classificateur automatique vous permet de supprimer automatiquement les modèles qui ne répondent pas à certains critères. Ces modèles ne sont pas répertoriés dans le rapport récapitulatif.

Figure 5-6
Noeud Classificateur automatique : Onglet Supprimer



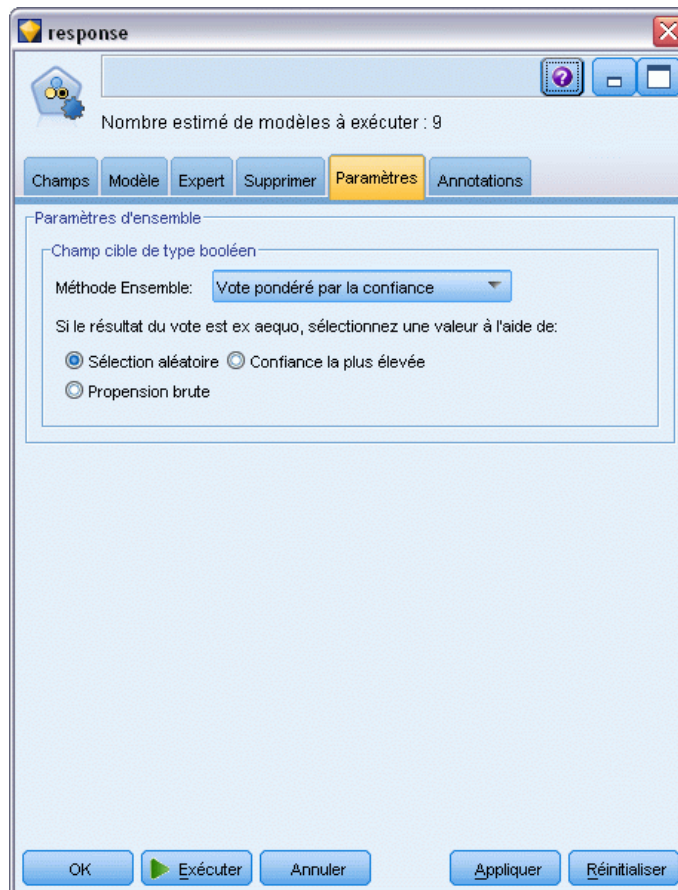
Vous pouvez spécifier un seuil minimal pour le degré de précision globale et un seuil maximal pour le nombre de variables utilisées dans le modèle. En outre, pour les cibles de type booléen, vous pouvez spécifier un seuil minimal de Lift (augmentation), de profit et de zone sous la courbe ; le Lift et le profit sont déterminés comme spécifiés dans l'onglet Modèle. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique - Options du modèle sur p. 101.](#)

Vous pouvez également configurer le noeud pour qu'il interrompe son exécution dès qu'un modèle remplissant tous les critères définis est généré. [Pour plus d'informations, reportez-vous à la section Noeuds de modélisation automatisés - Règles d'arrêt sur p. 98.](#)

Noeud Classificateur automatique - Options des paramètres

L'onglet Paramètres du noeud Classificateur automatique vous permet de préconfigurer les options de temps de score disponibles sur le nugget.

Figure 5-7
Noeud Classificateur automatique : onglet Paramètres



Méthode d'ensemble. Vous pouvez sélectionner l'une des méthodes d'ensemble suivantes pour les cibles :

- Voting
- Vote pondéré par la confiance
- Vote pondéré par la propension brute (cibles de type booléen uniquement)
- La confiance la plus élevée l'emporte
- Propension brute moyenne (cibles de type booléen uniquement).

Pour plus d'informations, reportez-vous à la section Paramètres du noeud Ensemble dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.

Si le vote est ex æquo, sélectionnez l'utilisation d'une valeur. Pour les méthodes de vote, vous pouvez indiquer le mode de résolution des ex æquo :

- **Sélection aléatoire.** Une des valeurs ex æquo est choisie au hasard.

- **Confiance la plus grande.** La valeur ex æquo prédite avec la plus grande confiance gagne. Remarque : ce n'est pas forcément la même que la plus grande confiance de toutes les valeurs prédites.
- **Propension brute.** (Uniquement pour les cibles de type booléen) La valeur ex æquo prédite avec la plus grande propension absolue, où la propension absolue est calculée avec :

$$\text{abs}(0.5 - \text{propensity}) * 2$$

Noeud Numérisation automatique

Le noeud Numérisation automatique évalue et compare les résultats d'intervalle numérique connu des modèles à l'aide de plusieurs méthodes différentes, vous permettant de tester diverses approches dans une seule modélisation. Vous pouvez sélectionner les algorithmes à utiliser et tester diverses combinaisons d'options. Par exemple, vous pouvez prévoir des valeurs immobilières par le biais des modèles de réseau de neurones, de régression linéaire, de C&RT et de CHAID et voir les meilleurs résultats. Vous pouvez également essayer différentes combinaisons de méthodes de régression Pas à pas, Ascendante et Descendante. Le noeud explore toutes les combinaisons d'options possibles, classe chaque modèle candidat en fonction de la mesure que vous spécifiez et enregistre le meilleur à utiliser dans le scoring ou une analyse supplémentaire. [Pour plus d'informations, reportez-vous à la section Noeuds de modélisation automatisés sur p. 96.](#)

Figure 5-8
Résultats de la numérisation automatique

Utiliser ?	Graphiques	Modèle	Durée de création (min)	Corrélation	Nbre champs utilisés	Erreur relative
<input checked="" type="checkbox"/>		Generaliz...	< 1	0,915	7	0,162
<input checked="" type="checkbox"/>		Regressi...	< 1	0,9	5	0,19
<input checked="" type="checkbox"/>		CHAID Tr...	< 1	0,892	5	0,204

Exemple : Une municipalité souhaite pouvoir évaluer de façon plus précise les taxes immobilières et ajuster les valeurs de certaines propriétés sans avoir à vérifier toutes les propriétés. Par le biais du noeud Numérisation automatique, l'analyste va pouvoir générer et comparer des modèles

qui prévoient les valeurs de propriétés selon le type de construction, le voisinage, la taille et d'autres facteurs connus. [Pour plus d'informations, reportez-vous à la section Valeurs de propriété \(Numérisation automatique\) dans le chapitre 5 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Un champ cible unique (dont le rôle est défini comme Cible), et au moins un champ d'entrée (dont le rôle est défini comme Entrée). La cible doit être un champ continu (intervalle numérique), tel que *âge* ou *revenu*. Les champs d'entrée peuvent être des champs continus ou des champs catégoriels. Néanmoins, certaines entrées peuvent ne pas être adaptées à certains types de modèles. Par exemple, les modèles d'arbre C&RT peuvent utiliser des champs catégoriels de type chaîne comme entrées, tandis que les modèles de régression linéaire ne peuvent utiliser ces champs et les ignoreront selon l'indication. Les conditions requises sont les mêmes que celles des noeuds de modélisation individuels. Par exemple, un modèles CHAID fonctionne de la même manière s'il est généré à partir d'un noeud CHAID ou d'un noeud Numérisation automatique.

Les champs d'effectif et de pondération. L'effectif et la pondération sont utilisés pour donner plus d'importance à certains enregistrements ; par exemple, l'utilisateur sait que l'ensemble de données de création sous-représente une section de la population parent (Pondération) ou parce qu'un enregistrement représente un nombre d'observations identiques (Effectif). Si cela est indiqué, les algorithmes CHAID et C&RT Tree peuvent utiliser un champ d'effectif. Un champs de pondération peut être utilisé par les algorithmes C&RT, CHAID, de régression et GenLin Les autres types de modèles ignoreront ces champs et créeront les modèles de toute façon. Les champs d'effectif et de pondération sont utilisés uniquement pour la création de modèle et ne sont pas pris en compte lors de l'évaluation des modèles. [Pour plus d'informations, reportez-vous à la section Utilisation des champs d'effectif et de pondération dans le chapitre 3 sur p. 42.](#)

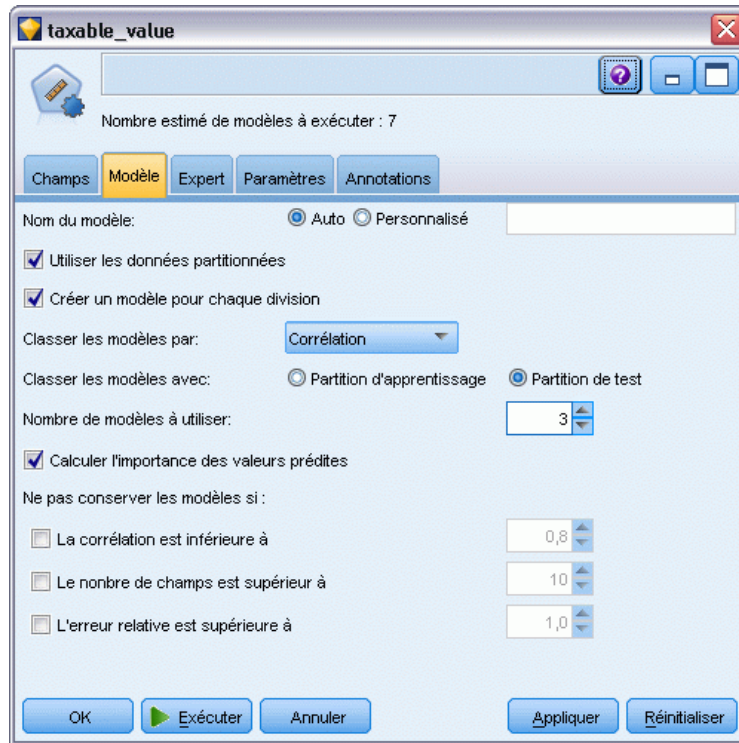
Types de modèle pris en charge

Les types de modèles pris en charge sont Réseau de neurones, Arbre C&R, CHAID, Régression, GenLin, Agrégation suivant le saut minimum et SVM. [Pour plus d'informations, reportez-vous à la section Noeud Numérisation automatique - Options expert sur p. 114.](#)

Noeud Numérisation automatique - Options du modèle

L'onglet Modèle du noeud Numérisation automatique vous permet de préciser le nombre de modèles à enregistrer, ainsi que les critères utilisés pour comparer les modèles.

Figure 5-9
Noeud Numérisation automatique : 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Classer les modèles par. Indique les critères utilisés pour comparer des modèles.

- **Corrélation.** La corrélation de Pearson entre la valeur observée pour chaque enregistrement et la valeur prédite par le modèle. La corrélation est une mesure d'association linéaire entre deux variables, avec des valeurs plus proches de 1 indiquant une relation plus forte. (Les valeurs de corrélation sont comprises entre -1 , pour une relation négative parfaite, et $+1$, pour une relation positive parfaite. La valeur 0 indique qu'il n'existe aucune relation linéaire, tandis qu'un modèle affectée d'une corrélation négative se trouvera au rang le plus bas).

- **Nombre de champs.** Le nombre de champs utilisés comme valeurs prédites dans le modèle. Choisir des modèles qui utilisent moins de champs peut simplifier la préparation des données voire améliorer les performances dans certains cas.
- **Erreur relative.** L'erreur relative est le rapport entre la variance des valeurs observées à partir des valeurs prédites par le modèle et la variance des valeurs observées à partir de la moyenne. En pratique, il compare les performances du modèle par rapport à un modèle **à valeur nulle** ou **à constantes** qui renvoie simplement la valeur moyenne du champ cible comme prévision. Pour un modèle performant, cette valeur doit être inférieure à 1, indiquant que le modèle est plus précis que le modèle à valeur nulle. Un modèle comportant une erreur relative supérieure à 1 est moins précis que le modèle à valeur nulle et il est donc moins utile. Dans le cas des modèles de régression linéaire, l'erreur relative est égale au carré de la corrélation et n'ajoute aucune information nouvelle. Dans le cas des modèles non linéaires, l'erreur relative n'est pas liée à la corrélation et fournit une mesure supplémentaire pour évaluer les performances du modèle.

Classer les modèles avec. Lorsque vous utilisez une partition, vous pouvez indiquer si les rangs doivent être basés sur l'ensemble de données d'apprentissage ou sur une partition de tests. Lorsque les ensembles de données sont volumineux, utiliser une partition pour le filtrage préliminaire des modèles permet d'améliorer considérablement les performances. [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*.](#)

Nombre de modèles à utiliser. Indique le nombre maximal de modèles à afficher dans le nugget de modèle généré par le noeud. Les modèles situés en tête de classement seront répertoriés conformément au critère de classement indiqué. Augmenter cette limite vous permet de comparer des résultats pour plusieurs modèles mais peut ralentir les performances. La valeur maximum autorisée est 100.

Calculer l'importance des valeurs prédites. Pour des modèles qui produisent une mesure appropriée d'importance, vous pouvez afficher un graphique qui indique l'importance relative de chaque valeur prédite dans l'estimation du modèle. En général, vous souhaitez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Notez que l'importance de la valeur prédite peut augmenter la durée nécessaire pour calculer certains modèles et ce n'est pas recommandé si vous souhaitez effectuer uniquement une large comparaison entre les différents modèles. Elle est plus utile une fois que vous avez réduit votre analyse à quelques modèles que vous souhaitez analyser plus en détail. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Ne pas conserver les modèles si. Indique les valeurs de seuil pour la corrélation, l'erreur relative et le nombre de champs. Les modèles qui ne correspondent pas à ces critères seront supprimés et ne seront pas répertoriés dans le rapport récapitulatif.

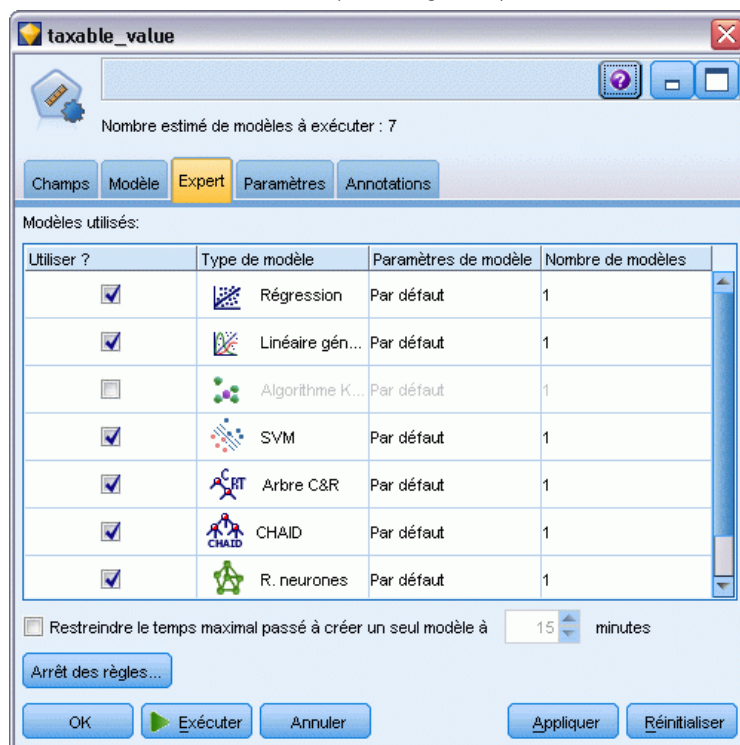
- **Corrélation inférieure à.** La corrélation minimum (en termes de valeur absolue) pour un modèle à inclure dans le rapport récapitulatif.
- **Le nombre de champs est supérieur à.** Le nombre maximum de champs à utiliser par un modèle à inclure.
- **L'erreur relative est supérieure à.** L'erreur relative maximum d'un modèle à inclure.

Vous pouvez également configurer le noeud pour qu'il interrompe son exécution dès qu'un modèle remplissant tous les critères définis est généré. [Pour plus d'informations, reportez-vous à la section Noeuds de modélisation automatisés - Règles d'arrêt sur p. 98.](#)

Noeud Numérisation automatique - Options expert

L'onglet Expert du noeud Numérisation automatique vous permet de sélectionner les algorithmes et les options à utiliser pour déterminer les règles d'arrêt.

Figure 5-10
Noeud Numérisation automatique : Onglet Expert



Modèles utilisés. Dans la colonne de gauche, cochez les cases correspondant aux types de modèle (algorithmes) à inclure dans la comparaison. Plus vous sélectionnez de types, plus un nombre important de modèles sont créés et plus le traitement est long.

Type de modèle. Répertorie les algorithmes disponibles (voir ci-dessous).

Paramètres de modèle. Vous pouvez utiliser les paramètres par défaut pour chaque type de modèle ou sélectionner Spécifier pour choisir des options pour chaque type de modèle. Les options sont semblables à celles proposées pour chaque noeud de modélisation. Cependant, vous pouvez sélectionner ici plusieurs options ou combinaisons. Par exemple, si vous comparez des modèles R. neurones, vous pouvez choisir toutes les méthodes d'apprentissage pour apprendre six modèles en une seule étape au lieu de sélectionner chaque méthode indépendamment.

Nombre de modèles créés. Indique le nombre de modèles générés pour chaque algorithme en fonction des paramètres actuels. Lorsque vous combinez des options, le nombre de modèles peut augmenter rapidement. Il est donc fortement recommandé de surveiller ce nombre, en particulier si vous utilisez des ensembles de données volumineux.

Restreindre le temps maximal passé à créer un seul modèle. (Pour les modèles K-Means, Kohonen, TwoStep, SVM, KNN, Bayes Net et Liste de décision) Définit une limite de temps maximale pour un modèle donné. Par exemple, si l'apprentissage d'un modèle donné prenait un temps particulièrement long du fait d'une interaction complexe, vous ne voudriez pas qu'il ralentisse l'exécution de la modélisation complète.

Algorithmes pris en charge



Le noeud R. neurones est un modèle simplifié de la manière dont le cerveau humain traite les informations. Le fonctionnement de ce modèle repose sur la simulation d'un grand nombre d'unités de traitement simples interconnectées, qui sont en quelque sorte des versions abstraites de nos neurones. Les réseaux de neurones sont de puissants estimateurs de fonctions qui ne requièrent qu'une connaissance limitée en matière de statistiques ou de mathématiques.



Le noeud Arbre Classification et Regression, (C&R) Tree, génère un arbre décision qui vous permet de prévoir ou de classifier les observations futures. La méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments en réduisant l'index d'impureté à chaque étape, un noeud de l'arbre étant considéré comme "pur" si 100 % de ses observations appartiennent à une catégorie spécifique du champ cible. Les champs cible et les champs d'entrée peuvent être des champs d'intervalle numériques ou des champs catégoriels numériques (nominal, ordinal ou booléen). Toutes les divisions sont binaires (deux sous-groupes uniquement). [Pour plus d'informations, reportez-vous à la section Noeud Arbre C&RT dans le chapitre 6 sur p. 158.](#)



Le noeud CHAID génère des arbres décision à l'aide des statistiques du Chi-deux pour identifier les séparations optimales. Contrairement aux noeuds C&R Tree et QUEST, CHAID peut générer des arbres non binaires, ce qui implique que certaines divisions possèdent plusieurs branches. Les champs cibles et les champs d'entrée peuvent être d'intervalle numérique (continu) ou catégoriels. La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en détail toutes les divisions possibles, mais dont les calculs sont plus longs. [Pour plus d'informations, reportez-vous à la section Noeud CHAID dans le chapitre 6 sur p. 159.](#)



La régression linéaire est une technique statistique couramment utilisée dans le domaine de la synthèse de données et de la prévision. Cette technique établit une ligne droite ou une surface afin de réduire les écarts entre les valeurs de sortie prévues et observées.



La procédure Modèles linéaires généralisés développe le modèle linéaire général de sorte que la variable dépendante soit linéairement reliée aux facteurs et covariables via une fonction de lien précise. En outre, le modèle permet à la variable dépendante de suivre une distribution non normale. Il couvre les fonctionnalités d'un grand nombre de modèles statistiques, notamment le modèle de régression linéaire, le modèle de régression logistique, le modèle log-linéaire pour les données d'effectif et le modèle de survie avec censure par intervalle. [Pour plus d'informations, reportez-vous à la section Noeud Modèles linéaires généralisés dans le chapitre 10 sur p. 327.](#)



Le noeud k -Voisin le plus proche (KNN) associe une nouvelle observation à la catégorie ou à la valeur des objets k les plus proches dans l'espace du prédicteur, où k est un entier. Les observations semblables sont proches l'une de l'autre et les observations dissemblables sont éloignées l'une de l'autre. [Pour plus d'informations, reportez-vous à la section Noeud KNN dans le chapitre 16 sur p. 507.](#)



Le noeud Support Vector Machine (SVM) vous permet de classer les données dans l'un de deux groupes sans sur-ajustement. SVM fonctionne bien avec les grands ensembles de données, comme ceux qui disposent d'un très grand nombre de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud SVM dans le chapitre 15 sur p. 501.](#)

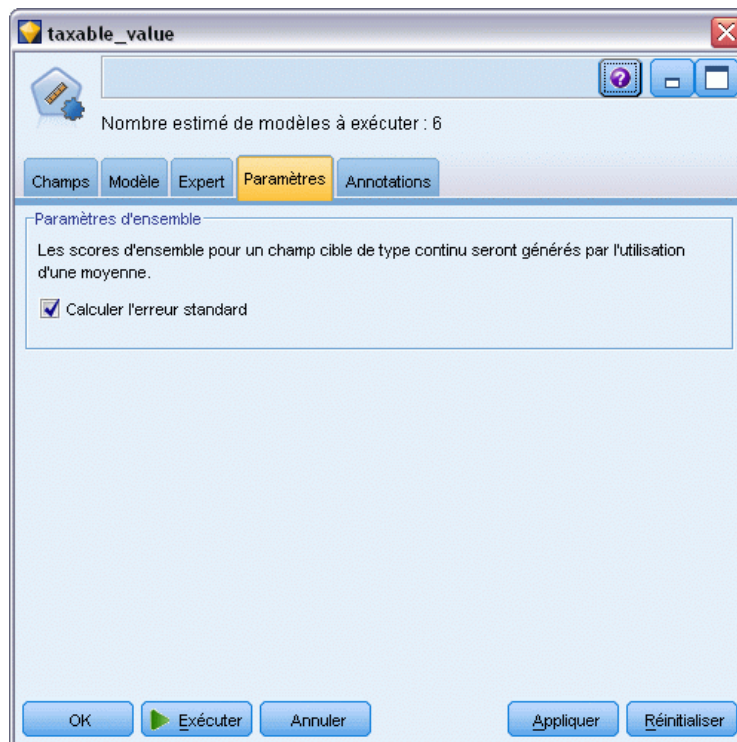


Les modèles de régression linéaire prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites. [Pour plus d'informations, reportez-vous à la section Modèles linéaires dans le chapitre 10 sur p. 269.](#)

Noeud Numérisation automatique - Options des paramètres

L'onglet Paramètres du noeud Numérisation automatique vous permet de préconfigurer les options de temps de score disponibles sur le nugget.

Figure 5-11
Noeud Numérisation automatique : onglet Paramètres



Calculer l'erreur standard. Pour une cible continue (intervalle numérique), un calcul d'erreur standard est exécuté par défaut pour calculer la différence entre les valeurs mesurées ou estimées et les valeurs réelles, et pour montrer la correspondance proche de ces évaluations.

Noeud Classification automatique

Le noeud Classification automatique évalue et compare les modèles de classification identifiant des groupes d'enregistrement ayant des caractéristiques similaires. Le noeud fonctionne de la même manière que les autres noeuds de modélisation automatisés, vous permettant de tester plusieurs combinaisons d'options en une seule modélisation. Les modèles peuvent être comparés à l'aide de mesures de bases permettant d'essayer de filtrer et de classer l'utilité des modèles de classification et de fournir une mesure en fonction de l'importance de champs particuliers.

Les modèles de classification sont souvent utilisés pour identifier les groupes pouvant être utilisés comme entrées dans des analyses ultérieures. Par exemple, vous voulez peut-être cibler des groupes de clients en fonction de caractéristiques démographiques telles que le revenu, ou des services qu'ils ont acheté dans le passé. Ceci peut être effectué sans connaître auparavant les groupes et leurs caractéristiques — vous ne savez peut-être pas combien de groupes rechercher ou quelles fonctions utiliser dans leur définition. Les modèles de classification sont souvent appelés modèles d'apprentissage non supervisé car ils n'utilisent pas de champ cible et ne renvoient pas de prédiction spécifique pouvant être évaluée comme vraie ou fausse. La valeur d'un modèle de classification est déterminée par sa capacité à capturer des groupements intéressants dans les données et à fournir des descriptions utiles de ces mêmes groupements. [Pour plus d'informations, reportez-vous à la section Modèles de classification dans le chapitre 11 sur p. 382.](#)

Figure 5-12
Résultats de la classification automatique

Utiliser ?	Graphiques	Modèle	Durée de création	Silhouette	Nombre de classes	Plus petite classe (N)	Plus petite classe (%)	Plus grande classe (N)	Plus grande classe (%)	Plus petite/Plus grande	Importance
<input checked="" type="checkbox"/>		K-... < 1	< 1	0,229	5	137	12	372	32	0,368	1
<input type="checkbox"/>		Tw... < 1	< 1	0,229	7	62	5	271	23	0,229	1
<input type="checkbox"/>		Ko... < 1	< 1	0,206	9	6	0	285	25	0,021	1

Conditions requises. Un ou plusieurs champs définissant des caractéristiques intéressantes. Les modèles de classification n'utilisent pas de champs cibles de la même manière que d'autres modèles car ils ne font pas de prédictions spécifiques pouvant être évaluées comme vraies ou fausses. Ils sont plutôt utilisés pour identifier des groupes d'observations pouvant être liés. Par exemple, vous ne pouvez pas utiliser un modèle de classification pour prédire si un client donné refusera ou répondra à une offre. Mais vous pouvez utiliser un modèle de classification pour attribuer des clients à des groupes en fonction de leur tendance à faire cela. Les champs de pondération et d'effectifs ne sont pas utilisés.

Champs d'évaluation. Quand aucune cible n'est utilisée, vous pouvez facultativement spécifier un ou plusieurs champs d'évaluation à utiliser pour comparer les modèles. L'utilité d'un modèle de classification peut être évaluée en mesurant la qualité (ou la mauvaise qualité) avec laquelle les classes différencient ces champs.

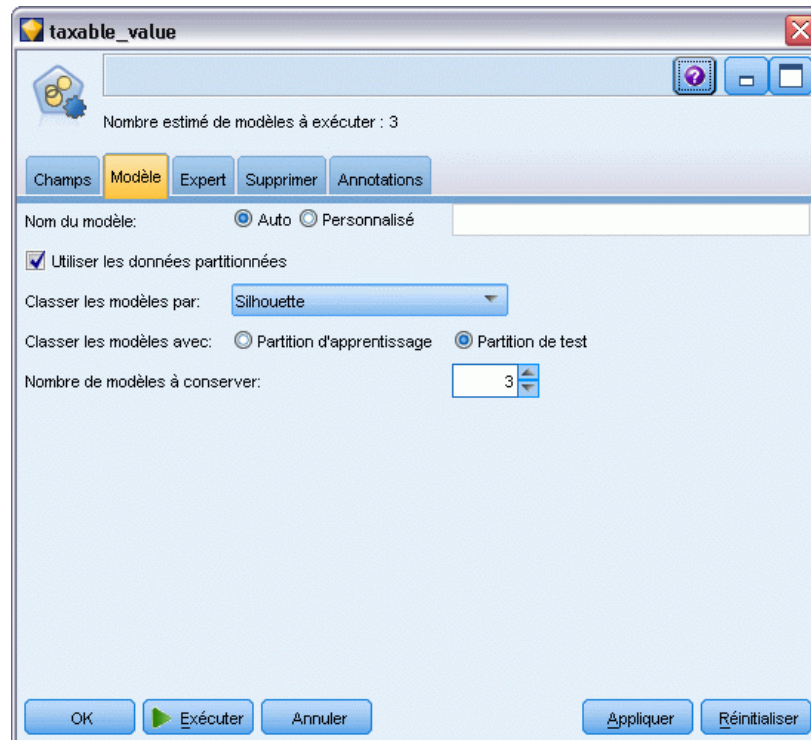
Types de modèle pris en charge

Les types de modèles pris en charge comprennent TwoStep, K-Means et Kohonen.

Noeud Classification - Options du modèle

L'onglet Modèle du noeud Classification automatique vous permet de préciser le nombre de modèles à enregistrer, ainsi que les critères utilisés pour comparer les modèles.

Figure 5-13
Noeud de classification automatique : 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.](#)

Classer les modèles par. Indique les critères utilisés pour comparer et classer des modèles.

- **Silhouette.** Index mesurant la cohésion et la séparation des classes. Reportez-vous à *Mesure de classement de la silhouette* ci-dessous pour plus d'informations.
- **Nombre de classes.** Le nombre de classes utilisées dans le modèle.
- **Taille de la plus petite classe.** La plus petite taille de classe.
- **Taille de la plus grande classe.** La plus grande taille de classe.
- **Plus petite / plus grande classe.** Rapport de la taille de la plus petite classe sur la plus grande classe.
- **Importance.** L'importance du champ Evaluation sur l'onglet Champs. Remarque : cela ne peut être calculé que si un champ Evaluation a été spécifié.

Classer les modèles avec. Lorsque vous utilisez une partition, vous pouvez indiquer si les rangs doivent être basés sur l'ensemble de données d'apprentissage ou sur l'ensemble de test. Lorsque les ensembles de données sont volumineux, utiliser une partition pour le filtrage préliminaire des modèles permet d'améliorer considérablement les performances.

Nombre de modèles à conserver. Indique le nombre maximal de modèles à répertorier dans le nugget généré par le noeud. Les modèles situés en tête de classement seront répertoriés conformément au critère de classement indiqué. Les performances peuvent être moindres si vous augmentez cette limite. La valeur maximum autorisée est 100.

Mesure de classement de la silhouette

La mesure de classement par défaut, Silhouette, possède une valeur par défaut de 0 car une valeur inférieure à 0 (c.à.d. négative) indique que la distance moyenne entre une observation et des points dans sa classe attribuée est supérieure à la distance moyenne minimale aux points dans une autre classe. Par conséquent, des modèles ayant une silhouette négative peuvent être supprimés en toute sécurité.

La mesure du classement est en réalité un coefficient de silhouette modifié, qui combine les concepts de cohésion de classe (favorisant des modèles qui contiennent des classes étroitement cohérentes) et de séparation de classe (favorisant des modèles qui contiennent des classes largement séparées). Le coefficient de silhouette moyen représente simplement la moyenne de toutes les observations du calcul suivant pour chaque observation individuelle :

$$(B - A) / \max(A, B)$$

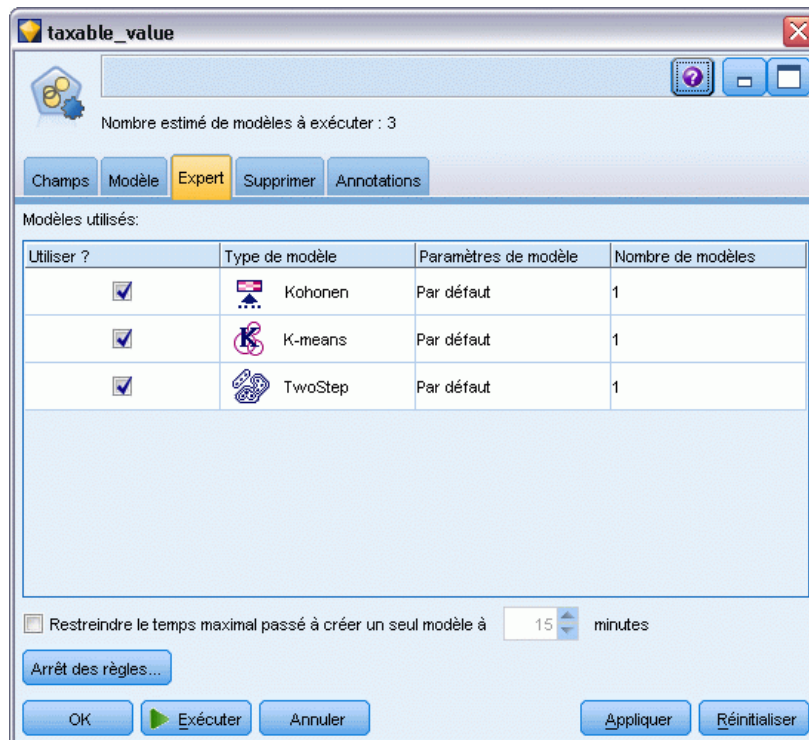
où A représente la distance de l'observation au centre de gravité de la classe à laquelle l'observation appartient ; et où B représente la distance minimale de l'observation au centre de gravité de chacune des autres classes.

Le coefficient de silhouette (et sa moyenne) est compris entre -1 (indiquant un modèle médiocre) et 1 (indiquant un excellent modèle). La moyenne peut être effectuée au niveau de la totalité des observations (ce qui donne la silhouette complète) ou au niveau des classes (ce qui donne la silhouette de classe). Les distances peuvent être calculées à l'aide des distances euclidiennes.

Noeud Classification - Options expert

L'onglet Expert du noeud Classification automatique vous permet d'appliquer une partition (si elle existe), de sélectionner les algorithmes à utiliser et de définir les règles d'arrêt.

Figure 5-14
Noeud de classification automatique : Onglet Expert



Modèles utilisés. Dans la colonne de gauche, cochez les cases correspondant aux types de modèle (algorithmes) à inclure dans la comparaison. Plus vous sélectionnez de types, plus un nombre important de modèles sont créés et plus le traitement est long.

Type de modèle. Répertorie les algorithmes disponibles (voir ci-dessous).

Paramètres de modèle. Vous pouvez utiliser les paramètres par défaut pour chaque type de modèle ou sélectionner Spécifier pour choisir des options pour chaque type de modèle. Les options sont semblables à celles proposées pour chaque noeud de modélisation. Cependant, vous pouvez sélectionner ici plusieurs options ou combinaisons. Par exemple, si vous comparez des modèles R. neurones, vous pouvez choisir toutes les méthodes d'apprentissage pour apprendre six modèles en une seule étape au lieu de sélectionner chaque méthode indépendamment.

Nombre de modèles créés. Indique le nombre de modèles générés pour chaque algorithme en fonction des paramètres actuels. Lorsque vous combinez des options, le nombre de modèles peut augmenter rapidement. Il est donc fortement recommandé de surveiller ce nombre, en particulier si vous utilisez des ensembles de données volumineux.

Restreindre le temps maximal passé à créer un seul modèle. (Pour les modèles K-Means, Kohonen, TwoStep, SVM, KNN, Bayes Net et Liste de décision) Définit une limite de temps maximale pour un modèle donné. Par exemple, si l'apprentissage d'un modèle donné prenait un temps particulièrement long du fait d'une interaction complexe, vous ne voudriez pas qu'il ralentisse l'exécution de la modélisation complète.

Algorithmes pris en charge



Le noeud K-Means classe l'ensemble de données dans différents groupes (ou classes). La méthode définit un nombre de classes fixe, affecte à plusieurs reprises des enregistrements à des classes et ajuste les centres de classe, jusqu'à ce que le modèle ne puisse plus être amélioré. Au lieu de tenter de prédire un résultat, le modèle *k*-means utilise un processus connu sous le nom d'apprentissage non supervisé pour découvrir des tendances dans l'ensemble de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud K-Means dans le chapitre 11 sur p. 390.](#)



Le noeud Kohonen génère un type de réseau de neurones qui peut être utilisé pour classer les données en groupes distincts. Lorsque l'apprentissage du réseau est terminé, les enregistrements similaires doivent être regroupés dans la connexion de sortie, tandis que les enregistrements différents sont à l'opposé. Vous pouvez étudier le nombre d'observations capturées par chaque unité du nugget de modèle afin d'identifier les unités fortes. Vous pouvez ainsi vous faire une idée du nombre de classes approprié. [Pour plus d'informations, reportez-vous à la section Noeud Kohonen dans le chapitre 11 sur p. 383.](#)

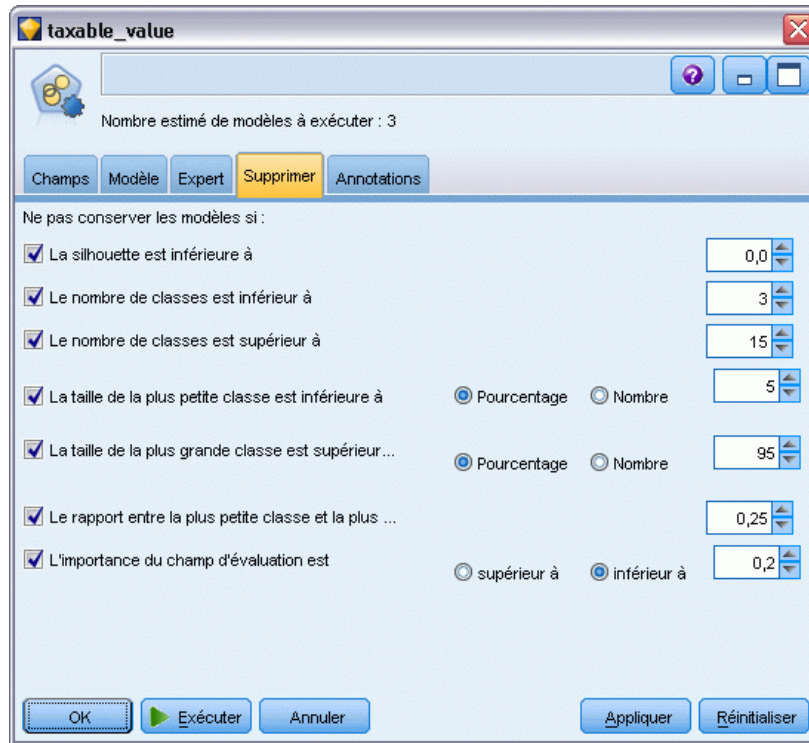


Le noeud TwoStep utilise une méthode de classification non supervisée en deux étapes. La première étape consiste en une exploration des données visant à compresser les données d'entrée brutes en sous-classes plus faciles à manipuler. Au cours de la seconde étape, l'utilisation d'une méthode de classification hiérarchique permet de fusionner progressivement les sous-classes en classes de plus en plus importantes. La technique TwoStep a l'avantage d'évaluer automatiquement le nombre de classes optimal pour les données d'apprentissage. Il peut prendre en charge de manière efficace des types de champ mixtes et des ensembles de données volumineux. [Pour plus d'informations, reportez-vous à la section Noeud Classification TwoStep dans le chapitre 11 sur p. 395.](#)

Noeud Classification - Options de suppression

L'onglet Supprimer du noeud Classification automatique vous permet de supprimer automatiquement les modèles qui ne répondent pas à certains critères. Ces modèles ne sont pas répertoriés dans le nugget de modèle.

Figure 5-15
Noeud de classification automatique : Onglet Supprimer



Vous pouvez spécifier la valeur minimale de la silhouette, le nombre de classes, la taille des classes et l'importance du champ d'évaluation utilisé dans le modèle. La silhouette, le numéro et la taille des classes sont déterminés comme spécifié dans le noeud de modélisation. [Pour plus d'informations, reportez-vous à la section Noeud Classification - Options du modèle sur p. 118.](#)

Vous pouvez également configurer le noeud pour qu'il interrompe son exécution dès qu'un modèle remplissant tous les critères définis est généré. [Pour plus d'informations, reportez-vous à la section Noeuds de modélisation automatisés - Règles d'arrêt sur p. 98.](#)

Nuggets de modèle automatisés

Quand un noeud de modélisation automatisé est exécuté, le noeud évalue des modèles candidats pour toutes les combinaisons d'options possibles, classe chaque modèle candidat en fonction de la mesure que vous spécifiez et enregistre les meilleurs modèles dans un nugget de modèle automatisé composite. Ce nugget de modèle contient un ensemble d'un ou plusieurs modèles générés par le noeud, pouvant être parcourus ou sélectionnés individuellement pour une utilisation dans un scoring. Le type de modèle et le moment de création sont répertoriés pour chaque modèle, ainsi que plusieurs autres mesures adaptées au type de modèle. Vous pouvez trier le tableau en fonction de n'importe laquelle de ces colonnes pour identifier rapidement les modèles les plus intéressants.

Figure 5-16
Résultats de la numérisation automatique

Utiliser ?	Graphiques	Modèle	Durée de création (min)	Corrélation	Nbre champs utilisés	Erreur relative
<input checked="" type="checkbox"/>		Generaliz...	< 1	0,915	7	0,162
<input checked="" type="checkbox"/>		Regressi...	< 1	0,9	5	0,19
<input checked="" type="checkbox"/>		CHAID Tr...	< 1	0,892	5	0,204

- Pour rechercher l'un des nuggets de modélisation individuels, double-cliquez sur l'icône du nugget. Vous pouvez alors générer un noeud de modélisation pour ce modèle dans l'espace de travail de flux ou une copie du nugget de modèle dans la palette des modèles.
- Les graphiques en miniature proposent une évaluation visuelle rapide pour chaque type de modèle, comme résumé ci-dessous. Vous pouvez double-cliquer sur une miniature pour générer un graphique en grandeur nature. Le graphique grandeur nature peut afficher jusqu'à 1 000 points et se basera sur un échantillon si l'ensemble de données en contient davantage. (Dans le cas des diagrammes de dispersion uniquement, le graphique est régénéré à chaque fois qu'il s'affiche ; aussi, tout changement des données en amont—telles que la mise à jour d'un échantillon aléatoire si l'option Définir graine aléatoire n'est pas sélectionnée—pourra se refléter à chaque fois que le diagramme de dispersion est redessiné.)
- Utilisez la barre d'outils pour afficher ou masquer des colonnes précises dans l'onglet Modèle ou pour modifier la colonne de tri du tableau. (Pour modifier le critère de tri, vous pouvez également cliquer sur les en-têtes de colonne.)
- Utilisez le bouton Supprimer pour supprimer de façon permanente tous les modèles non utilisés.
- Pour réorganiser les colonnes, cliquez sur un en-tête de colonne puis faites glisser la colonne vers l'emplacement de votre choix.
- Si vous utilisez une partition, vous pouvez afficher les résultats de la partition d'apprentissage ou de test, selon le cas. [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.](#)

Les colonnes indiquées dépendent du type de modèle comparé, comme suit :

Cibles binaires

- Dans le cas des modèles binaires, le graphique en miniature affiche la distribution des valeurs réelles, superposées aux valeurs prédites, pour fournir une indication visuelle rapide du nombre d'enregistrements prédits correctement dans chacune des catégories.
- Les critères de classement correspondent aux options dans le noeud de modélisation Classificateur automatique. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique - Options du modèle sur p. 101.](#)
- Pour le profit maximal, le centile où le maximum est atteint est également signalé.
- Pour les graphiques Lift cumulatifs, vous pouvez modifier le centile sélectionné via la barre d'outils.

Cibles nominales

- Dans le cas des modèles nominaux (ensemble), le graphique en miniature affiche la distribution des valeurs réelles, superposées aux valeurs prédites, pour fournir une indication visuelle rapide du nombre d'enregistrements prédits correctement dans chacune des catégories.
- Les critères de classement correspondent aux options dans le noeud de modélisation Classificateur automatique. [Pour plus d'informations, reportez-vous à la section Noeud Classificateur automatique - Options du modèle sur p. 101.](#)

Cibles continues

- Dans le cas des modèles continus (intervalle numérique), les graphiques prédits par rapport aux valeurs observées de chaque modèle fournissent une indication visuelle rapide de leurs corrélation. Pour un modèle performant, les points devraient être regroupés le long de la diagonale plutôt que dispersés de façon aléatoire dans le graphique.
- Les critères de classement correspondent aux options dans le noeud de modélisation Numérisation automatique. [Pour plus d'informations, reportez-vous à la section Noeud Numérisation automatique - Options du modèle sur p. 111.](#)

Cibles de classe

- Dans le cas de modèles de classe, les graphiques sont comptés par rapport aux classes de chaque modèle, ce qui fournit un aperçu visuel rapide de la distribution de la classe.
- Les critères de classement correspondent aux options dans le noeud de modélisation Classification automatique. [Pour plus d'informations, reportez-vous à la section Noeud Classification - Options du modèle sur p. 118.](#)

Sélection de modèles pour le scoring

La colonne Utiliser ? vous permet de sélectionner des modèles à utiliser dans le scoring.

- Pour les cibles binaires, nominales et numériques, vous pouvez sélectionner plusieurs modèles de scoring et combiner les scores dans un nugget de modèle sous forme d'ensemble unique. En combinant des prévisions à partir de différents modèles, vous pouvez éviter les limites des

modèles individuels, vous permettant ainsi d'obtenir une plus grande précision globale de chacun des modèles.

- Pour les modèles de classification, un seul modèle de scoring peut être sélectionné à la fois. Par défaut, le premier du classement est sélectionné en premier.

Génération de noeuds et de modèles

Vous pouvez générer une copie du nugget de modèle automatisé composite ou du noeud de modélisation automatisé à partir duquel il a été créé. Par exemple, cela peut s'avérer utile si vous ne disposez pas du flux d'origine à partir duquel le nugget de modèle automatisé a été créé. Vous pouvez aussi générer un nugget ou un noeud de modélisation pour tous les modèles individuels répertoriés dans le nugget de modèle automatisé.

Nugget de modélisation automatisé

- ▶ Dans le menu Générer, sélectionnez Modèle vers palette pour ajouter le nugget de modèle automatisé à la palette Modèles. Vous pouvez enregistrer ou utiliser le modèle généré tel qu'il est, sans réexécuter le flux.
- ▶ Vous pouvez également sélectionner Générer un noeud de modélisation dans le menu Générer pour ajouter le noeud de modélisation à l'espace de travail de flux. Ce noeud permet de réestimer les modèles sélectionnés sans répéter toute l'exécution de la modélisation.

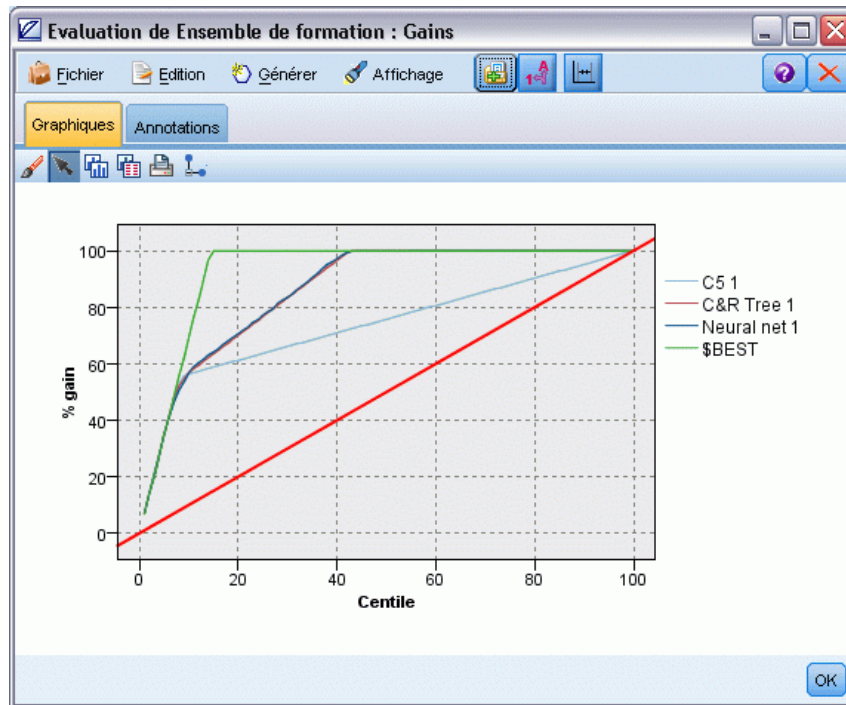
Nugget de modélisation individuel

- ▶ Dans le menu Modèle, double-cliquez sur le nugget individuel dont vous avez besoin. Une copie de ce nugget s'ouvre dans une nouvelle boîte de dialogue.
- ▶ Dans le menu Générer, sélectionnez Modèle vers palette pour ajouter le nugget de modélisation individuel à la palette Modèles.
- ▶ Vous pouvez également sélectionner Générer un noeud de modélisation dans le menu Générer dans la nouvelle boîte de dialogue pour ajouter le noeud de modélisation individuel à l'espace de travail de flux.

Génération de graphiques Evaluation

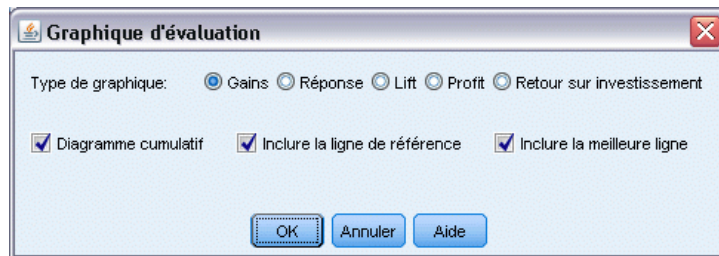
Pour les modèles binaires uniquement, vous pouvez générer des graphiques d'évaluation qui offrent un affichage visuel pour évaluer et comparer les performances de chaque modèle. Les graphiques d'évaluation ne sont pas disponibles pour les modèles générés à partir des noeuds Numérisation automatique ou Classification automatique. [Pour plus d'informations, reportez-vous à la section Noeud Evaluation dans le chapitre 5 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

Figure 5-17
Graphique de réponses (cumulatif) affichant la meilleure ligne et la ligne de référence



- Dans la colonne *Utiliser ?* du nugget de modèle automatisé Classificateur automatique, sélectionnez les modèles que vous souhaitez évaluer.
- Dans le menu Générer, sélectionnez Evaluation Chart(s).

Figure 5-18
Génération d'un graphique Evaluation



- Sélectionnez le type de graphique et les autres options souhaitées. [Pour plus d'informations, reportez-vous à la section Onglet Nuage d'évaluation dans le chapitre 5 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Graphiques d'évaluation

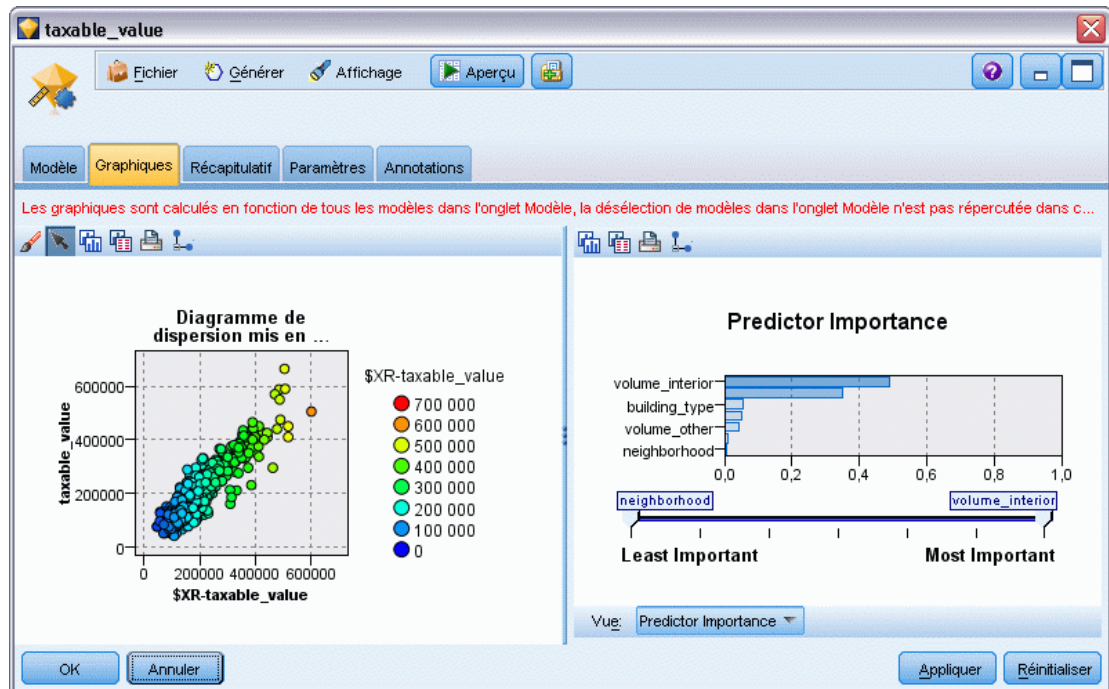
Dans l'onglet Modèle du nugget de modèle automatisé, vous pouvez faire défiler les résultats pour afficher des graphiques individuels pour chacun des modèles affichés. Pour les nuggets Classificateur automatique et Numérisation automatique, l'onglet Graphique affiche un graphique

et l'importance des valeurs prédites qui reflètent tous deux les résultats de tous les modèles combinés. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Pour Classificateur automatique, un graphique Proportion est affiché où des courbes (aussi appelées diagramme de dispersion) sont affichées pour la Numérisation automatique. [Pour plus d'informations, reportez-vous à la section Fonctions communes des noeuds Graphiques dans le chapitre 5 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Figure 5-19

Numérisation automatique - Graphique Courbes pour les modèles combinés au sein du nugget de modèle automatisé

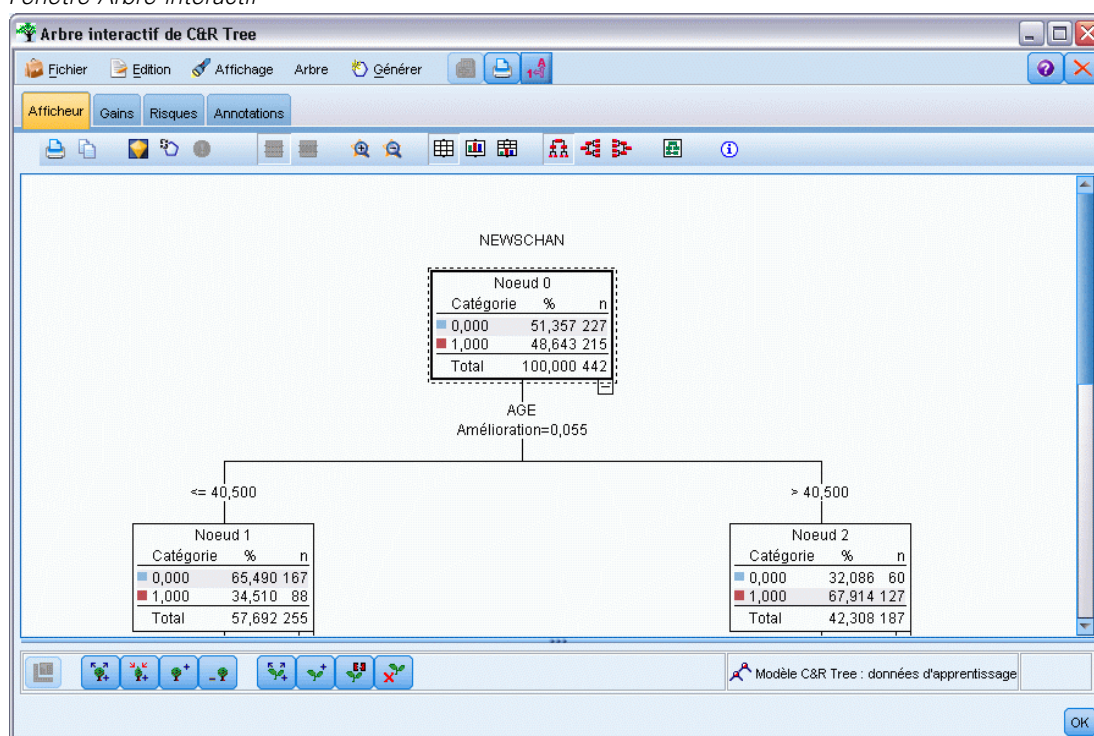


Arbres décision

Modèles d'arbre décision

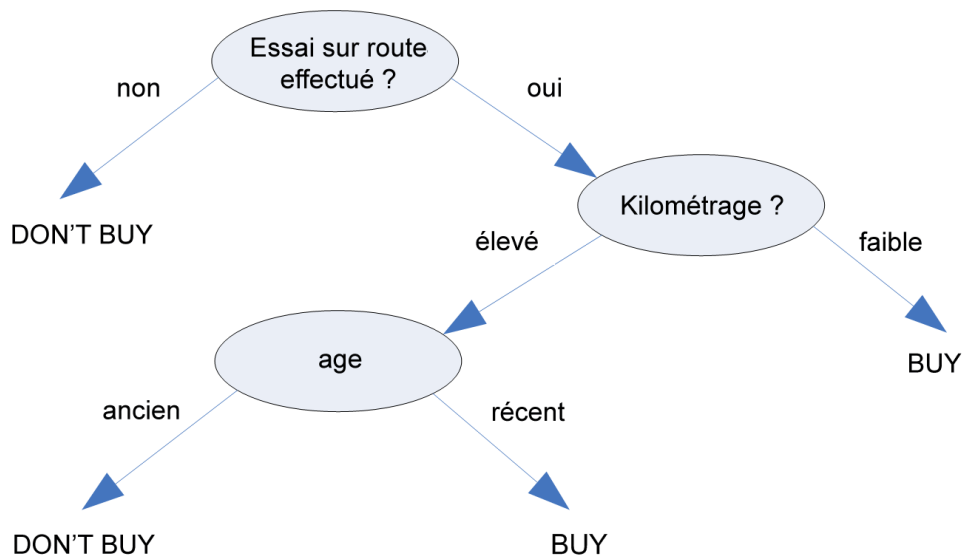
Les modèles d'arbre décision vous permettent de développer des systèmes de classification prévoyant ou classant les observations futures à partir d'un ensemble de règles de décision. Si vos données sont divisées en classes qui vous intéressent (par exemple, rapport prêts à haut risque, prêts à faible 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.

Figure 6-1
Fenêtre Arbre interactif



Cette approche, parfois nommée **induction de règle**, présente plusieurs avantages. Tout d'abord, le raisonnement qui sous-tend le modèle apparaît de façon très claire lorsque vous parcourez l'arbre, ce qui n'est pas le cas avec certaines méthodes de modélisation, qualifiées de « boîtes noires » dont il est parfois difficile de saisir la logique interne.

Figure 6-2
Arbre décision simple pour l'achat d'une voiture



En outre, le processus inclut automatiquement dans sa règle les attributs ayant une influence sur la prise de décision. Les autres attributs sont alors ignorés. Vous pouvez ainsi obtenir des informations très utiles sur les données et les utiliser pour que les champs pertinents apparaissent avant de vous familiariser avec une autre technique d'apprentissage, telle que le réseau de neurones.

Les nuggets de modèle d'arbre décision peuvent être convertis en une suite de règles If-Then (un **ensemble de règles**), permettant souvent d'afficher les informations de manière plus intelligible. La présentation de l'arbre décision permet de visualiser la façon dont les attributs dans les données peuvent **fractionner** ou **partitionner** la population en sous-ensembles pertinents. La présentation de l'ensemble de règles est utile pour voir comment un groupe d'éléments particulier aboutit à une conclusion spécifique. Par exemple, la règle suivante fournit le **profil** d'un groupe de voitures de bon rapport qualité-prix :

```

IF tested = 'yes'
AND mileage = 'low'
THEN -> 'BUY'.
  
```

Algorithmes de création d'arbre

Quatre algorithmes permettent d'analyser la classification et l'analyse de segmentation. Ils procèdent tous quasiment de la même façon : ils examinent tous les champs de l'ensemble de données pour rechercher celui qui aboutit à la meilleure classification ou prévision lors de la division des données en sous-groupes. Le processus est appliqué de manière récursive : les sous-groupes sont divisés en unités de plus en plus petites jusqu'à ce que l'arbre soit terminé (en fonction de critères d'arrêt donnés). En fonction de l'algorithme employé, les champs d'entrée et les champs cible utilisés pour créer l'arbre peuvent être des champs continus (intervalle

numérique) ou des champs catégoriels. Si une cible continue est utilisée, un arbre de régression est généré ; s'il s'agit d'une cible catégorielle, un arbre de classification est créé.



Le noeud Arbre Classification et Regression, (C&R) Tree, génère un arbre décision qui vous permet de prévoir ou de classifier les observations futures. La méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments en réduisant l'index d'impureté à chaque étape, un noeud de l'arbre étant considéré comme "pur" si 100 % de ses observations appartiennent à une catégorie spécifique du champ cible. Les champs cible et les champs d'entrée peuvent être des champs d'intervalle numériques ou des champs catégoriels numériques (nominal, ordinal ou booléen). Toutes les divisions sont binaires (deux sous-groupes uniquement). [Pour plus d'informations, reportez-vous à la section Noeud Arbre C&RT sur p. 158.](#)



Le noeud CHAID génère des arbres décision à l'aide des statistiques du Chi-deux pour identifier les séparations optimales. Contrairement aux noeuds C&R Tree et QUEST, CHAID peut générer des arbres non binaires, ce qui implique que certaines divisions possèdent plusieurs branches. Les champs cibles et les champs d'entrée peuvent être d'intervalle numérique (continu) ou catégoriels. La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en détail toutes les divisions possibles, mais dont les calculs sont plus longs. [Pour plus d'informations, reportez-vous à la section Noeud CHAID sur p. 159.](#)



Le noeud QUEST est une méthode de classification supervisée binaire permettant de créer des arbres décision, développée pour réduire le temps de traitement nécessaire aux analyses C&R Tree importantes, tout en limitant la tendance, observée parmi les méthodes d'arbre de classification, à favoriser les entrées autorisant un nombre supérieur de divisions. Les champs d'entrée peuvent être des intervalles numériques (continues) mais les champs cible doivent être catégoriels. Toutes les divisions sont binaires. [Pour plus d'informations, reportez-vous à la section Noeud QUEST sur p. 160.](#)



Le noeud C5.0 crée un arbre décision ou un ensemble de règles. Le fonctionnement de ce modèle repose sur un découpage de l'échantillon basé sur le champ qui fournit le gain d'informations le plus important à chaque niveau. Le champ cible doit être catégoriel. Les divisions multiples en plus de deux sous-groupes sont autorisées. [Pour plus d'informations, reportez-vous à la section Noeud C5.0 sur p. 180.](#)

Emplois généraux de l'analyse en arbre

Voici quelques exemples généraux d'emploi de l'analyse en arbre :

Segmentation. Permet d'identifier les personnes susceptibles d'être membres d'une classe spécifique.

Stratification : Permet d'attribuer les observations à une ou plusieurs catégories, comme par exemple les groupes à risque faible, moyen et élevé.

Prévision. Permet de créer des règles et de les utiliser pour prévoir les événements ultérieurs. La prédiction peut également concerner des tentatives pour relier les attributs de prévision aux valeurs d'une variable continue.

Factorisation et filtrage des variables. Permet de sélectionner un sous-ensemble utile de valeurs prédites à partir d'un ensemble volumineux de variables, afin de l'utiliser pour élaborer un modèle paramétrique formel.

Identification d'interaction. Permet d'identifier les relations qui appartiennent uniquement à certains sous-groupes et de les définir dans un modèle paramétrique formel.

Fusion de catégories et variables continues en bandes. Permet de regrouper des catégories de variables prédites et des variables continues avec une perte d'information minimale.

Générateur d'arbres interactifs

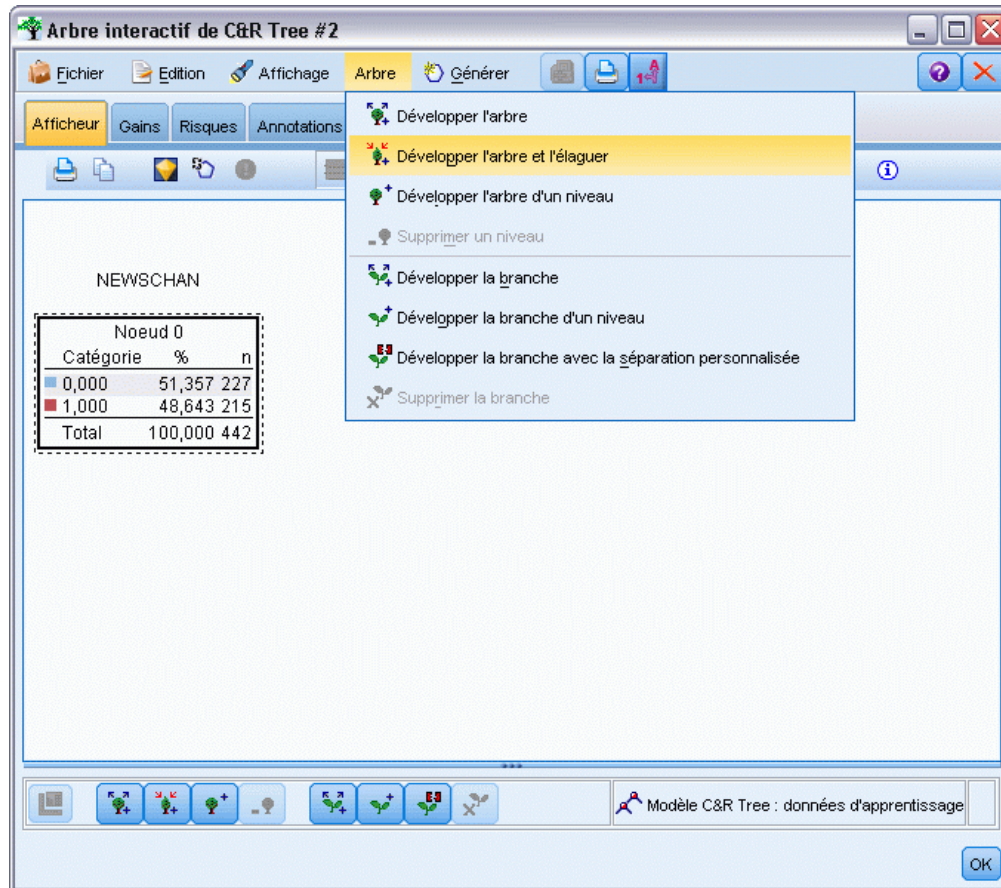
Vous pouvez créer automatiquement un modèle d'arbre, ce qui permet à l'algorithme de choisir la division optimale à chaque niveau. Vous avez également la possibilité d'utiliser le générateur d'arbres interactifs pour prendre le contrôle et mettre en pratique votre connaissance du marché, et pour affiner ou simplifier l'arbre avant d'enregistrer le nugget de modèle.

- ▶ Générez un flux et ajoutez l'un des noeuds Arbre décision C&R, CHAID ou QUEST.

Remarque : La création d'arbres interactifs n'est pas prise en charge pour les arbres C5.0.

- ▶ Ouvrez le noeud et, dans l'onglet Champs, sélectionnez des champs cible et des champs variables indépendantes, puis spécifiez des options de modèle supplémentaires, si nécessaire. Pour obtenir des instructions spécifiques, reportez-vous à la documentation de chaque noeud de création d'arbre.
- ▶ Dans le volet Objectifs de l'onglet Options de création, sélectionnez Lancer une session interactive.
- ▶ Cliquez sur Exécuter afin de lancer le générateur d'arbres.

Figure 6-3
Fenêtre Générateur d'arbre interactif



L'arbre actuel, dont le nœud racine occupe la première position, apparaît. Avant de générer des modèles, vous pouvez modifier et élaguer cet arbre niveau par niveau, et accéder aux gains, aux risques et aux informations liées.

Commentaires

- Grâce aux nœuds Arbre C&RT, CHAID et QUEST, les champs ordinaux utilisés dans le modèle concerné doivent disposer d'un stockage numérique (et non d'une chaîne). Si nécessaire, vous pouvez utiliser le nœud Recoder pour les convertir. [Pour plus d'informations, reportez-vous à la section Nœud Recoder dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

- Si vous le souhaitez, vous pouvez vous servir d'un champ de partition pour diviser les données en échantillons d'apprentissage et de test. [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*.](#)
- Au lieu d'utiliser le générateur d'arbres, vous pouvez également générer un modèle directement depuis le noeud de modélisation comme pour les autres modèles IBM® SPSS® Modeler. [Pour plus d'informations, reportez-vous à la section Création directe d'un modèle d'arbre sur p. 155.](#)

Développement et élagage de l'arbre

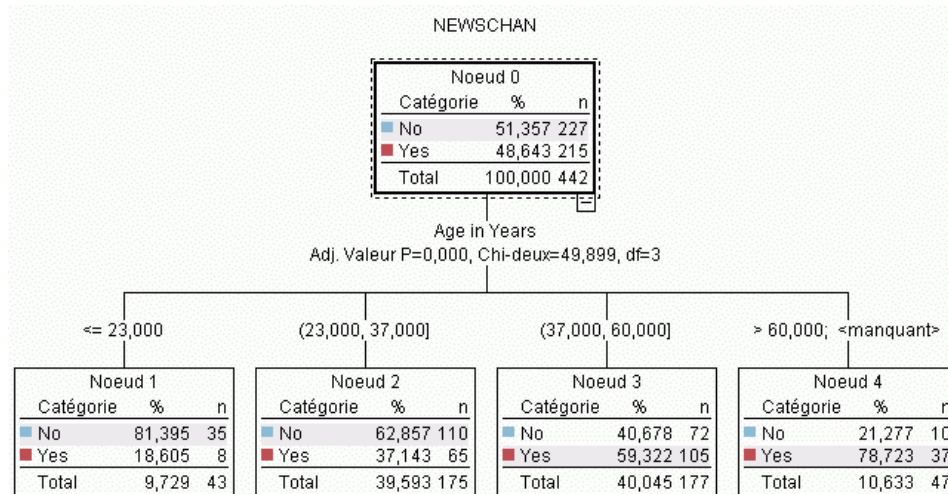
L'onglet Afficheur du générateur d'arbres contient l'arbre actuel qui commence au noeud racine.

- ▶ Pour développer l'arbre, choisissez les options suivantes :
Arbre > Développer l'arbre

Le système crée l'arbre en divisant chaque branche de façon récursive jusqu'à ce qu'un ou plusieurs critères d'arrêt soient remplis. A chaque division, la variable indépendante optimale est sélectionnée automatiquement selon la méthode de modélisation utilisée.
- ▶ Vous pouvez également choisir Développer l'arbre d'un niveau afin d'ajouter un niveau.
- ▶ Pour ajouter une branche sous un nœud, sélectionnez-le, puis choisissez Développer la branche.
- ▶ Pour choisir la variable indépendante utilisée pour une division, sélectionnez le nœud souhaité, puis Développer la branche avec la séparation personnalisée. [Pour plus d'informations, reportez-vous à la section Définition de divisions personnalisées sur p. 134.](#)
- ▶ Pour élaguer une branche, sélectionnez un noeud, puis Supprimer la branche en vue de supprimer ce noeud.
- ▶ Pour supprimer le niveau inférieur de l'arbre, sélectionnez Supprimer un niveau.
- ▶ Pour les arbres C&R et QUEST uniquement, sélectionnez Développer l'arbre et l'élaguer afin d'élaguer l'arbre à partir d'un algorithme de complexité des coûts. Cet algorithme ajuste l'évaluation des risques en fonction du nombre de noeuds terminaux, ce qui permet généralement d'obtenir un arbre plus simple. [Pour plus d'informations, reportez-vous à la section Noeud Arbre C&RT sur p. 158.](#)

Lire les règles de division dans l'onglet Afficheur

Figure 6-4
Règles de division affichées dans l'onglet Afficheur



Lorsque vous affichez des règles de division dans l'onglet Afficheur, les crochets signifient que la valeur voisine est comprise dans l'intervalle alors que les parenthèses indiquent que la valeur voisine est exclue de l'intervalle. L'expression (23,37] signifie par conséquent de 23 non compris à 37 compris ; autrement dit, de la valeur immédiatement supérieure à 23 à 37. Dans l'onglet Modèle, la même condition serait affichée ainsi :

Age > 23 and Age <= 37

Interruption du développement d'un arbre. Pour interrompre le développement d'un arbre (si l'opération s'avère plus longue que prévu, par exemple), cliquez sur le bouton Arrêter l'exécution dans la barre d'outils.

Figure 6-5
Bouton Arrêter l'exécution



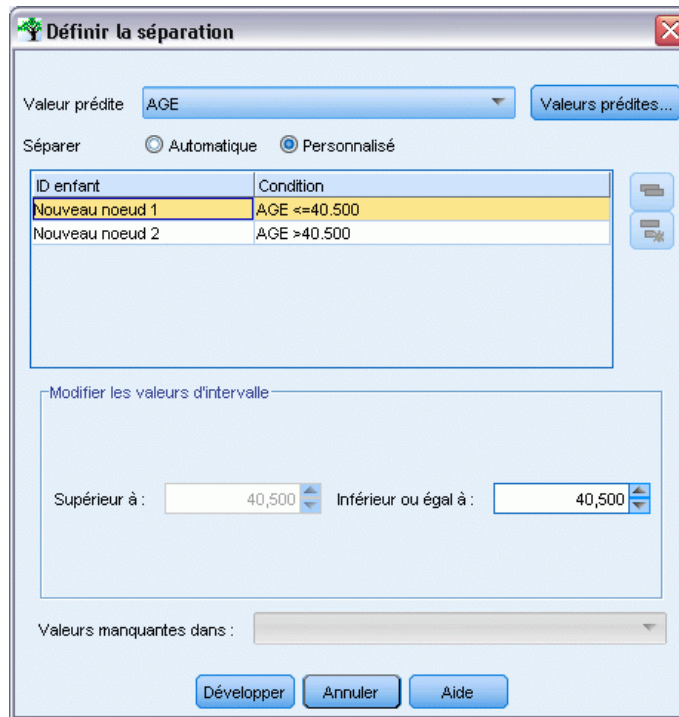
Le bouton n'est activé que pendant le développement de l'arbre. L'opération de développement en cours cesse au stade d'exécution où elle se trouve. Les noeuds déjà ajoutés sont conservés, les modifications apportées ne sont pas enregistrées ou la fenêtre n'est pas fermée. Le générateur d'arbres reste ouvert, ce qui vous permet de générer un modèle, de mettre à jour les directives ou d'exporter les résultats au format adéquat, au besoin.

Définition de divisions personnalisées

La boîte de dialogue Définir la séparation permet de sélectionner la variable indépendante et d'indiquer les conditions de chaque division.

- Dans le générateur d'arbres, sélectionnez un nœud dans l'onglet Afficheur puis, dans les menus proposés, choisissez l'une des options suivantes :
Arbre > Développer la branche avec la séparation personnalisée

Figure 6-6
Boîte de dialogue Définir la séparation



- Sélectionnez la variable indépendante de votre choix dans la liste déroulante ou cliquez sur le bouton Variables indépendantes pour visualiser les détails de chaque variable indépendante. [Pour plus d'informations, reportez-vous à la section Visualisation des détails d'une variable indépendante sur p. 136.](#)
- Vous pouvez accepter les conditions par défaut de chaque division ou sélectionner Personnalisé pour définir les conditions de la division, selon vos besoins.
 - Pour les variables indépendantes continues (intervalles numériques), vous pouvez utiliser les champs Modifier les valeurs d'intervalle pour indiquer l'intervalle de valeurs incluses dans chaque nouveau nœud.
 - Pour les variables indépendantes catégorielles, vous pouvez utiliser les champs Modifier les valeurs définies ou Modifier les valeurs ordinales pour indiquer les valeurs (ou l'intervalle de valeurs si une variable indépendante ordinaire est utilisée) correspondant à chaque nouveau nœud.
- Sélectionnez Développer afin de redévelopper la branche via la variable indépendante sélectionnée.
En général, l'arbre peut être divisé à l'aide d'une variable indépendante, quelles que soient les règles d'arrêt. Deux cas dérogent à cette règle, à savoir lorsque le nœud est pur (c'est-à-dire que 100 % des observations sont incluses dans la même catégorie cible et qu'il ne reste aucun

élément à diviser) ou que la variable indépendante sélectionnée est constante (il n'existe aucune structure de division).


Valeurs manquantes vers. Pour les arbres CHAID uniquement, si des valeurs manquantes sont disponibles pour une variable indépendante donnée, vous pouvez définir une division personnalisée afin d'attribuer ces valeurs à un nœud enfant. (Pour l'arbre C&RT et QUEST, les valeurs manquantes sont traitées via les substitutions définies dans l'algorithme. [Pour plus d'informations, reportez-vous à la section Détails et substitutions d'une division sur p. 137.](#))

Visualisation des détails d'une variable indépendante

La boîte de dialogue Sélectionner la variable indépendante affiche les statistiques relatives aux variables indépendantes disponibles (parfois appelées « concurrents ») qui peuvent être utilisées pour la division en cours.

Figure 6-7

Boîte de dialogue du nœud Sélectionner la variable indépendante



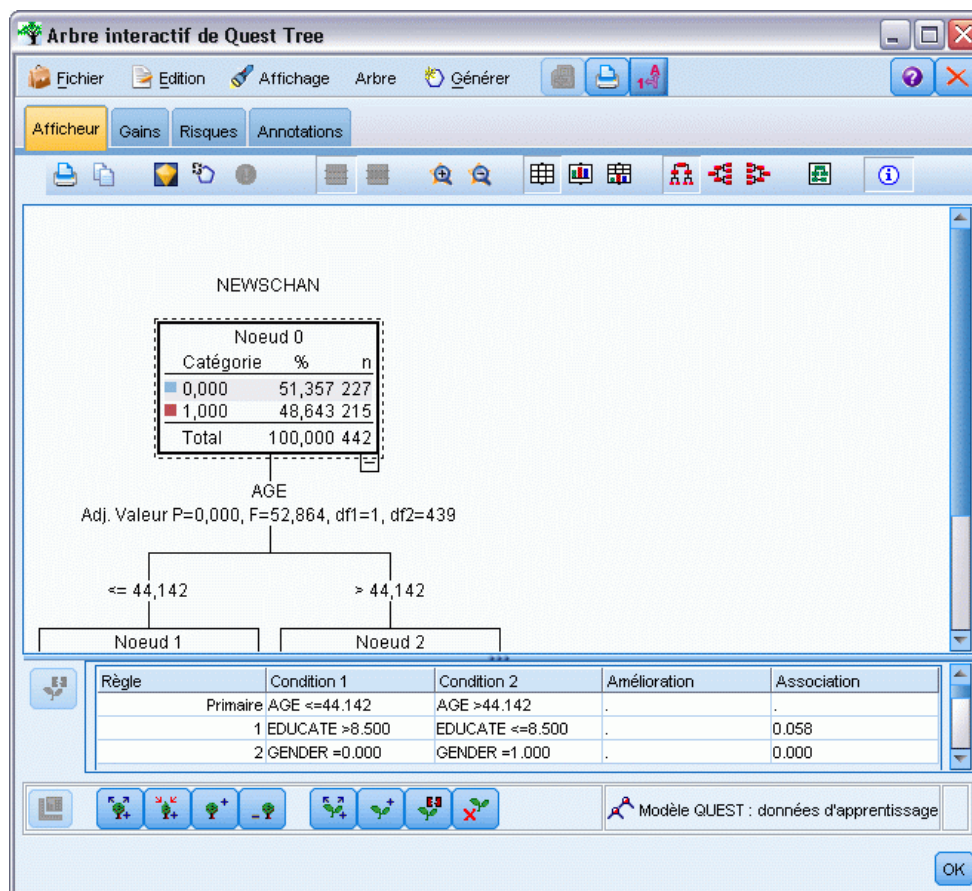
Valeur prédite	Noeuds	Amélioration
AGE	2	0.055
GENDER	2	0.011
INC	2	0.010
ORGS	2	0.004
EDUCATE	2	0.004
TVDAY	2	0.004
CHILDS		

- Pour les arbres CHAID et les arbres Exhaustive CHAID, les statistiques du Chi-deux sont fournies pour chaque variable indépendante catégorielle. Si une variable indépendante représente un intervalle numérique, la statistique F est affichée. Les statistiques du Chi-deux permettent de mesurer le degré d'indépendance du champ cible par rapport au champ de séparation. Une statistique de Chi-deux élevée est généralement liée à une probabilité faible, ce qui signifie que les deux champs sont probablement dépendants l'un de l'autre — indication que le champ de séparation est un champ correct. Des degrés de liberté sont également indiqués car ils prennent en compte le fait qu'il est plus simple pour un découpage en trois groupes d'avoir une statistique élevée et une probabilité faible que pour un découpage en deux groupes.
- Pour l'arbre C&RT et QUEST, l'amélioration de chaque variable indépendante apparaît. Plus cette amélioration est élevée, plus la réduction de l'impureté entre les nœuds parent et enfant est importante si cette variable indépendante est utilisée. (Un nœud pur est un nœud dont toutes les observations sont incluses dans une même catégorie cible. Plus l'arbre est pur, plus le modèle est adapté aux données.) En d'autres termes, un chiffre d'amélioration élevé indique généralement un découpage utile pour ce type d'arbre. La mesure d'impureté employée est spécifiée dans le nœud de création d'arbre.

Détails et substitutions d'une division

Vous pouvez sélectionner un noeud dans l'onglet Afficheur, puis cliquer sur le bouton d'affichage des informations de séparation à droite dans la barre d'outils pour visualiser les détails de la division de ce noeud. La règle de division utilisée apparaît, ainsi que les statistiques correspondantes. Pour les arbres catégoriels C&RT, l'amélioration et l'association sont affichées. L'association est une mesure de correspondance entre une substitution et le champ de division principal, la "meilleure" substitution étant celle qui est la plus proche du champ de division principal. Pour l'arbre C&RT et QUEST, sont également répertoriées les substitutions utilisées à la place de la variable indépendante principale.

Figure 6-8
Fenêtre du Générateur d'arbres interactif contenant les informations de division



- Pour modifier la division du noeud sélectionné, vous pouvez cliquer sur l'icône située à gauche du panneau de substitutions afin d'ouvrir la boîte de dialogue Définir la séparation. (Une méthode plus rapide consiste à choisir une substitution dans la liste avant de cliquer sur l'icône pour la sélectionner en tant que champ de division principal.)

Substitutions. Cet onglet affiche toutes les substitutions du champ de division principal pour le noeud sélectionné. Les substitutions sont des champs de remplacement utilisés si la valeur prédite principale d'un enregistrement déterminé est manquante. Le nombre de substitutions maximal autorisé pour une division donnée est indiqué dans le noeud de création de modèle, mais

dépend en fait des données d'apprentissage. En règle générale, plus il manque de données, plus les substitutions ont de chances d'être utilisées. Pour les autres modèles d'arbres décision, cet onglet est vide.

Remarque : Pour faire partie du modèle, les substitutions doivent être identifiées au cours de la phase d'apprentissage. Si l'échantillon d'apprentissage ne comporte aucune valeur manquante, aucune substitution n'est identifiée. Les enregistrements contenant des valeurs manquantes repérés au cours du test ou du scoring sont alors automatiquement placés dans le noeud enfant comptant le plus grand nombre d'enregistrements. Si vous prévoyez la détection de valeurs manquantes lors du test ou du scoring, veillez à ce que ces valeurs soient également manquantes dans l'échantillon d'apprentissage. Les substitutions ne sont pas disponibles avec les arbres CHAID.

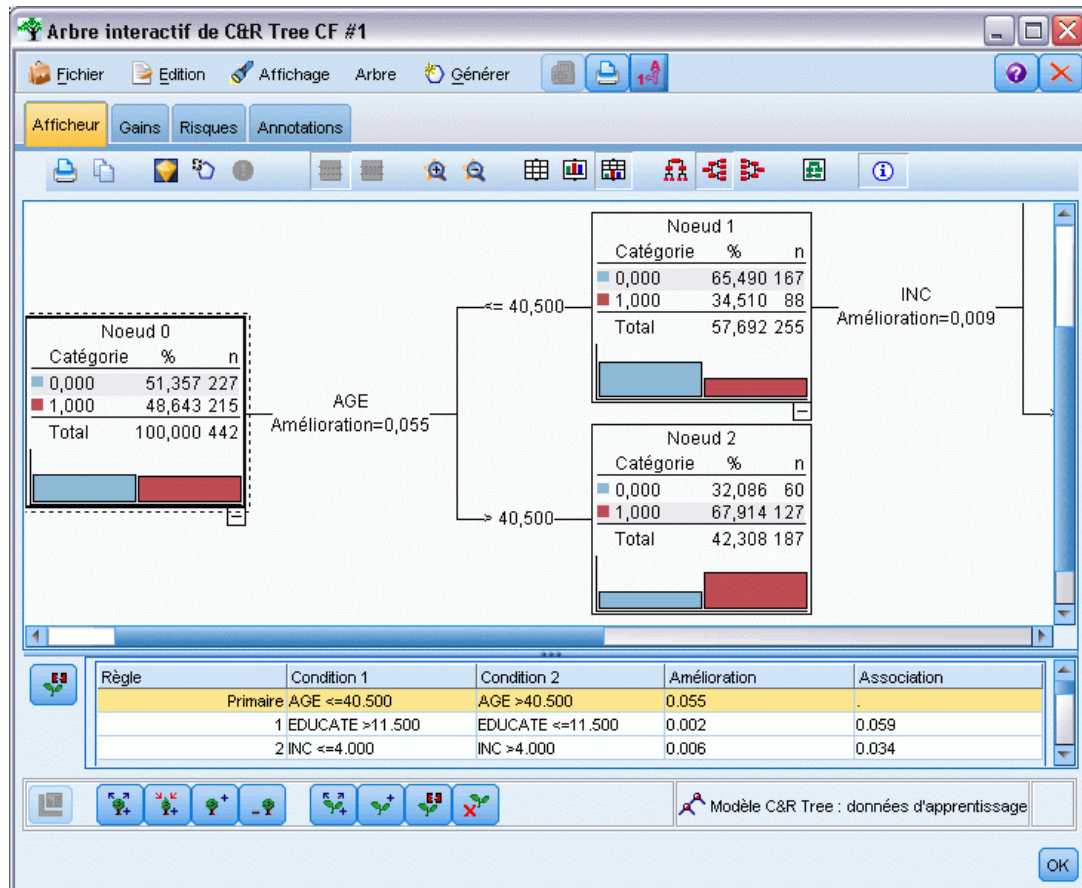
Aucune substitution n'est employée pour les arbres CHAID, mais lorsque vous définissez une division personnalisée, vous pouvez affecter des substitutions à un nœud enfant. [Pour plus d'informations, reportez-vous à la section Définition de divisions personnalisées sur p. 134.](#)

Personnalisation de l'affichage d'un arbre

L'onglet Afficheur du générateur d'arbres affiche l'arbre actuel. Par défaut, toutes les branches de cet arbre sont développées, mais vous pouvez également en réduire certaines et personnaliser d'autres paramètres, au besoin.

Figure 6-9

Affichage de l'arbre (de gauche à droite) contenant les détails de la règle de division, les graphiques des noeuds et les étiquettes



- Cliquez sur le signe moins (-) dans le coin inférieur droit d'un nœud parent pour masquer tous ses nœuds enfant. Cliquez sur le signe plus (+) dans le coin inférieur droit d'un nœud parent pour afficher tous ses nœuds enfant.
- Utilisez le menu Affichage ou la barre d'outils pour modifier l'orientation de l'arbre (de haut en bas, de gauche à droite ou de droite à gauche).
- Cliquez sur le bouton « Afficher les étiquettes de champ et de valeur » de la barre d'outils principale pour afficher ou masquer les étiquettes de champ et de valeur.
- Utilisez les boutons représentant une loupe pour effectuer un zoom avant ou arrière dans l'affichage. Vous pouvez également cliquer sur le bouton Carte de l'arbre à l'extrémité droite dans la barre d'outils pour afficher un diagramme de l'ensemble de l'arbre.
- Si un champ de partition est utilisé, vous pouvez passer des partitions d'apprentissage aux partitions de test (menu Affichage> Partition). Lorsque l'échantillon de test apparaît, vous pouvez afficher l'arbre, mais pas le modifier. (La partition actuelle apparaît dans la barre d'état dans l'angle inférieur droit de la fenêtre.)

- Cliquez sur le bouton d’affichage des informations de séparation (le bouton “i” à l’extrême droite de la barre d’outils) pour afficher les détails de la division actuelle. [Pour plus d’informations, reportez-vous à la section Détails et substitutions d’une division sur p. 137.](#)
- Affichez les statistiques ou les graphiques, ou les deux dans chaque nœud (reportez-vous à la section ci-dessous).

Affichage des statistiques et des graphiques

Statistiques des nœuds. Pour un champ cible catégoriel, le tableau de chaque nœud indique le nombre et le pourcentage d’enregistrements contenus dans chaque catégorie, ainsi que le pourcentage de la totalité de l’échantillon représenté par le nœud. Pour un champ cible continu (intervalle numérique), le tableau indique la moyenne, l’écart-type, le nombre d’enregistrements et la valeur prédite du champ cible.

Graphiques du nœud. Pour un champ cible catégoriel, le graphique est un diagramme en barres représentant les pourcentages de chaque catégorie du champ cible. Avant chaque ligne du tableau se trouve un échantillon des couleurs qui représentent chaque catégorie du champ cible dans les graphiques du nœud. Pour un champ cible continu (intervalle numérique), le graphique indique un histogramme du champ cible des enregistrements contenus dans le nœud.

Gains

L’onglet Gains contient les statistiques de tous les nœuds terminaux de l’arbre. Les gains permettent de mesurer la différence de la moyenne ou de la proportion d’un nœud donné par rapport à la moyenne globale. En règle générale, plus cette différence est significative, plus l’arbre est utile comme outil de prise de décision. Par exemple, la valeur d’index ou de « Lift » (augmentation) de 148 % d’un nœud implique que les enregistrements qu’il comporte ont 1,5 fois plus de chances d’être inclus dans la catégorie cible que l’ensemble de données complet.

Pour les nœuds Arbre C&R et QUEST où un ensemble de prévention de surajustement est spécifié, deux ensembles de statistiques sont affichés :

- ensemble de développement d’arbre : l’échantillon d’apprentissage avec l’ensemble de prévention de surajustement supprimé
- ensemble de prévention de surajustement

Pour les autres arbres interactifs Arbre C&R et QUEST, et pour tous les arbres interactifs CHAID, seules les statistiques d’ensembles de développement d’arbre sont affichées.

Figure 6-10
Onglet Gains

Ensemble de développement d'arbre							Ensemble de prévention de surajustement						
Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)	Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)
2	73,00	48,34	45,00	64,29	61,64	132,97	2	37,00	52,86	23,00	69,70	62,16	131,86
1	78,00	51,66	25,00	35,71	32,05	69,14	1	33,00	47,14	10,00	30,30	30,30	64,28

L'onglet Gains vous permet d'effectuer les tâches suivantes :

- Afficher les statistiques nœud par nœud, cumulatives ou en quantiles.
- Afficher les gains ou les profits.
- Passer de l'affichage des tableaux à celui des graphiques.
- Sélectionner la catégorie cible (pour les cibles catégorielles uniquement).
- Trier le tableau dans l'ordre croissant ou décroissant en fonction du pourcentage d'index. Si les statistiques de plusieurs partitions sont fournies, des tris sont systématiquement appliqués à l'échantillon d'apprentissage plutôt qu'à celui de test.

En général, les sélections effectuées dans le tableau des gains sont mises à jour dans l'affichage de l'arbre, et vice versa. Par exemple, si vous sélectionnez une ligne de ce tableau, le nœud correspondant est sélectionné dans l'arbre.

Gains de classification

Pour les arbres de classification (dotés d'une variable cible catégorielle), le pourcentage d'index de gains vous indique la différence de la proportion d'une catégorie cible à chaque nœud par rapport à la proportion globale.

Noeud par noeud

Dans ce mode, le tableau affiche une ligne pour chaque nœud terminal. Par exemple, si le taux de réponse global à votre campagne de publicité directe est de 10 %, mais que 20 % des enregistrements inclus dans le nœud X correspondent à des réponses positives, le pourcentage d'index de ce nœud s'élève à 200 %. Cela signifie que les personnes interrogées de ce groupe sont deux fois plus susceptibles d'acheter des articles que l'ensemble de la population.

Pour les noeuds Arbre C&R et QUEST où un ensemble de prévention de surajustement est spécifié, deux ensembles de statistiques sont affichés :

- ensemble de développement d'arbre : l'échantillon d'apprentissage avec l'ensemble de prévention de surajustement supprimé
- ensemble de prévention de surajustement

Pour les autres arbres interactifs Arbre C&R et QUEST, et pour tous les arbres interactifs CHAID, seules les statistiques d'ensembles de développement d'arbre sont affichées.

Figure 6-11
Statistiques de gains nœud par nœud

Ensemble de développement d'arbre							Ensemble de prévention de surajustement						
Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)	Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)
2	73,00	48,34	45,00	64,29	61,64	132,97	2	37,00	52,86	23,00	69,70	62,16	131,86
1	78,00	51,66	25,00	35,71	32,05	69,14	1	33,00	47,14	10,00	30,30	30,30	64,28

Noeuds. ID du nœud actuel (affiché dans l'onglet Afficheur).

Noeud : n. Nombre total d'enregistrements de ce nœud.

Noeud (%). Pourcentage d'enregistrements de l'ensemble de données compris dans ce nœud.

Gain : n. Nombre d'enregistrements dotés de la catégorie cible sélectionnée compris dans ce nœud. Cette option permet de savoir combien d'enregistrements de l'ensemble de données inclus dans la catégorie cible sont compris dans ce nœud.

Gain (%). Pourcentage d'enregistrements de la catégorie cible, dans l'ensemble de données complet, compris dans ce nœud.

Réponse (%). Pourcentage d'enregistrements du nœud actuel inclus dans la catégorie cible. Les réponses utilisées dans ce contexte sont parfois appelées « correspondances ».

Index (%). Pourcentage de réponses du nœud actuel exprimé en tant que pourcentage du pourcentage de réponses de l'ensemble de données complet. Par exemple, la valeur d'index de 300 % signifie que les enregistrements de ce nœud ont 3 fois plus de chances d'être inclus dans la catégorie cible que l'intégralité de cet ensemble de données.

Statistiques cumulatives

En mode cumulatif, le tableau contient un nœud par ligne, mais les statistiques sont cumulatives, triées dans l'ordre croissant ou décroissant par pourcentage d'index. Par exemple, si un tri décroissant est appliqué, le nœud possédant le pourcentage d'index le plus élevé est répertorié en premier. Les statistiques figurant sur les lignes, quant à elles, sont cumulatives pour la ligne en question et les lignes précédentes.

Figure 6-12
Gains cumulatifs triés dans l'ordre décroissant par pourcentage d'index

Ensemble de développement d'arbre							Ensemble de prévention de surajustement						
Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)	Noeuds	Noeud : n	Noeud (%)	Gain : n	Gain (%)	Réponse ...	Index (%)
2	73,00	48,34	45,00	64,29	61,64	132,97	2	37,00	52,86	23,00	69,70	62,16	131,86
1	151,00	100,00	70,00	100,00	46,36	100,00	1	70,00	100,00	33,00	100,00	47,14	100,00

Le pourcentage d'index cumulatif se réduit ligne par ligne à mesure que sont ajoutés les noeuds avec des pourcentages de réponse de plus en plus faibles. L'index cumulatif de la dernière ligne est toujours égal à 100 % car, à ce stade, l'ensemble de données complet est inclus.

Quantiles

Dans ce mode, chaque ligne du tableau représente un quantile au lieu d'un nœud. Il peut s'agir de quartiles, de quintiles (cinquièmes), de déciles (dixièmes), de vingtiles (vingtièmes) ou de centiles (centièmes). Plusieurs noeuds peuvent être répertoriés dans un seul quantile s'ils doivent être utilisés pour obtenir ce pourcentage (par exemple, si les quartiles apparaissent mais que les deux premiers noeuds contiennent moins de 50 % des observations). Les autres valeurs du tableau sont cumulatives et peuvent être interprétées de la même façon qu'en mode cumulatif.

Figure 6-13
Gains triés par quartile dans l'ordre décroissant par pourcentage d'index

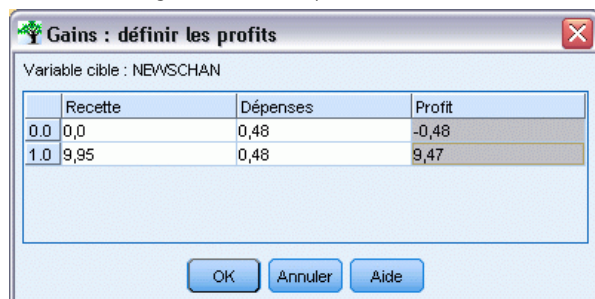
Ensemble de développement d'arbre							Ensemble de prévention de surajustement						
Noeuds	Centile	Centile : n	Gain : n	Gain (%)	Réponse ...	Index (%)	Noeuds	Centile	Centile : n	Gain : n	Gain (%)	Réponse ...	Index (%)
2	25,00	38,00	23,00	33,46	61,64	132,97	2	25,00	18,00	11,00	33,91	62,16	131,86
2,1	50,00	76,00	46,00	65,86	60,48	130,45	2	50,00	35,00	22,00	65,93	62,16	131,86
1	75,00	113,00	58,00	82,60	51,17	110,38	2,1	75,00	53,00	28,00	84,39	52,54	111,46
1	100,00	151,00	70,00	100,00	46,36	100,00	1	100,00	70,00	33,00	100,00	47,14	100,00

Profits de classification et retour sur investissement

Pour les arbres de classification, les statistiques des gains peuvent également être affichées en tant que profit et retour sur investissement. La boîte de dialogue définir les profits permet d'indiquer les revenus et les frais de chaque catégorie.

- Dans l'onglet Gains, cliquez sur le bouton de profit (étiqueté \$/\$) dans la barre d'outils pour accéder à cette boîte de dialogue.

Figure 6-14
Boîte de dialogue définir les profits



- Saisissez des valeurs de revenu et de frais à chaque catégorie du champ cible.

Par exemple, si vous devez payer 0,48 € pour envoyer une offre à chaque client, et que le revenu d'une réponse positive s'élève à 9,95 € pour un abonnement de trois mois, chaque réponse *négative* vous coûte 0,48 € et chaque réponse *positive* vous rapporte 9,47 € (9,95–0,48).

Dans le tableau des gains, les **profits** sont calculés en tant que total des revenus moins les frais pour chaque enregistrement d'un nœud terminal. Le **retour sur investissement**, quant à lui, représente le total des profits divisé par le total des frais d'un nœud.

Commentaires

- Les valeurs de profit n'affectent que les valeurs moyennes de profit et de retour sur investissement figurant dans le tableau des gains, pour que l'affichage des statistiques soit plus adapté à votre résultat net. Elles n'ont pas d'effet sur la structure de base du modèle d'arbre. Il convient de ne pas confondre les profits avec les coûts de mauvaise réaffectation, spécifiés dans le nœud de génération d'arbre et pris en compte dans le modèle, afin d'empêcher toute erreur qui risquerait de vous coûter cher.
- Les spécifications de profit ne sont pas conservées d'une session de création d'arbre interactif à l'autre.

Gains de régression

Pour les arbres de régression, vous avez le choix entre les modes d'affichage nœud par nœud, cumulatif nœud par nœud ou en quantiles. Les valeurs moyennes sont fournies dans le tableau. Les graphiques ne sont disponibles que pour les quantiles.

Graphiques de gains

Vous pouvez afficher un graphique dans l'onglet Gains à la place d'un tableau.

- Dans l'onglet Gains, cliquez sur l'icône Quantiles (la troisième en partant de la gauche dans la barre d'outils). (Les graphiques ne sont pas disponibles pour les statistiques cumulatives ou nœud par nœud.)
- Cliquez sur l'icône Graphiques.
- Sélectionnez les unités affichées (centiles, déciles, etc.) dans la liste déroulante, selon vos besoins.

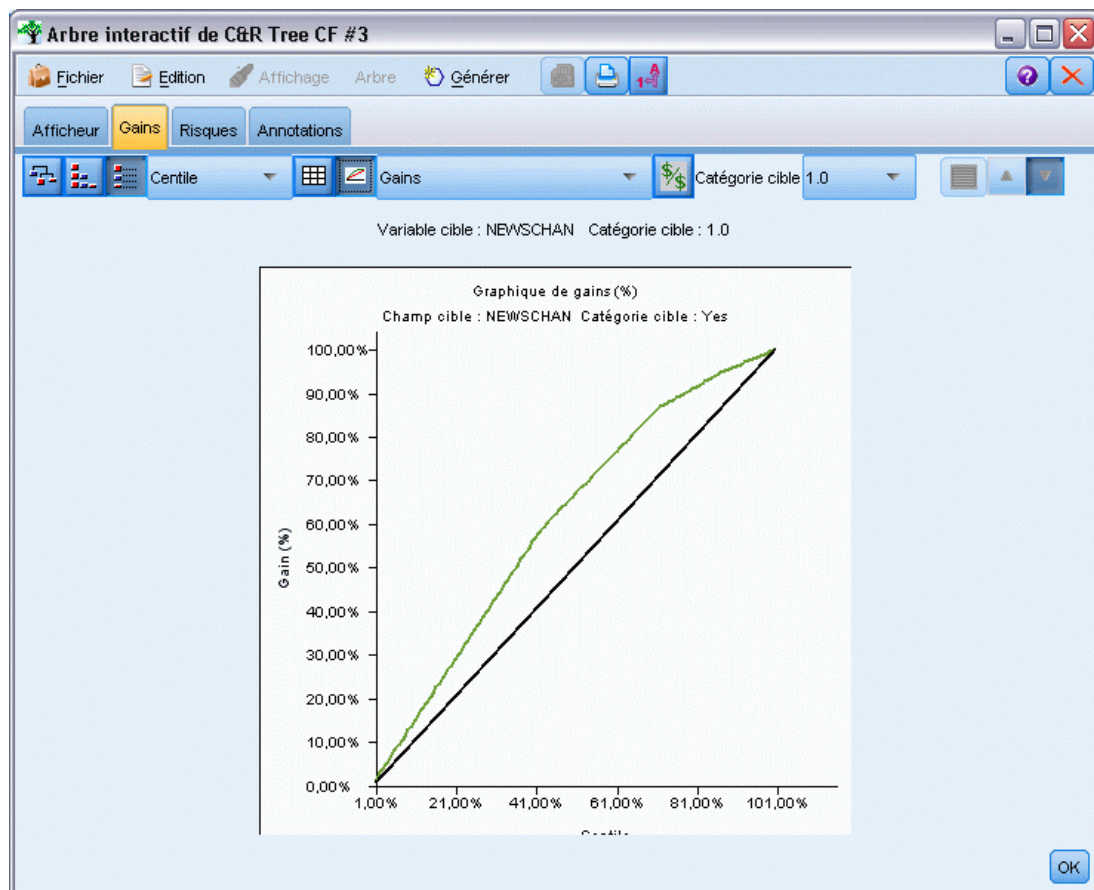
- Sélectionnez Gains, Réponse ou Lift pour modifier la mesure affichée.

Diagramme des gains

Le graphique de gains représente les valeurs affichées dans la colonne *Gain (%)* du tableau. Les gains sont définis en tant que proportion des correspondances de chaque incrément par rapport au nombre total de correspondances de l'arbre via l'équation suivante :

(correspondances de l'incrément / nombre total de correspondances) x 100 %

Figure 6-15
Graphiques de gains



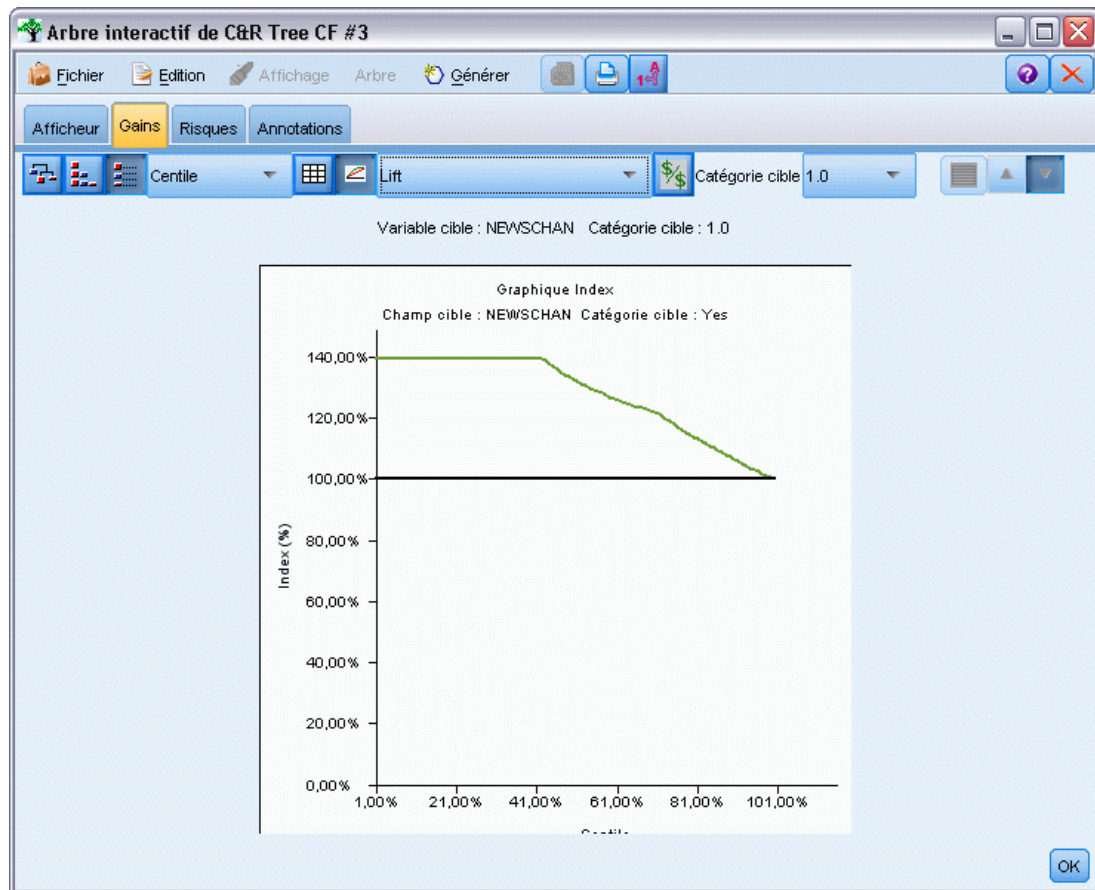
Ce graphique permet d'illustrer efficacement dans quelle mesure vous devez élargir vos perspectives pour obtenir un pourcentage précis de correspondances de l'arbre. La diagonale représente la réponse attendue pour l'échantillon complet si vous n'utilisiez pas le modèle. Dans ce cas, le taux de réponse serait constant, car une personne est susceptible de répondre comme une autre. Pour doubler ce nombre, vous devriez interroger deux fois plus de personnes. La courbe fournit la valeur d'amélioration possible du taux de réponse en n'incluant que les valeurs comprises dans les centiles plus élevés basés sur les gains. Par exemple, si les premiers 50 % sont ajoutés, vous pouvez obtenir plus de 70 % des réponses positives. Plus la courbe s'accroît, plus les gains sont élevés.

Graphique Lift

Le graphique Lift représente les valeurs affichées dans la colonne *Index (%)* du tableau. Ce graphique compare le pourcentage des enregistrements de chaque incrément qui se sont traduits par des correspondances et le pourcentage total de correspondances dans les ensembles de données d'apprentissage, via l'équation suivante :

(correspondances de l'incrément / enregistrements de l'incrément) / (nombre total de correspondances / nombre total d'enregistrements)

Figure 6-16
Graphique Lift

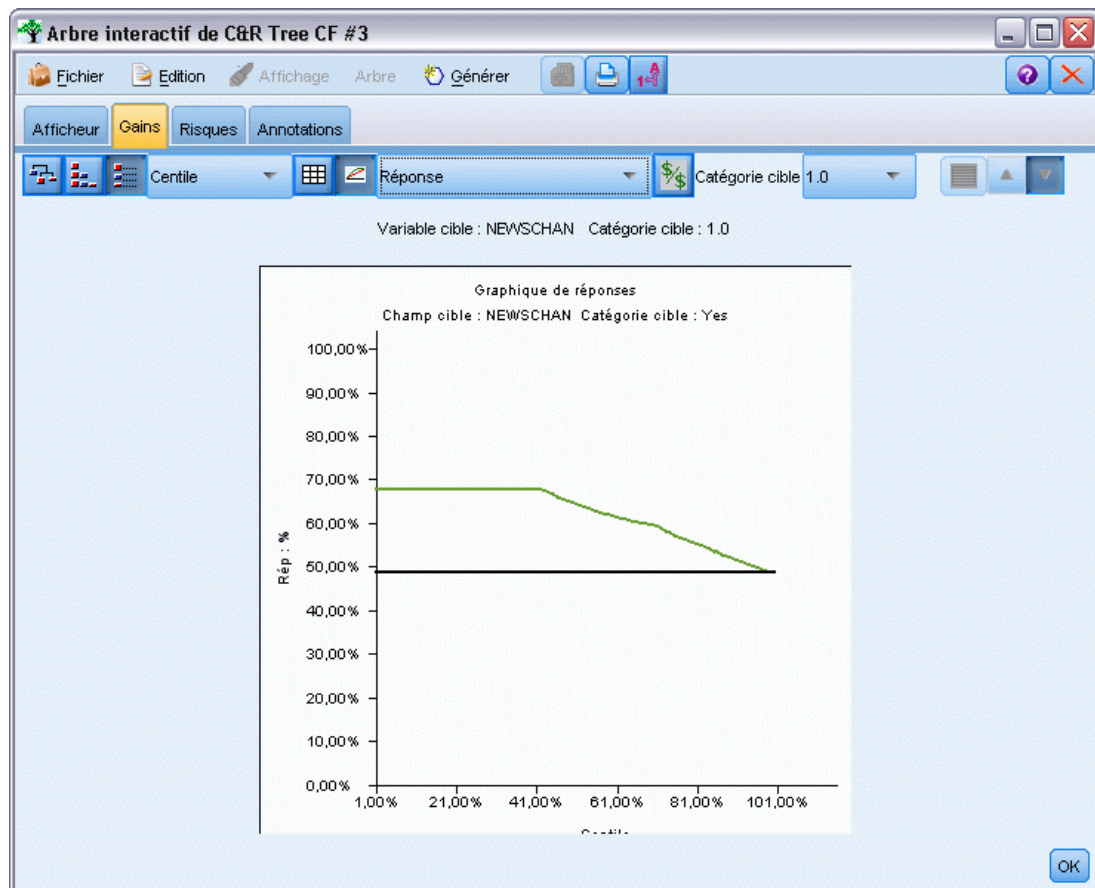


Graphique de réponses

Le graphique de réponses représente les valeurs affichées dans la colonne *Response (%)* du tableau. La réponse est un pourcentage d'enregistrements de l'incrément constituant des correspondances via l'équation suivante :

(réponses de l'incrément / enregistrements de l'incrément) x 100 %

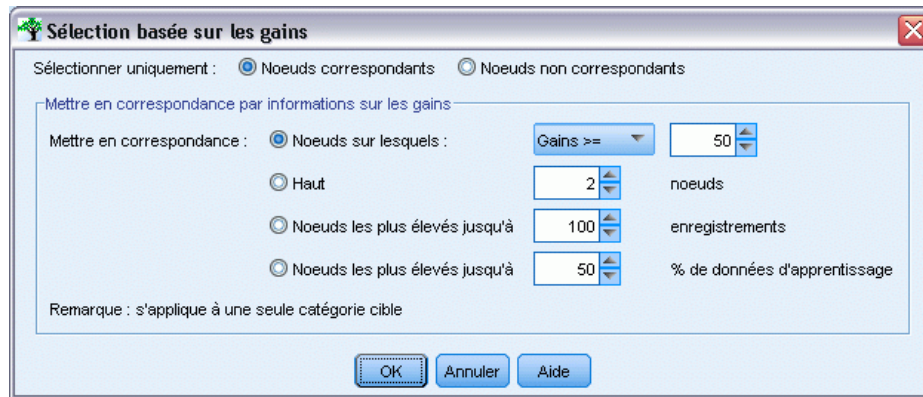
Figure 6-17
Graphique de réponses



Sélection basée sur les gains

La boîte de dialogue Sélection basée sur les gains vous permet de sélectionner automatiquement les noeuds terminaux offrant les gains les plus élevés (ou les plus faibles) selon la règle ou le seuil indiqué. Ensuite, vous pouvez générer un nœud Sélectionner en fonction des noeuds sélectionnés.

Figure 6-18
Boîte de dialogue Sélection basée sur les gains



- ▶ Dans l'onglet Gains, sélectionnez le mode d'affichage cumulatif ou noeud par noeud, puis la catégorie cible sur laquelle doit reposer la sélection. (La sélection repose sur l'affichage du tableau actuel et n'est pas disponible pour les quantiles.)
- ▶ Dans les différents menus de l'onglet Gains, sélectionnez l'une des options suivantes :
Edition > Sélectionnez les noeuds finals > Sélection basée sur les gains

Sélectionner uniquement. Vous pouvez sélectionner des noeuds correspondants *ou* non correspondants (par exemple, pour sélectionner *tous les enregistrements sauf* les 100 premiers).

Mettre en correspondance par informations sur les gains. Met en correspondance les noeuds en fonction des statistiques de gains de la catégorie cible actuelle, y compris les noeuds suivants :

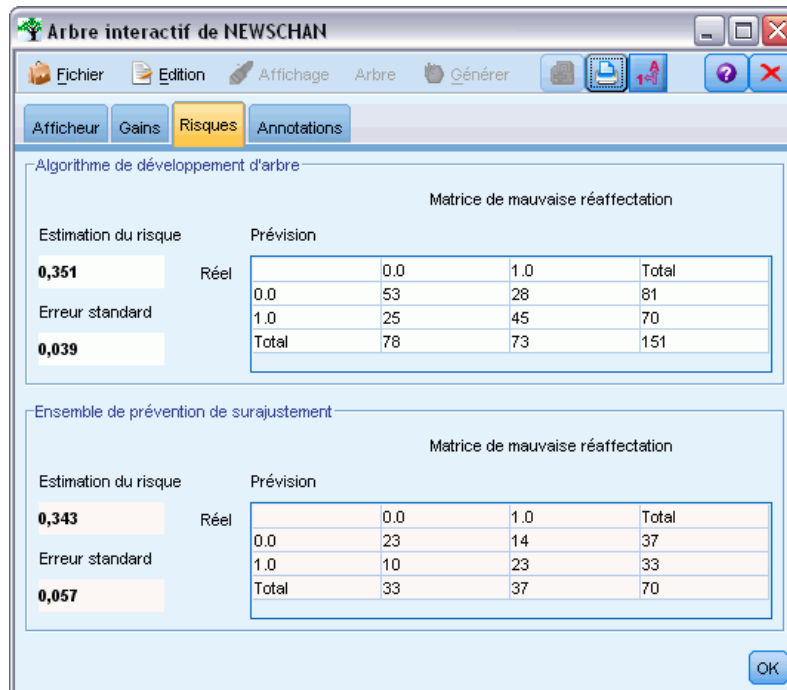
- Les noeuds dans lesquels les gains, la réponse ou le Lift (index) correspondent au seuil indiqué (par exemple, un taux de réponse supérieur ou égal à 50 %).
 - Les n premiers noeuds basés sur le gain de la catégorie cible.
 - Les premiers noeuds permettant d'atteindre un certain nombre d'enregistrements.
 - Les premiers noeuds permettant d'atteindre un pourcentage de données d'apprentissage donné.
- ▶ Cliquez sur OK pour mettre à jour la sélection dans l'onglet Afficheur.
 - ▶ Pour créer un noeud Sélectionner à partir de la sélection actuelle dans l'onglet Afficheur, choisissez Noeud Sélectionner dans le menu Générer. [Pour plus d'informations, reportez-vous à la section Génération de noeuds Filtrer et Sélectionner sur p. 154.](#)

Remarque : Dans la mesure où vous êtes en train de sélectionner des noeuds au lieu d'enregistrements ou de pourcentages, vous ne pouvez pas obtenir systématiquement une correspondance parfaite avec le critère de sélection fourni. Le système sélectionne des noeuds complets *jusqu'à ce qu'il parvienne au* niveau indiqué. Par exemple, si vous sélectionnez les 12 premières observations, et que 10 figurent dans le premier noeud et 2 dans le second noeud, seul le premier noeud est sélectionné.

Risques

Les risques vous informent des probabilités de mauvaise réaffectation à n'importe quel niveau. L'onglet Risques affiche une estimation ponctuelle des risques et (pour les résultats catégoriels) un tableau de réaffectations incorrectes.

Figure 6-19
Tableau de réaffectations incorrectes d'une cible catégorielle



- Pour les prévisions numériques, le risque correspond à une estimation pondérée de la variance à chaque nœud terminal.
- Pour les prévisions catégorielles, le risque représente la proportion d'observations mal classifiées, ajustées pour les probabilités a priori ou les coûts de mauvaise réaffectation.

Enregistrement de modèles d'arbre et de résultats

Vous pouvez enregistrer ou exporter les résultats des sessions de création d'arbre interactif de différentes manières, entre autres :

- Générer un modèle à partir de l'arbre actuel (Générer > Générer le modèle).
- Enregistrer les directives permettant de développer l'arbre actuel. Lors de l'exécution suivante du nœud de création d'arbre, l'arbre actuel se redéveloppe automatiquement, y compris les divisions personnalisées que vous avez définies.
- Exporter les informations relatives au modèle, aux gains et aux risques. [Pour plus d'informations, reportez-vous à la section Exportation des informations concernant les modèles, gains et risques sur p. 153.](#)

A partir du générateur d'arbres ou d'un nugget de modèle d'arbre, vous pouvez effectuer les opérations suivantes :

- Générer un noeud Filtrer ou Sélectionner à partir de l'arbre actuel. [Pour plus d'informations, reportez-vous à la section Génération de noeuds Filtrer et Sélectionner sur p. 154.](#)
- Générer un noeud Ensemble de règles représentant l'arborescence sous la forme d'un ensemble de règles définissant les branches terminales de l'arbre. [Pour plus d'informations, reportez-vous à la section Génération d'un noeud Ensemble de règles à partir d'un noeud Arbre décision sur p. 154.](#)
- En outre, vous pouvez exporter le modèle au format PMML pour les nuggets de modèle d'arbre uniquement. [Pour plus d'informations, reportez-vous à la section Palette Modèles dans le chapitre 3 sur p. 52.](#) Notez que si le modèle comporte des divisions personnalisées, ces informations ne sont pas conservées dans le fichier PMML exporté. (Les divisions sont conservées, mais il n'est plus possible de les distinguer de celles choisies par l'algorithme.)
- Générer un graphique à partir d'une partie sélectionnée de l'arbre actuel. *Remarque* : ceci fonctionne pour un nugget uniquement lorsqu'il est relié à d'autres noeuds dans un flux. [Pour plus d'informations, reportez-vous à la section Génération de graphiques sur p. 195.](#)

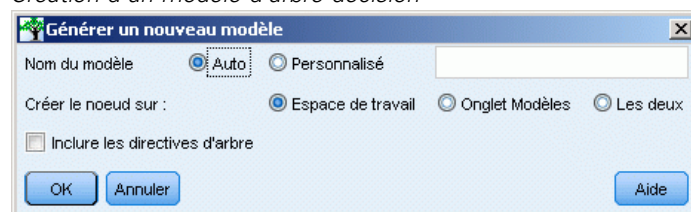
Remarque : il est impossible d'enregistrer l'arbre interactif. Pour éviter de perdre votre travail, générez un modèle et/ou mettez à jour les directives d'arbre avant de fermer la fenêtre du générateur d'arbres.

Création d'un modèle à partir du Générateur d'arbres

Pour générer un modèle à partir de l'arbre actuel, dans les différents menus du générateur d'arbres, sélectionnez :

Générer > Modèle

Figure 6-20
Création d'un modèle d'arbre décision



Vous avez le choix entre les options suivantes :

Nom du modèle. Vous pouvez spécifier un nom personnalisé ou créer le nom automatiquement à partir de celui du nœud de modélisation.

Créer le noeud sur. Vous pouvez ajouter le nœud sur l'espace de travail, sur la palette GM ou sur les deux.

Inclure les directives d'arbre. Pour inclure les directives de l'arbre actuel dans le modèle généré, sélectionnez cette case. Cela vous permet de régénérer l'arbre si nécessaire. [Pour plus d'informations, reportez-vous à la section Directives de développement d'arbre sur p. 151.](#)

Directives de développement d'arbre

Pour les modèles d'arbre C&RT, CHAID et QUEST, les directives d'arbre définissent les conditions de développement de l'arbre niveau par niveau. Les directives sont appliquées lorsque le générateur d'arbres interactifs est lancé à partir du noeud.

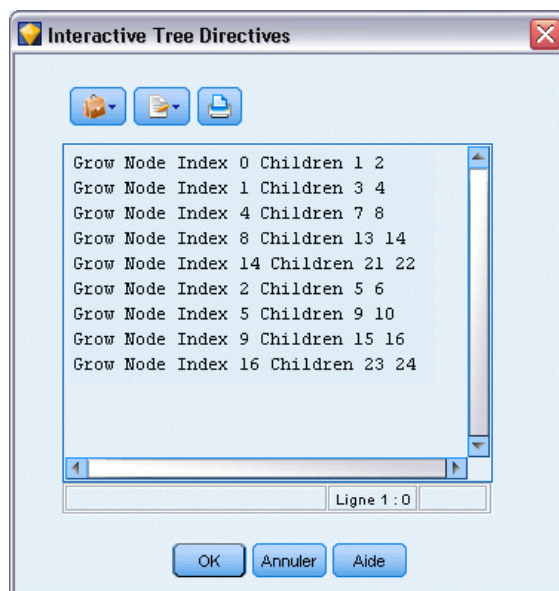
- Les directives permettent de recréer en toute sécurité un arbre créé lors d'une session interactive précédente. [Pour plus d'informations, reportez-vous à la section Mise à jour des directives d'arbre sur p. 153.](#) Vous pouvez également éditer ces directives manuellement, mais vous devez alors être très vigilant.
- Les directives correspondent strictement à la structure de l'arbre qu'elles décrivent. Ainsi, toute modification apportée aux données ou aux options de modélisation sous-jacentes peut entraîner l'échec d'un ensemble de directives auparavant valide. Par exemple, si l'algorithme CHAID change une division en deux groupes en division en trois groupes sur la base des données mises à jour, les directives reposant sur l'ancienne division sont incorrectes.

Remarque : Si vous optez pour la génération directe d'un modèle (sans utiliser le générateur d'arbres), les directives d'arbre ne sont pas prises en compte.

Modification des directives

- ▶ Pour afficher ou modifier les directives enregistrées, ouvrez le noeud de générateur d'arbres, puis sélectionnez le volet Objectif de l'onglet Options de création.
- ▶ Sélectionnez Lancer une session interactive pour activer les commandes, sélectionnez Utiliser les directives d'arbre, puis cliquez sur Directives.

Figure 6-21
Directives de développement d'arbre



Syntaxe des directives

Les directives définissent les conditions de développement de l'arbre, en commençant par le nœud racine. Par exemple, pour développer l'arbre d'un niveau :

```
Grow Node Index 0 Children 1 2
```

Aucune variable indépendante n'étant fournie, l'algorithme sélectionne la division optimale.

Notez que la première division doit systématiquement être effectuée sur le nœud racine (Index 0) et les valeurs d'index des deux enfants doivent être indiquées (1 et 2 en l'occurrence). Vous ne pouvez pas spécifier `Grow Node Index 2 Children 3 4` sauf si vous avez d'abord développé la racine qui a créé le nœud 2.

Pour développer l'arbre :

```
Développer l'arbre
```

Pour développer et élaguer l'arbre (Arbre C&RT uniquement) :

```
Grow_And_Prune Tree
```

Pour indiquer une division personnalisée pour une variable indépendante continue :

```
Grow Node Index 0 Children 1 2 Spliton
( "EDUCATE", Interval ( NegativeInfinity, 12.5)
  Interval ( 12.5, Infinity ) )
```

Pour diviser une variable indépendante nominale en deux valeurs :

```
Grow Node Index 2 Children 3 4 Spliton
( "GENDER", Group( "0.0" )Group( "1.0" ) )
```

Pour diviser une variable indépendante nominale en plusieurs valeurs :

```
Grow Node Index 6 Children 7 8 Spliton
( "ORGS", Group( "2.0","4.0" )
  Group( "0.0","1.0","3.0","6.0" ) )
```

Pour effectuer une division sur une variable indépendante ordinale :

```
Grow Node Index 4 Children 5 6 Spliton
( "CHILDS", Interval ( NegativeInfinity, 1.0)
  Interval ( 1.0, Infinity ) )
```

Remarque : Lorsque vous spécifiez des divisions personnalisées, les noms et valeurs de champ (EDUCATE, GENDER, CHILDS, etc.) distinguent les majuscules des minuscules.

Directives des arbres CHAID

Les directives des arbres CHAID sont particulièrement sensibles aux modifications apportées aux données ou au modèle car, à la différence des arbres C&RT et QUEST, elles ne doivent pas nécessairement utiliser des divisions binaires. Par exemple, la syntaxe suivante a l'air tout à fait valide, mais échoue si l'algorithme divise le noeud racine en plus de deux enfants :

```
Grow Node Index 0 Children 1 2  
Grow Node Index 1 Children 3 4
```

Pour les arbres CHAID, il est possible que le noeud 0 ait 3 ou 4 enfants, ce qui entraîne l'échec de la deuxième ligne de la syntaxe.

Utilisation des directives dans les scripts

Des directives peuvent également être intégrées entre guillemets triples dans les scripts. [Pour plus d'informations, reportez-vous à la section Blocs de texte littéral dans le chapitre 3 dans *Guide de génération de scripts et d'automatisation de IBM SPSS Modeler 15*.](#)

Mise à jour des directives d'arbre

Pour protéger votre travail contre une session de création d'arbre interactif, vous pouvez enregistrer les directives utilisées pour générer l'arbre actuel. Contrairement à l'enregistrement d'un nugget de modèle, qui n'est plus modifiable, cela vous permet de recréer l'arbre dans l'état où il se trouve pour le modifier ultérieurement.

- ▶ Pour mettre à jour les directives, dans les différents menus du générateur d'arbres, sélectionnez :
Fichier > Mettre à jour les directives

Les directives sont enregistrées dans le noeud de modélisation utilisé pour créer l'arbre (arbre C&RT, QUEST ou CHAID) et permettent de régénérer l'arbre actuel. [Pour plus d'informations, reportez-vous à la section Directives de développement d'arbre sur p. 151.](#)

Exportation des informations concernant les modèles, gains et risques

A partir du générateur d'arbres, vous pouvez exporter les statistiques concernant les modèles, gains et risques aux formats texte, HTML ou graphique, selon vos besoins.

- ▶ Dans la fenêtre du générateur d'arbres, sélectionnez l'onglet ou l'affichage à exporter.
- ▶ A partir des menus, sélectionnez :
Fichier > Exporter
- ▶ Choisissez Texte, HTML ou Graphiques, selon vos besoins. Ensuite, dans le sous-menu, sélectionnez les éléments à exporter.

Le cas échéant, l'exportation repose sur les sélections actuelles.

Exportation aux formats texte ou HTML. Vous pouvez exporter les statistiques concernant les gains ou les risques de la partition d'apprentissage ou de test (si elle a été définie). L'exportation est basée sur les sélections actuelles répertoriées dans l'onglet Gains — par exemple, vous pouvez opter pour des statistiques noeud par noeud, cumulatives ou en quantiles.

Exportation des graphiques. Vous pouvez exporter l'arbre actuel affiché dans l'onglet Afficheur ou les graphiques de gains de la partition d'apprentissage ou de test (si elle a été définie). Les formats disponibles sont les suivants : *.JPEG*, *.PNG* et *.BMP*. Pour les gains, l'exportation repose sur les sélections actuelles répertoriées dans l'onglet Gains (disponible seulement si un graphique est affiché).

Génération de noeuds Filtrer et Sélectionner

- Dans la fenêtre du générateur d'arbres ou en naviguant dans un nugget de modèle d'arbre décision, sélectionnez les options suivantes :

Générer > Noeud Filtrer

ou

> Noeud Sélectionner

Noeud Filtrer. Génère un nœud filtrant les champs non utilisés par l'arbre actuel. Cette méthode rapide permet de réduire l'ensemble de données pour inclure uniquement les champs que l'algorithme sélectionne comme étant importants. Si un noeud Typer figure en amont de ce noeud Arbre décision, tous les champs avec le rôle *Cible* sont transmis par le nugget de modèle Filtrer.

Noeud Sélectionner. Génère un nœud sélectionnant tous les enregistrements compris dans le noeud actuel. Cette option exige que vous sélectionniez plusieurs branches dans l'onglet Afficheur.

Le nugget de modèle est placé dans l'espace de travail de flux.

Génération d'un noeud Ensemble de règles à partir d'un noeud Arbre décision

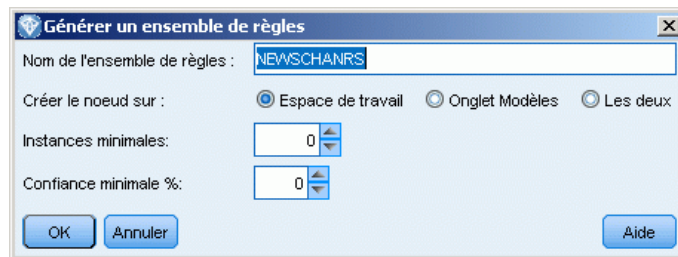
Vous pouvez générer un nugget de modèle Ensemble de règles représentant l'arborescence sous la forme d'un ensemble de règles définissant les branches terminales de l'arbre. Les ensembles de règles sont capables de conserver la plupart des informations importantes d'un arbre décision, selon un modèle moins complexe cependant. La principale différence est que dans un ensemble de règles, un enregistrement spécifique peut faire l'objet de plusieurs règles ou bien d'aucune. Par exemple, vous pouvez voir l'ensemble des règles prévoyant un résultat *néгатif*, suivi de celles prévoyant un résultat *positif*. Si plusieurs règles peuvent s'appliquer à un enregistrement, chacune de ces règles fait l'objet d'un « vote » pondéré basé sur le degré de confiance associée à cette règle. La prévision finale est alors calculée en combinant les votes pondérés de toutes les règles qui s'appliquent à l'enregistrement en question. Si aucune règle ne s'applique à un enregistrement, une prévision par défaut lui est alors attribuée.

Vous ne pouvez créer des ensembles de règles que depuis des arbres dotés de champs cible catégoriels (mais pas depuis des arbres de régression).

- Dans la fenêtre du générateur d'arbres ou en naviguant dans un nugget de modèle d'arbre décision, sélectionnez les options suivantes :

Générer > Ensemble de règles

Figure 6-22
Boîte de dialogue Générer un ensemble de règles



Nom de l'ensemble de règles. Permet d'attribuer un nom au nouveau nugget de modèle Ensemble de règles.

Créer le noeud sur. Détermine l'emplacement du nouveau nugget de modèle Ensemble de règles. Sélectionnez Canevas, Palette GM ou Les deux.

Instances minimales. Spécifiez le nombre minimal d'instances (nombre d'enregistrements auxquels les règles s'appliquent) à conserver dans le nugget de modèle Ensemble de règles. Les règles dont la prise en charge est inférieure à la valeur spécifiée ne seront pas incluses dans le nouvel ensemble de règles.

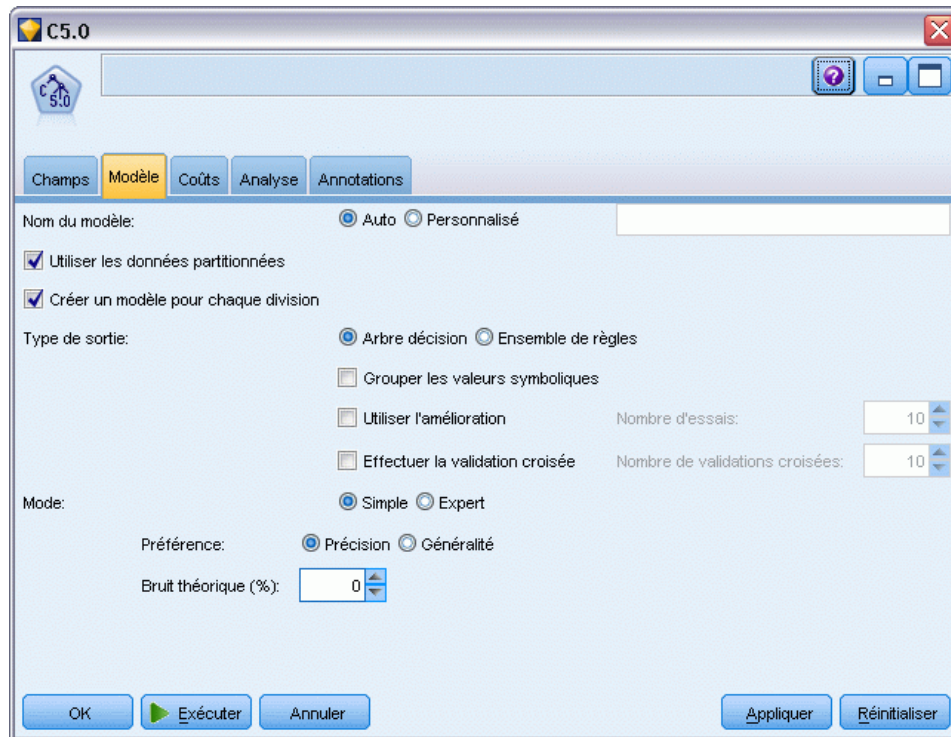
Confiance minimale. Indique le degré de confiance minimum que doivent avoir les règles pour apparaître dans le nugget de modèle Ensemble de règles. Les règles avec une confiance inférieure à la valeur spécifiée ne seront pas incluses dans le nouvel ensemble de règles.

Création directe d'un modèle d'arbre

Au lieu d'utiliser le générateur d'arbres interactif, vous pouvez créer un modèle d'arbre de décision directement depuis le noeud lorsque le flux est exécuté. Cette logique est cohérente avec la plupart des autres noeuds de création de modèle. Pour les modèles d'arbre C5.0, non pris en charge par le générateur d'arbres interactifs, seule cette méthode peut être utilisée.

- Créez un flux et ajoutez l'un des noeuds d'arbre de décision (C&RT, CHAID, QUEST ou C5.0).

Figure 6-23
Création directe d'un d'arbre C5.0



- Pour un arbre C&R, QUEST ou CHAID, dans le volet Objectif de l'onglet Options de création, sélectionnez l'un des objectifs principaux. Si vous sélectionnez Créer un seul arbre, vérifiez que Mode est configuré sur Générer le modèle.

Pour C5.0, dans l'onglet Modèle, configurez Type de sortie sur Arbre de décision.

- Sélectionnez des champs cible et des champs variables indépendantes, puis spécifiez des options de modèle supplémentaires, si nécessaire. Pour obtenir des instructions spécifiques, reportez-vous à la documentation de chaque nœud de création d'arbre.
- Exécutez le flux afin de générer le modèle.

Commentaires

- Lorsque des arbres sont créés à l'aide de cette méthode, les directives de développement d'arbre ne sont pas prises en compte.
- Qu'elles soient interactives ou directes, ces deux méthodes de création d'arbres décision génèrent en fin de compte des modèles similaires. Il est simplement question de savoir quel type de contrôle appliquer à un moment donné.

Noeuds Arbre de décision

Les noeuds Arbres de décision de IBM® SPSS® Modeler offrent un accès aux algorithmes de création d'arbres présentés précédemment :

- Arbre C&RT
- QUEST
- CHAID
- C5.0

Pour plus d'informations, reportez-vous à la section Modèles d'arbre décision sur p. 128.

Les algorithmes sont similaires par le fait qu'ils créent tous un arbre de décision en divisant de manière récursive les données en sous-groupes de plus en plus petits. Cependant, il existe quelques différences importantes

Champs d'entrée. Les champs d'entrée (variables indépendantes) peuvent être d'un des types suivants (niveaux de mesure) : continu, catégoriel, booléen, nominal ou ordinal. [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.](#)

Champs cible. Seul un champ cible peut être spécifié. Pour les arbres C&RT et CHAID, le champ cible peut être continu, catégoriel, booléen, nominal ou ordinal. Pour les arbres QUEST, il peut être catégoriel, booléen ou nominal. Pour les arbres C5.0 le champ cible peut être booléen, nominal ou ordinal.

Type de division. Les arbres C&R et QUEST ne prennent en charge que les divisions binaires (c'est-à-dire que chaque noeud de l'arbre peut être divisé en deux branches au maximum). Par contraste, CHAID et C5.0 prennent en charge la division en plus de deux branches à la fois.

Méthode utilisée pour la division. Les algorithmes diffèrent par les critères utilisés pour décider des divisions. Lorsqu'un arbre C&R prédit une sortie catégorielle, une mesure de dispersion est utilisée (par défaut, le coefficient Gini, bien que vous puissiez le modifier). La méthode de déviation des moindres carrés est utilisée pour les cibles continues. CHAID utilise un test Chi-deux ; QUEST utilise un test Chi-deux pour les variables indépendantes catégorielles et l'analyse de la variance des entrées continues. Pour C5.0 une mesure théorique des informations est utilisée, le rapport de gain d'informations.

Gestion des valeurs manquantes. Tous les algorithmes autorisent des valeurs manquantes pour les champs variables indépendantes, bien qu'ils utilisent différentes méthodes pour les gérer. Les arbres C&R et QUEST utilisent des champs variables indépendantes de remplacement, si nécessaire, pour faire progresser un enregistrement avec des valeurs manquantes dans l'arbre pendant l'apprentissage. CHAID crée une catégorie séparée pour les valeurs manquantes et autorise leur utilisation dans la création de l'arbre. C5.0 utilise une méthode de fractionnement qui transfère une partie fractionnaire d'un enregistrement vers les branches inférieures de l'arbre à partir d'un noeud dans lequel la division est basée sur un champ comportant une valeur manquante.

Elagage. Les arbres C&R, QUEST et C5.0 offrent la possibilité de développer complètement l'arbre puis de l'élaguer en supprimant les découpages de niveau inférieur qui ne contribuent pas de manière significative à la précision de l'arbre. Cependant, tous les algorithmes des arbres de décision vous permettent de contrôler la taille minimale d'un sous-groupe, ce qui vous permet d'éviter d'avoir des branches comportant peu d'enregistrements de données.

Création d'arbres interactifs. Les arbres C&R, QUEST et CHAID offrent la possibilité de lancer une session interactive. Ceci vous permet de construire votre arbre niveau par niveau, de modifier des divisions ou d'élaguer cet arbre avant de créer le modèle. C5.0 ne possède pas d'option interactive.

Probabilités a priori. Les arbres C&R et QUEST prennent en charge la spécification des probabilités a priori des catégories lors de la prévision d'un champ cible catégoriel. Les probabilités a priori sont des estimations de l'effectif relatif global de chaque catégorie cible dans la population d'où sont extraites les données d'apprentissage. Autrement dit, ce sont les estimations de probabilité que vous feriez pour chaque valeur cible possible avant de connaître quoi que ce soit sur les variables indépendantes. CHAID et C5.0 ne prennent pas en charge la spécification des probabilités a priori.

Ensembles de règles. Pour les modèles comportant des champs cibles catégoriels, les noeuds des arbres de décision offrent la possibilité de créer le modèle sous la forme d'un ensemble de règles, ce qui peut parfois s'avérer plus facile à interpréter qu'un arbre de décision complexe. Pour les arbres C&R, QUEST et CHAID, vous pouvez générer un ensemble de règles à partir d'une session interactive ; pour C5.0, pouvez spécifier cette option dans le noeud de modélisation. En outre, tous les modèles d'arbres de décision vous permettent de générer un ensemble de règles à partir d'un nugget de modèle. [Pour plus d'informations, reportez-vous à la section Génération d'un noeud Ensemble de règles à partir d'un noeud Arbre décision sur p. 154.](#)

Noeud Arbre C&RT

Le noeud Arbre C&RT (Classification et régression) est une méthode de classification et de prévision basée sur un système d'arborescence. Similaire au noeud C5.0, cette méthode utilise la technique de partition récursive afin de diviser les données d'apprentissage en segments présentant des champs de sortie similaires. Le noeud Arbre C&RT examine en premier lieu les champs d'entrée, afin de définir la meilleure segmentation : celle-ci est mesurée en fonction de la réduction de l'index d'impureté résultant de la segmentation. Le découpage définit deux sous-groupes qui sont à leur tour découpés en deux nouveaux sous-groupes : le découpage se poursuit jusqu'à ce que l'un des critères d'arrêt soit atteint. Toutes les divisions sont binaires (deux sous-groupes uniquement).

Elagage

Vous pouvez développer les arbres C&RT, puis l'élaguer à partir d'un algorithme de complexité des coûts. Cet algorithme ajuste l'évaluation des risques en fonction du nombre de noeuds terminaux. Cette méthode, qui permet à l'arbre de se développer avant d'être élagué par le biais de critères plus complexes, peut générer des arbres réduits offrant de meilleures propriétés de validation croisée. En général, augmenter le nombre de noeuds terminaux atténue le risque pour

les données (d'apprentissage) actuelles, mais le risque réel peut s'avérer bien plus important si le modèle s'étend aux données non visibles. Dans un cas extrême, supposez que vous disposez d'un autre nœud terminal pour chaque enregistrement des données d'apprentissage. Le pourcentage d'estimation des risques est nul, car chaque enregistrement est inclus dans son propre nœud, mais le risque de mauvaise réaffectation des données (de test) non visibles serait sans aucun doute supérieur à 0. La mesure de complexité des coûts est donc utilisée à des fins de compensation.

Exemple : Une entreprise de télévision câblée a requis la réalisation d'une étude marketing afin de déterminer les clients susceptibles de s'abonner à un service d'information interactif par câble. En utilisant les données de l'étude, vous pouvez créer un flux dans lequel le champ cible représente l'intention de souscrire à l'abonnement et les champs variables indépendantes comprennent l'âge, le sexe, l'éducation, la catégorie de revenus, les heures passées devant la télévision par jour et le nombre d'enfants. En appliquant un nœud Arbre CRT au flux, vous pourrez prédire et classifier les réponses pour obtenir le taux de réponses le plus élevé pour votre campagne.

Conditions requises. L'apprentissage d'un modèle d'arbre C&RT requiert l'utilisation d'au moins un champ *Entrée* et d'un champ *Cible*. Les champs cible et d'entrée peuvent être continus (intervalle numérique) ou catégoriels. Les champs paramétrés sur *Les deux* ou *Aucun* sont ignorés. Les types des champs utilisés dans le modèle doivent être totalement instanciés et les champs ordinaux (ensemble ordonné) dont il se sert doivent disposer d'un stockage numérique (et non d'une chaîne). Si nécessaire, vous pouvez utiliser le nœud Recoder pour les convertir. [Pour plus d'informations, reportez-vous à la section Nœud Recoder dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

Puissance. Les modèles d'arbre C&RT s'avèrent relativement robustes en présence de problèmes (par exemple, des données manquantes ou un nombre trop important de champs). Leur temps d'apprentissage est généralement court. De plus, les modèles d'arbre C&RT sont généralement plus faciles à comprendre que d'autres types de modèle dans la mesure où les règles extraites de ces modèles sont relativement simples à interpréter. Contrairement au nœud C5.0, le nœud Arbre C&RT prend en charge aussi bien les champs de sortie continus que catégoriels.

Noeud CHAID

CHAID (Chi-squared Automatic Interaction Detection) est une méthode de classification permettant de créer des arbres décision à l'aide de statistiques du Chi-deux, afin d'identifier les divisions optimales.

CHAID vérifie d'abord les tableaux croisés entre chaque champ d'entrée, ainsi que les résultats et les tests de signification par le biais d'un test d'indépendance Chi-deux. Si plusieurs de ces relations s'avèrent significatives d'un point de vue statistique, CHAID sélectionne le champ d'entrée le plus significatif (valeur p minimale). Si une valeur d'entrée possède plusieurs catégories, celles-ci sont comparées. Les catégories dont les résultats sont identiques sont réduites simultanément. Cette opération est effectuée en joignant la paire de catégories qui présente la plus faible différence de signification. Ce processus de fusion de catégories s'interrompt si toutes les autres catégories s'avèrent différentes au niveau de test indiqué. Pour les champs d'entrée nominaux, les catégories peuvent être fusionnées. Pour un ensemble d'ordinaux, seules les catégories contiguës peuvent l'être.

La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en profondeur toutes les divisions possibles pour chaque variable indépendante, mais dont les calculs sont plus longs.

Conditions requises. Les champs cible et d'entrée peuvent être des champs continus ou catégoriels. Les nœuds peuvent être divisés en plusieurs sous-groupes à chaque niveau. Les champs ordinaux utilisés dans le modèle doivent disposer d'un stockage numérique (et non d'une chaîne). Si nécessaire, vous pouvez utiliser le nœud Recoder pour les convertir. [Pour plus d'informations, reportez-vous à la section Nœud Recoder dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

Puissance. Contrairement aux nœuds C&R Tree et QUEST, CHAID peut générer des arbres non binaires, ce qui implique que certaines divisions possèdent plusieurs branches. Par conséquent, cette méthode a tendance à créer un arbre plus large que les méthodes de développement binaire. CHAID s'applique à tous les types d'entrées, et accepte les pondérations d'observation et les variables de fréquence.

Noeud QUEST

QUEST— (Quick, Unbiased, Efficient Statistical Tree) —est une méthode de classification binaire permettant de créer des arbres décision. L'une des principales raisons pour lesquelles cette méthode a été développée était de réduire le temps de traitement nécessaire aux analyses C&RT importantes, qui utilisaient alors de nombreuses variables ou observations. QUEST avait également pour objectif de limiter la tendance, observée parmi les méthodes d'arbre de classification, à favoriser les entrées autorisant un nombre supérieur de divisions, à savoir des champs d'entrée continus (intervalle numérique) ou ceux dotés de nombreuses catégories.

- QUEST utilise une séquence de règles, basée sur des tests de signification, pour évaluer les champs d'entrée d'un nœud. A des fins de sélection, vous pouvez être amené à n'effectuer qu'un seul test sur chaque entrée d'un nœud. Contrairement à C&RT, cette méthode ne vérifie pas toutes les divisions et, à la différence de C&RT et CHAID, elle ne teste pas non plus les combinaisons de catégories lorsqu'un champ d'entrée est évalué pour la sélection. L'analyse s'en trouve ainsi accélérée.
- Les divisions sont définies en exécutant une analyse discriminante quadratique via l'entrée sélectionnée dans les groupes qui se composent des catégories cible. Là encore, cette méthode accélère le processus lors d'une recherche complète (C&RT) pour déterminer la division optimale.

Conditions requises. Les champs d'entrée peuvent être continus (intervalles numériques), mais les champs cible doivent être catégoriels. Toutes les divisions sont binaires. Il est impossible d'utiliser les champs de pondération. Les champs ordinaux (ensemble ordonné) utilisés dans le modèle doivent disposer d'un stockage numérique (et non d'une chaîne). Si nécessaire, vous pouvez utiliser le nœud Recoder pour les convertir. [Pour plus d'informations, reportez-vous à la section Nœud Recoder dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

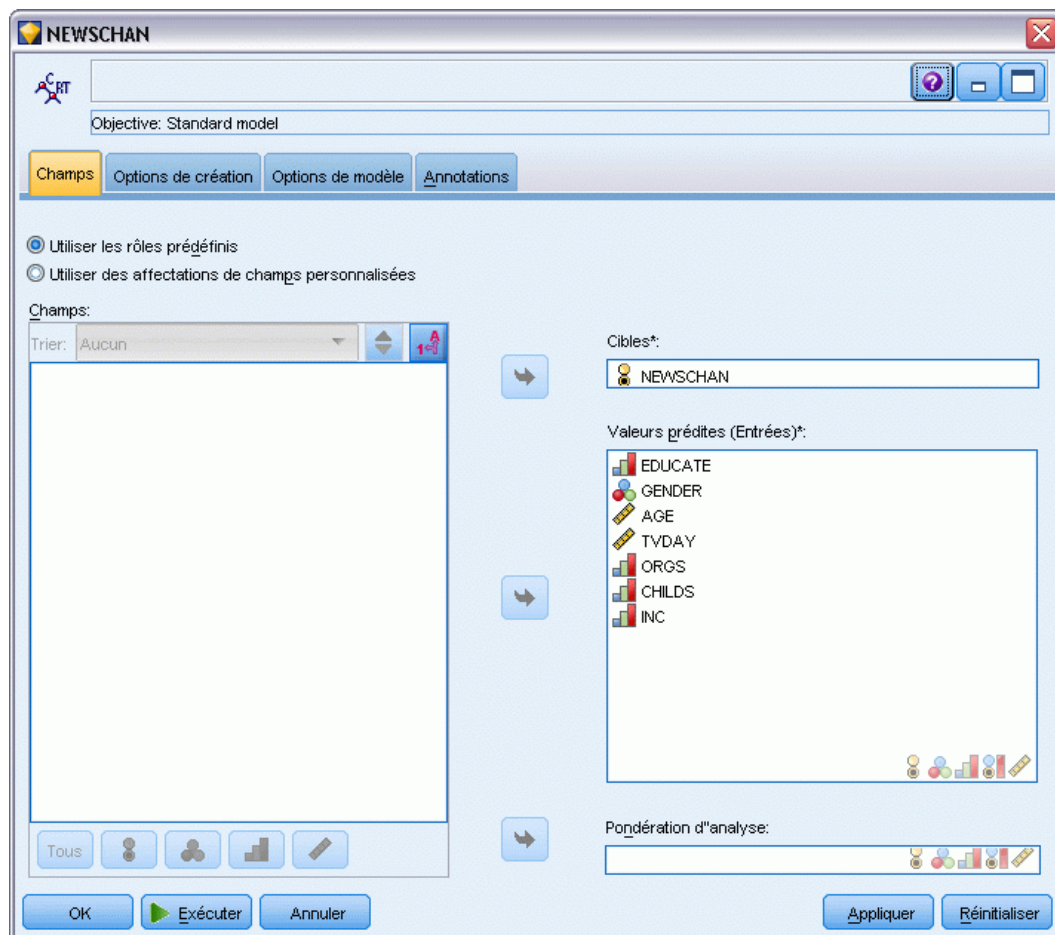
Puissance. A l'instar de CHAID, mais contrairement à C&RT, QUEST a recours à des tests statistiques pour décider si un champ d'entrée doit éventuellement être utilisé. Cette méthode traite également à part les problèmes de sélection et de division d'entrées, en appliquant différents

critères à chacun. QUEST contraste avec CHAID, méthode dans laquelle le résultat du test statistique déterminant la sélection de variables génère également la division. De même, l'arbre C&RT utilise la mesure d'incrément d'impureté pour sélectionner le champ d'entrée et définir la division.

Options des champs du noeud Arbre décision

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 6-24
Noeud d'arbre C&RT - Onglet Champs



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.

Valeurs prédites (Entrées). Sélectionnez un ou plusieurs champs comme entrée pour la prédiction.

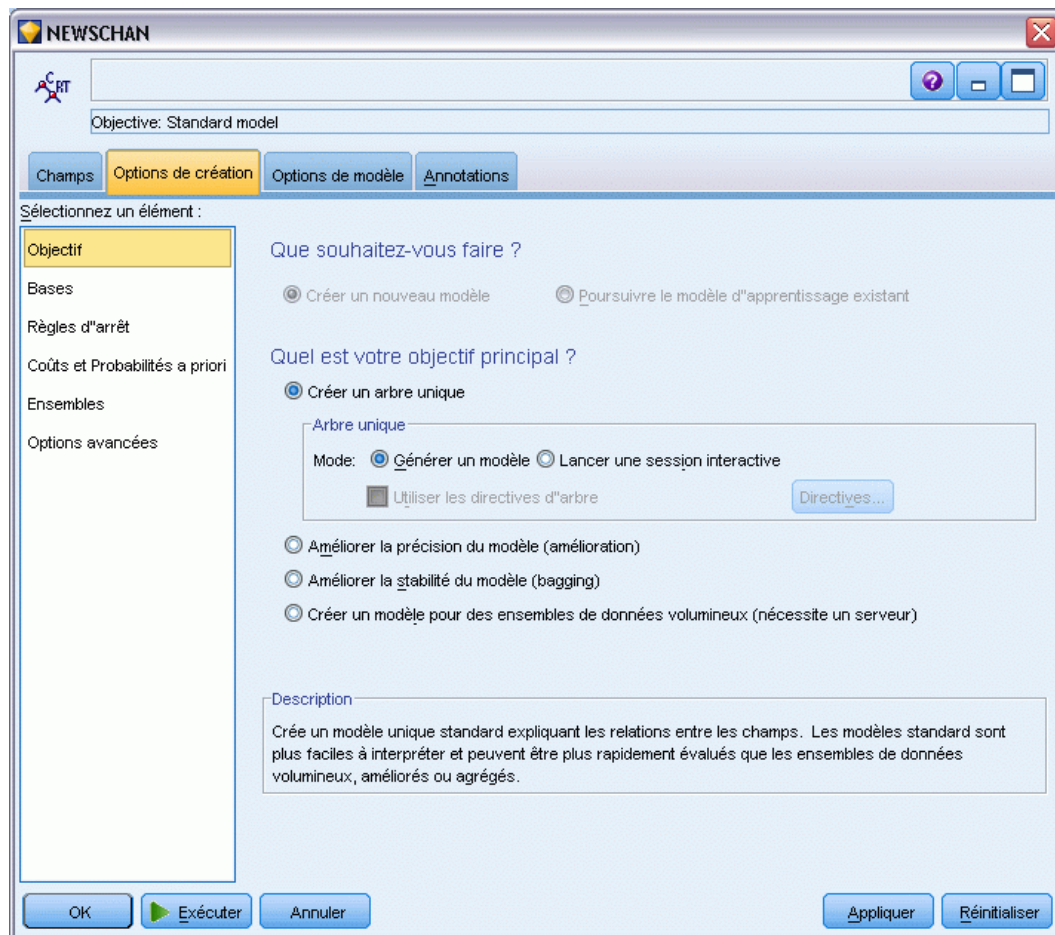
Pondération d'analyse. (CHAID et C&RT uniquement) Pour utiliser un champ en tant que pondération d'observation, spécifiez le champ ici. Les pondérations d'observation sont utilisées pour représenter les différences de variance dans les niveaux du champ de résultat. [Pour plus d'informations, reportez-vous à la section Utilisation des champs d'effectif et de pondération dans le chapitre 3 sur p. 42.](#)

Options de création du noeud Arbre décision

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.

Vous pouvez sélectionner l'option de création d'un nouveau modèle ou de mise à jour d'un modèle existant. Vous pouvez aussi définir le principal objectif du noeud : pour créer un modèle standard, pour en créer un avec une précision ou une stabilité accrue, ou pour créer un nouveau modèle à utiliser avec des ensembles de données très volumineux.

Figure 6-25
Noeud d'arbre C&RT - Onglet Options de création



Que souhaitez-vous faire ?

Créer un nouveau modèle. (Par défaut) Crée un modèle entièrement nouveau à chaque fois que vous exécutez un flux contenant ce noeud de modélisation.

Poursuivre l'apprentissage du modèle existant. Par défaut, un modèle complètement nouveau est créé chaque fois qu'un noeud de modélisation est exécuté. Si vous sélectionnez cette option, l'apprentissage se poursuit sur le dernier modèle généré par le noeud. Cela permet de mettre à jour ou d'actualiser un modèle existant sans avoir à accéder aux données d'origine et peut permettre des performances nettement plus rapides car *seuls* les enregistrements nouveaux ou mis à jour sont acheminés dans le flux. Les détails relatifs au modèle précédent sont stockés avec le noeud de modélisation, ce qui permet d'utiliser cette option même si le nugget du modèle précédent n'est plus disponible dans le flux ou dans la palette de modèles.

Remarque : L'option est activée uniquement si vous sélectionnez Créer un modèle pour des ensembles de données très volumineux comme objectif.

Quel est votre objectif principal ?

- **Créer un seul arbre.** Crée un seul modèle d'arbre de décision standard. Les modèles standards sont généralement plus faciles à interpréter et peuvent être plus rapides à évaluer que des modèles créés en utilisant d'autres options d'objectif.

Mode. Indique la méthode utilisée pour créer le modèle. Générer le modèle génère un modèle automatiquement lorsque le flux est exécuté. Lancer une session interactive ouvre le générateur d'arbres, qui vous permet de créer votre propre arbre niveau par niveau, de modifier des divisions ou d'élaguer cet arbre selon vos besoins, avant de créer le nugget de modèle.

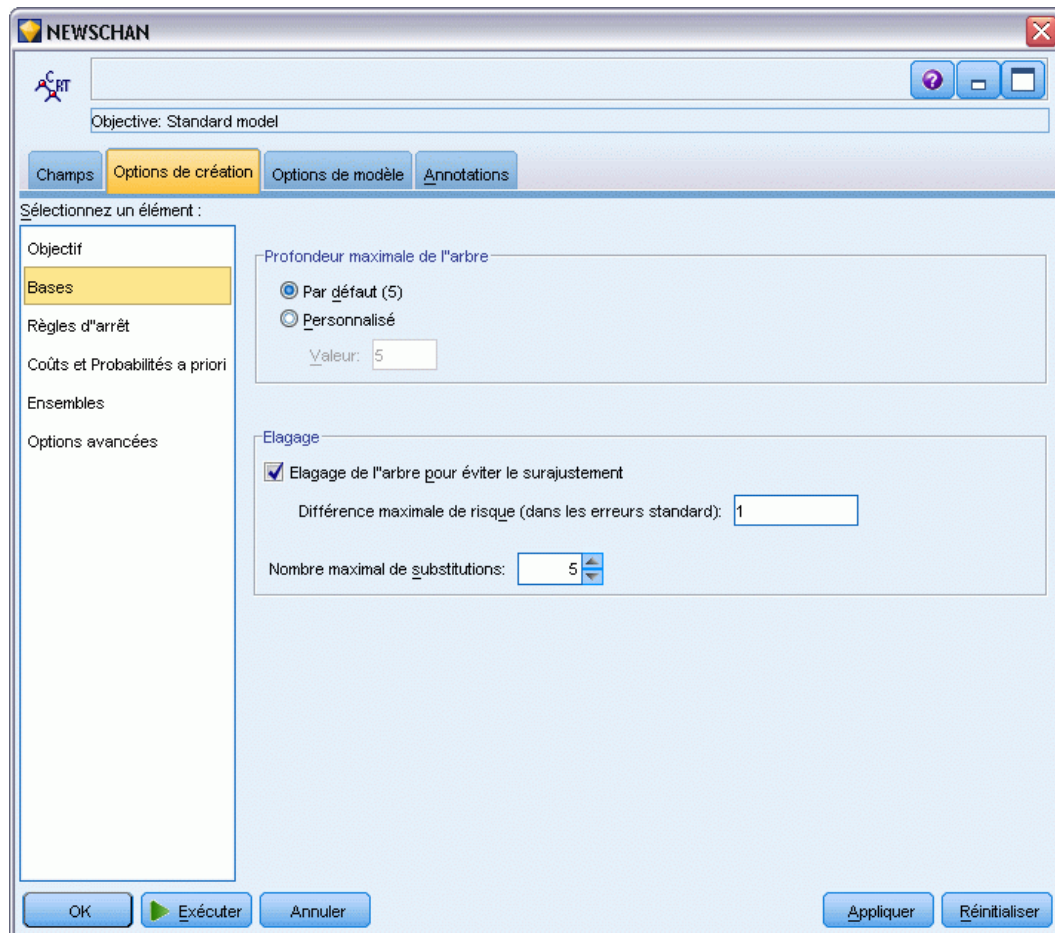
Utiliser les directives d'arbre. Sélectionnez cette option pour spécifier les directives à appliquer lors de la génération d'un arbre interactif depuis le nœud. Par exemple, vous pouvez spécifier les divisions effectuées aux premier et deuxième niveaux. Ces divisions sont appliquées automatiquement quand le générateur d'arbres est lancé. Il est également possible d'enregistrer les directives d'une session de création d'arbre interactif pour régénérer l'arbre ultérieurement. [Pour plus d'informations, reportez-vous à la section Mise à jour des directives d'arbre sur p. 153.](#)

- **Améliorer la précision du modèle (amélioration).** Sélectionnez cette option si vous souhaitez utiliser une méthode spéciale, appelée **amélioration**, afin d'améliorer le taux de précision du modèle. L'amélioration repose sur la création séquentielle de plusieurs modèles. Le premier modèle est généré de manière classique. Un deuxième modèle est généré ; il contient essentiellement les enregistrements qui n'ont pas été correctement classés par le premier modèle. Un troisième modèle contenant les erreurs du deuxième arbre est ensuite généré, et ainsi de suite. Au final, l'ensemble des modèles est appliqué aux observations à l'aide d'une procédure de vote pondéré qui permet de combiner les différentes prévisions en une prévision globale. Si le processus d'amélioration peut accroître de façon significative la précision d'un modèle d'arbre de décision, son utilisation requiert également un temps d'apprentissage plus long.
- **Améliorer la stabilité du modèle (agrégation).** Sélectionnez cette option si vous souhaitez utiliser une méthode spéciale, appelée **agrégation**, (agrégation de bootstrap) afin d'améliorer la stabilité du modèle et d'éviter le surajustement. Cette option crée plusieurs modèles et les combine afin d'obtenir des prédictions plus fiables. La création et l'évaluation des modèles obtenus à l'aide de cette option peut prendre davantage de temps que des modèles standard.
- **Créer un modèle pour des ensembles de données très volumineux.** Sélectionnez cette option lorsque vous travaillez avec des ensembles de données trop volumineux pour créer un modèle qui utilise une des autres options d'objectif. Cette option divise les données en blocs de données plus petits et construit un modèle pour chaque bloc. Les modèles les plus précis sont alors automatiquement sélectionnés et combinés pour créer un nugget de modèle unique. Vous pouvez réaliser une mise à jour d'un modèle incrémentiel si vous sélectionnez l'option Poursuivre l'apprentissage du modèle existant sur cet écran. *Remarque :* Cette option destinée aux ensembles de données très volumineux nécessite une connexion à IBM® SPSS® Modeler Server. [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.](#)

Noeuds Arbre de décision - Bases

C'est ici que vous pouvez spécifier les options de base concernant la création de l'arbre de décision.

Figure 6-26
Options d'arbre décision de base



Algorithme de développement d'arbre (CHAID uniquement) Sélectionnez le type d'algorithme CHAID que vous souhaitez utiliser. La méthode Exhaustive CHAID correspond à une modification du CHAID qui examine plus en profondeur toutes les divisions possibles pour chaque variable indépendante, mais dont les calculs sont plus longs.

Profondeur maximale de l'arbre. Indiquez le nombre maximal de niveaux inférieurs au nœud Racine (nombre de découpages récursifs que doit subir l'échantillon). La valeur par défaut est 5, sélectionnez Personnaliser et saisissez une valeur pour spécifier un nombre de niveaux différent.

Élagage (C&RT et QUEST uniquement)

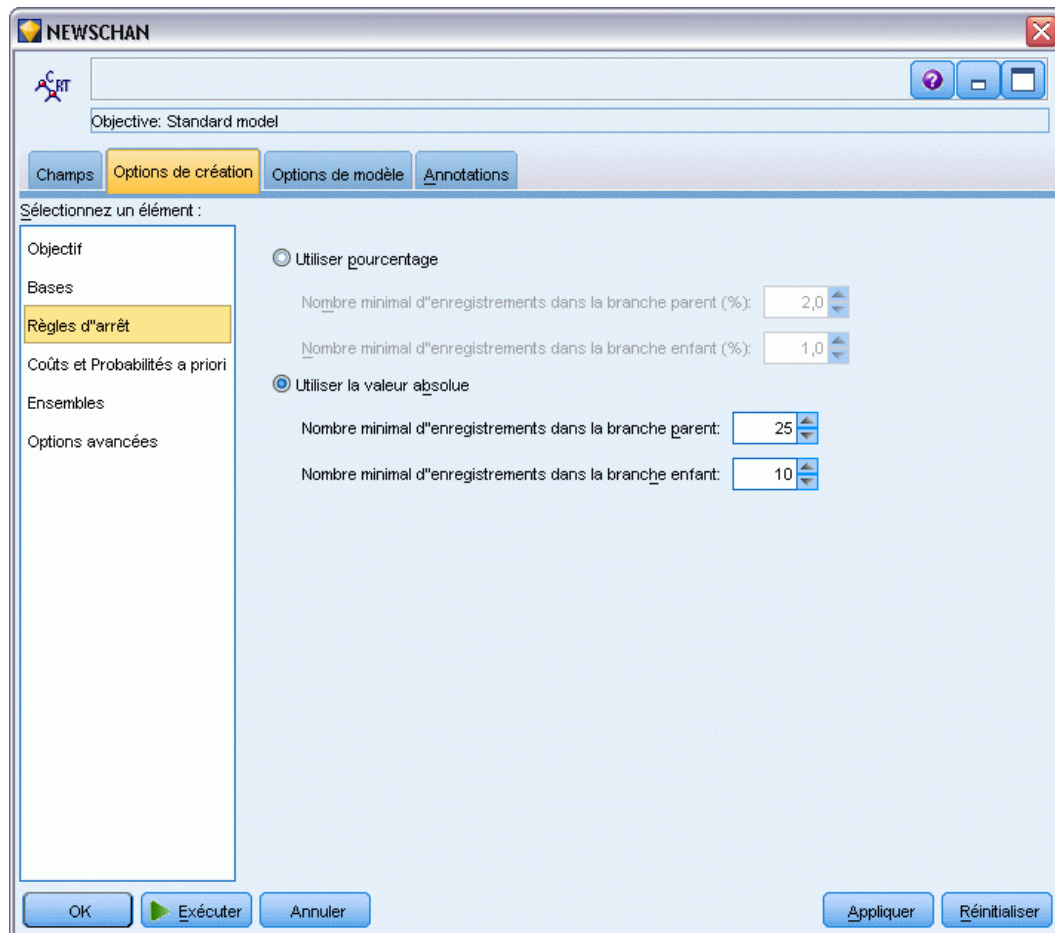
Élagage de l'arbre pour éviter le surajustement. L'élagage consiste à supprimer les découpages de niveau inférieur qui ne contribuent pas de manière significative à la précision de l'arbre. L'élagage peut permettre de simplifier l'arbre, ce qui peut faciliter l'interprétation et, dans certains cas, améliorer la généralisation. Si vous souhaitez que l'ensemble de l'arbre ne subisse aucun élagage, laissez cette option désélectionnée.

- **Différence maximale pour le risque (dans Erreurs standard).** Permet de spécifier une règle d'élagage plus flexible. La règle d'erreur standard permet de sélectionner l'arbre le plus élémentaire dont le risque estimé est proche (mais éventuellement supérieur) de celui du sous-arbre présentant le risque le plus faible. La valeur indique la taille de l'écart autorisé pour le risque estimé entre l'arbre élagué et l'arbre présentant le risque le plus faible en termes de risque estimé. Par exemple, si vous spécifiez 2, un arbre dont le risque estimé est ($2 \times$ erreur standard) supérieur à celui de l'arbre entier peut être sélectionné.

Nombre maximal de substitutions. Les substitutions permettent de résoudre les problèmes de valeurs manquantes. Pour chaque découpage dans l'arbre, l'algorithme identifie les champs d'entrée les plus proches du champ de découpage sélectionné. Ces champs deviennent les **substitutions** pour ce découpage. Lorsqu'un enregistrement doit être classé mais qu'une valeur est manquante dans un champ de découpage, le nœud utilisera la valeur correspondante dans un champ substitut afin d'opérer le découpage. Si l'augmentation de ce paramètre permet une gestion plus polyvalente des valeurs manquantes, elle accroît également l'utilisation de la mémoire et allonge la durée de l'apprentissage.

Noeuds Arbre de décision - Règles d'arrêt

Figure 6-27
Options des règles d'arrêt

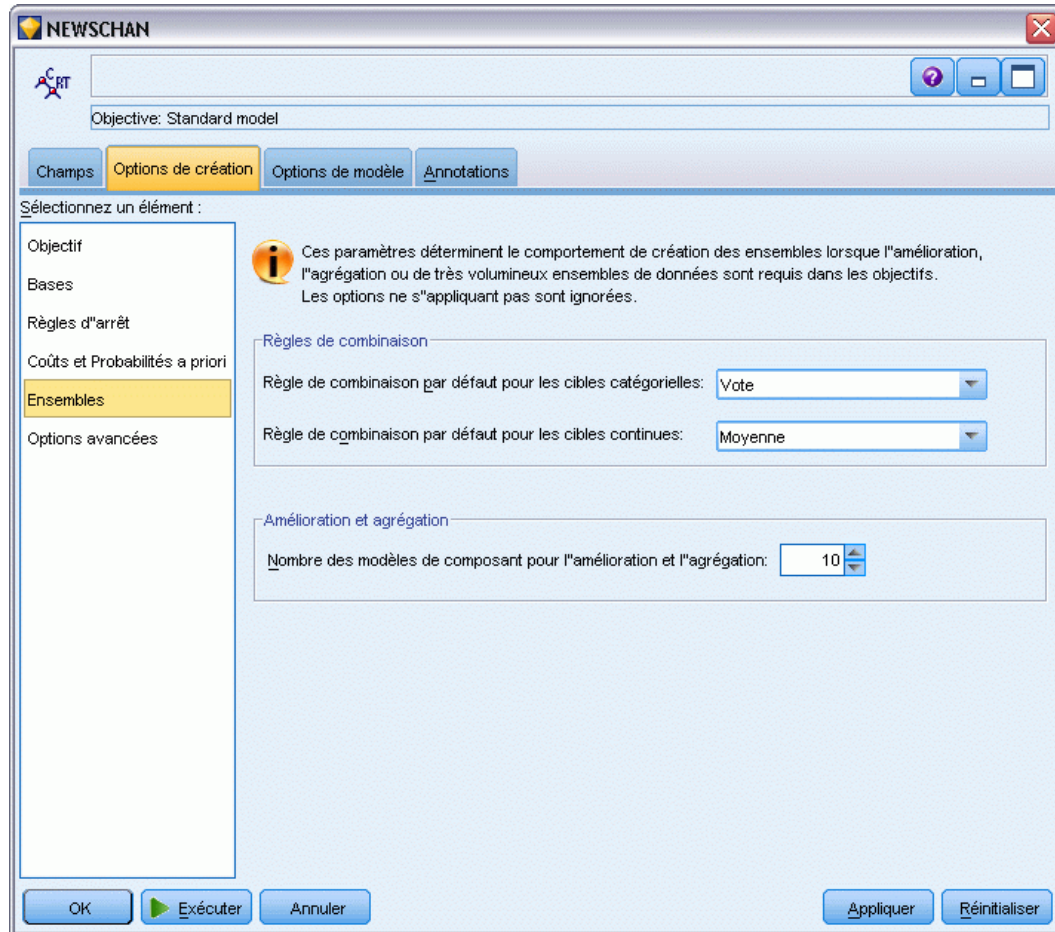


Ces options permettent de contrôler la création de l'arbre. Les règles d'arrêt déterminent quand des branches spécifiques de l'arbre doivent être arrêtées. Vous pouvez définir des tailles de branche minimale pour éviter que le découpage ne crée des sous-groupes trop petits. Nombre minimal d'enregistrements dans la branche parent permet d'éviter le découpage si le nombre d'enregistrements dans le nœud devant être découpé (les nœuds **parents**) est inférieur à la valeur spécifiée. Nombre minimal d'enregistrements dans la branche enfant permet d'éviter le découpage si le nombre d'enregistrements dans une branche créée par le découpage (les nœuds **enfant**) est inférieur à la valeur spécifiée.

- **Utiliser pourcentage.** Permet de spécifier, sous forme de pourcentage, la taille globale des données d'apprentissage.
- **Utiliser la valeur absolue.** Permet de spécifier les tailles en tant que nombre absolu d'enregistrements.

Noeuds Arbre de décision - Ensembles

Figure 6-28
Options des ensembles



Ces paramètres déterminent le comportement d'assemblage qui se produit lors du boosting, du bagging ou lorsque que des ensembles volumineux de données sont requis dans les objectifs. Les options qui ne s'appliquent pas à l'objectif sélectionné sont ignorées.

Bagging and Very Large Datasets. When scoring an ensemble, this is the rule used to combine the predicted values from the base models to compute the ensemble score value.

- **Règles de combinaison par défaut pour les cibles catégorielles.** Des valeurs prédites d'ensemble pour ces cibles catégorielles peuvent être combinées à l'aide du vote, de la probabilité la plus élevée ou de la probabilité de moyenne la plus élevée. Le **vote** sélectionne la catégorie qui a la plus forte probabilité le plus souvent sur les mêmes modèles de base. La **probabilité la plus élevée** sélectionne la catégorie qui atteint la probabilité la plus élevée sur tous les modèles de base. La **probabilité de moyenne la plus élevée** sélectionne la catégorie dont

la valeur est la plus élevée lorsqu'est effectuée la moyenne des probabilités de catégorie sur les modèles de base.

- **Règles de combinaison par défaut pour les cibles continues.** Des valeurs prédites d'ensemble pour des cibles continues peuvent être combinées à l'aide de la moyenne ou de la médiane des valeurs prédites à partir des modèles de base.

Veillez noter que lorsque l'objectif consiste à améliorer la précision du modèle, les sélections de règles de combinaisons sont ignorées. Le boosting utilise toujours un vote majoritaire pondéré pour évaluer des cibles catégorielles et une médiane pondérée pour évaluer des cibles continues.

Boosting et bagging. Spécifiez le nombre de modèles de base à créer lorsque l'objectif est d'améliorer la précision ou la stabilité du modèle ; pour le bagging, il s'agit du nombre d'échantillons de bootstrap. Il doit s'agir d'un entier positif.

Noeuds C&RT et QUEST - Coûts & Probabilités a priori

Figure 6-29
Définition des coûts de mauvaise réaffectation et des probabilités a priori

The screenshot shows the NEWSCHAN software interface. The 'Options de création' tab is selected, and the 'Coûts et Probabilités a priori' section is active. The 'Objectif' is set to 'Standard model'. The 'Coûts de mauvaise réaffectation' section has the checkbox 'Utiliser les coûts de mauvaise réaffectation' checked. A table shows the cost matrix for 'Prévu' vs 'Réel'.

	0	1
Réel 0	0,0	1,0
Réel 1	1,0	0,0

The 'Probabilités a priori' section has the radio button 'Personnalisé' selected. A table shows the prior probabilities for 'Valeur' vs 'Probabilité'.

Valeur	Probabilité
0	0,5
1	0,5

Buttons at the bottom include 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser'.

Coûts de mauvaise réaffectation

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.

Probabilités a priori

Ces options vous permettent de spécifier les probabilités a priori des catégories lors de la prévision d'un champ cible symbolique. Les **probabilités a priori** sont des estimations de l'effectif relatif global de chaque catégorie cible dans la population d'où sont extraites les données d'apprentissage. Autrement dit, ce sont les estimations de probabilité que vous feriez pour chaque valeur cible possible *avant* de connaître quoi que ce soit sur les valeurs prédites. Vous pouvez les configurer de trois façons différentes.

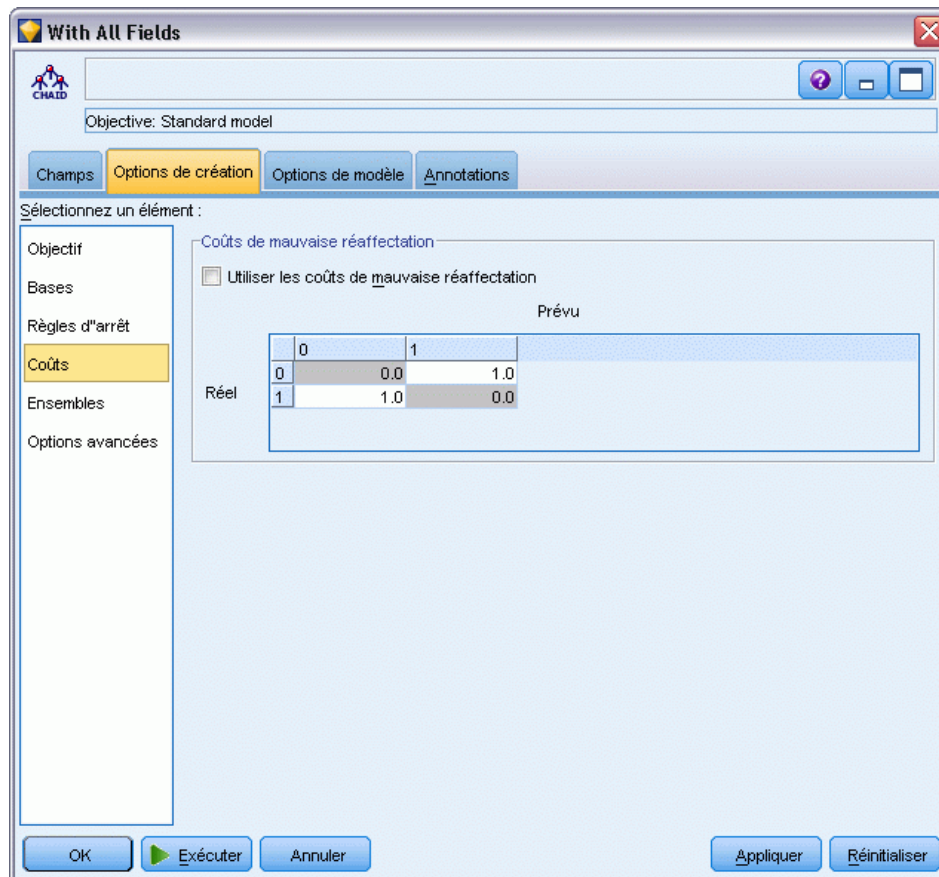
- **En fonction des données d'apprentissage.** Il s'agit de l'option par défaut. Les probabilités a priori sont basées sur les effectifs relatifs des catégories dans les données d'apprentissage.
- **Identiques pour toutes les classes.** Les probabilités a priori sont définies par $1/k$, où k correspond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

Le total des probabilités a priori de toutes les catégories doit être égal à 1 (la **contrainte de probabilité**). 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 probabilité. 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.

Ajuster les probabilités a priori en utilisant les coûts de mauvaise réaffectation. Cette option vous permet d'ajuster les probabilités a priori, en fonction des coûts de mauvaise réaffectation (onglet Coûts). Cela vous permet d'intégrer des informations sur les coûts au processus d'accroissement pour les arbres qui utilisent la mesure d'impureté Associer par paire. (Lorsque cette option n'est pas sélectionnée, les informations sur les coûts sont uniquement utilisées pour la classification des enregistrements et le calcul des risques estimés pour les arbres, en fonction de la mesure Associer par paire.)

Noeud CHAID - Coûts

Figure 6-30
Coûts de mauvaise réaffectation dans le noeud CHAID



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*.

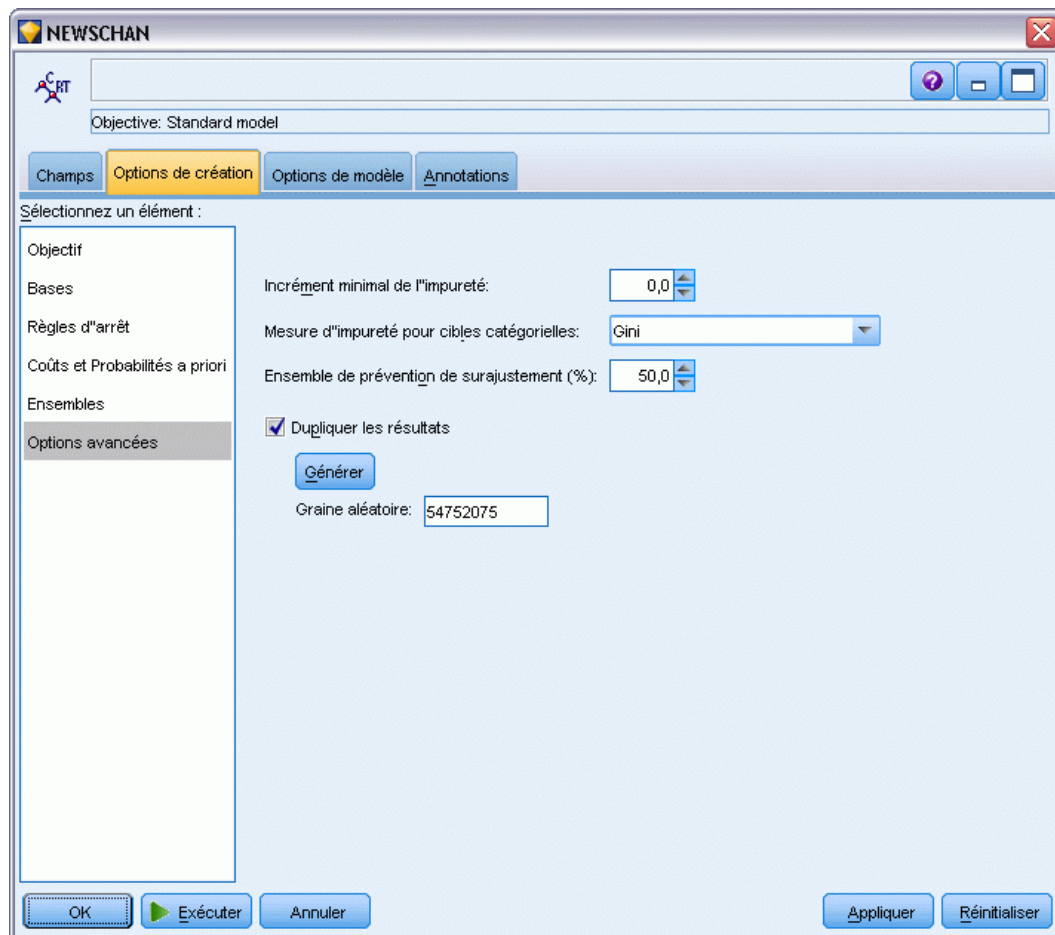
La matrice de mauvaise réaffectation des coûts affiche le coût de chaque combinaison possible de catégories prédites et de catégories réelles. Par défaut, tous les coûts de mauvaise réaffectation sont paramétrés sur 1. Pour entrer des valeurs de coût personnalisées, sélectionnez Utiliser les coûts de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

Noeud Arbre C&RT - Options avancées

Les options avancées vous permettent d'affiner le processus de génération d'arbres.

Figure 6-31
Définition des options avancées du noeud d'arbre C&RT



Incrément minimal de l'impureté. Indiquez une valeur pour générer un nouveau découpage dans 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. Pour les cibles catégorielles, un noeud est considéré comme étant « pur » si 100 % de ses observations sont incluses dans une catégorie spécifique du champ cible. 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, alors le découpage ne sera pas effectué.

Mesure d'impureté pour cibles catégorielles. Pour les champs cible catégoriels, indiquez la méthode utilisée pour mesurer l'impureté de l'arbre. (Pour les cibles continues, cette option n'est pas prise en compte et la mesure d'impureté **au moindre carré** est toujours utilisée.)

- Gini est une mesure d'impureté générale dont le calcul est basé sur les probabilités d'appartenance à une catégorie de la branche.

- La mesure Associer par paire est une autre mesure d'impureté qui met l'accent sur le découpage binaire : elle permet d'obtenir des branches de tailles plus similaires.
- Ordonné ajoute la contrainte supplémentaire selon laquelle seules les catégories cible contiguës peuvent être regroupées, comme c'est le cas pour les cibles ordinales uniquement. Si cette option est sélectionnée pour une cible nominale, la mesure Associer par paire standard est utilisée par défaut.

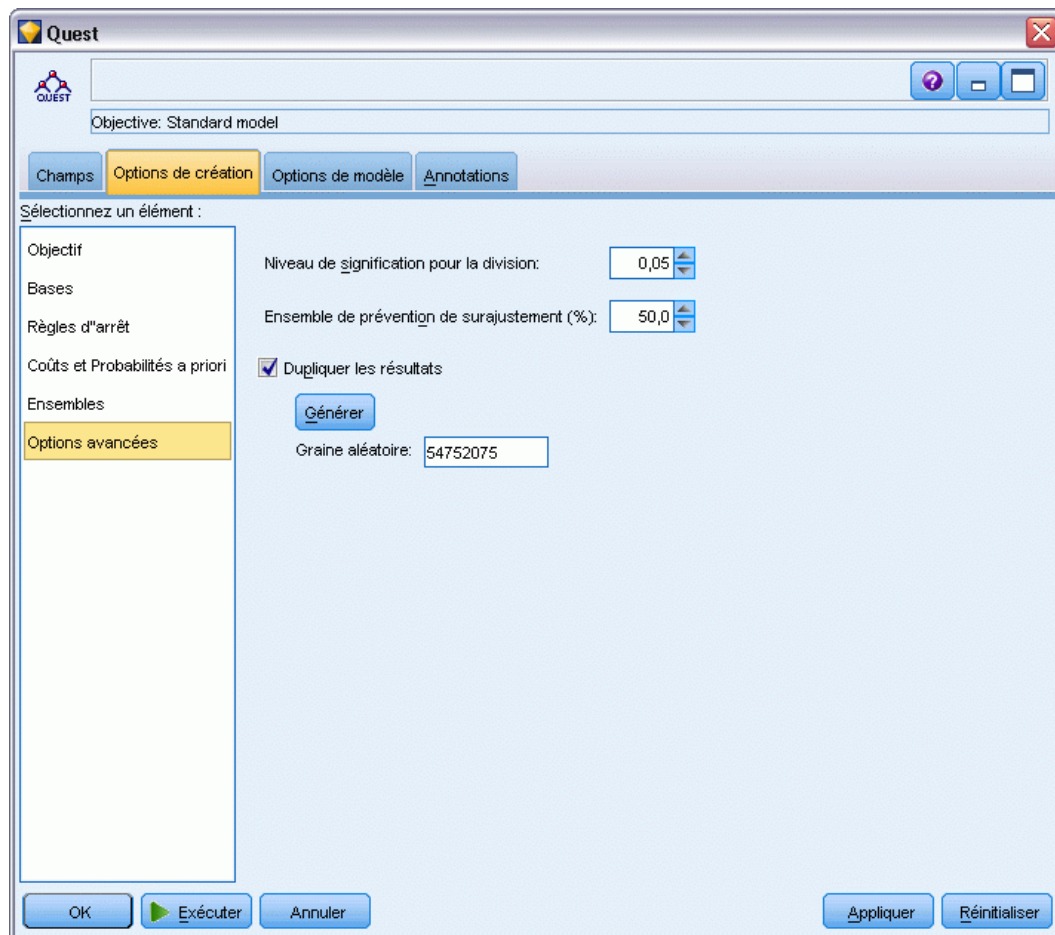
Ensemble de prévention de surajustement. L'algorithme sépare les enregistrements de manière interne en un ensemble de création de modèle et un ensemble de prévention de surajustement, qui est un ensemble d'enregistrements de données servant à effectuer le suivi des erreurs pendant l'apprentissage afin d'éviter que la méthode ne réalise la modélisation d'une variation aléatoire dans les données. Spécifier un pourcentage d'enregistrements. La valeur par défaut est 30.

Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus.

Noeud QUEST - Options avancées

Les options avancées vous permettent d'affiner le processus de génération d'arbres.

Figure 6-32
Définition des options avancées du noeud QUEST



Niveaux de signification de la division. Indique le niveau de signification (α) de la division des noeuds. Cette valeur doit être comprise entre 0 et 1. Les valeurs inférieures ont tendance à créer des arbres avec moins de noeuds.

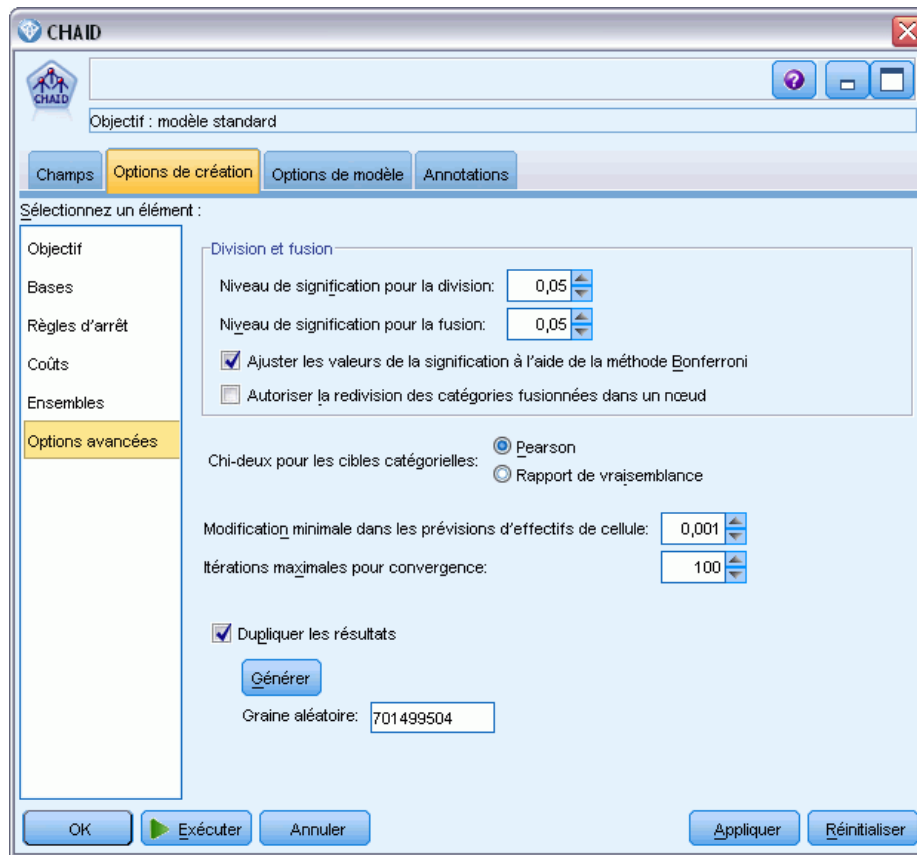
Ensemble de prévention de surajustement. L'algorithme sépare les enregistrements de manière interne en un ensemble de création de modèle et un ensemble de prévention de surajustement, qui est un ensemble d'enregistrements de données servant à effectuer le suivi des erreurs pendant l'apprentissage afin d'éviter que la méthode ne réalise la modélisation d'une variation aléatoire dans les données. Spécifier un pourcentage d'enregistrements. La valeur par défaut est 30.

Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus.

Noeud CHAID - Options avancées

Les options avancées vous permettent d'affiner le processus de génération d'arbres.

Figure 6-33
Définition des options avancées du noeud CHAID



Niveaux de signification de la division. Indique le niveau de signification (α) de la division des noeuds. Cette valeur doit être comprise entre 0 et 1. Les valeurs inférieures ont tendance à créer des arbres avec moins de noeuds.

Niveau de signification pour la fusion. Indique le niveau de signification (α) de la fusion des catégories. Cette valeur doit être supérieure à 0, et inférieure ou égale à 1. Pour empêcher toute fusion de catégories, spécifiez la valeur 1. Pour les cibles continues, cela signifie que le nombre de catégories de la variable de l'arbre final correspond au nombre d'intervalles fourni. Cette option n'est pas disponible pour la méthode Exhaustive CHAID.

Ajuster les valeurs de signification à l'aide de la méthode Bonferroni. Corrige les valeurs de signification lorsque les différentes combinaisons de catégories d'une variable indépendante sont testées. Ces valeurs sont corrigées en fonction du nombre de tests, lui-même en rapport direct avec le nombre de catégories et le niveau de mesure d'une variable indépendante. En général, ce cas de figure est souhaitable, car il permet de contrôler le taux de faux positif. Si vous désactivez cette option, l'analyse effectuée est plus à même de détecter les différences réelles, mais au prix d'un taux de faux positif plus élevé. Il peut être notamment conseillé de désactiver cette option pour les petits échantillons.

Autoriser une nouvelle division des catégories fusionnées dans un noeud. L'algorithme CHAID tente de fusionner les catégories pour créer l'arbre le plus élémentaire décrivant le modèle. Si elle est sélectionnée, cette option permet de rediviser les catégories fusionnées si cette solution se révèle plus favorable.

Chi-deux pour les cibles catégorielles. Pour les cibles catégorielles, vous pouvez indiquer la méthode permettant de calculer les statistiques Chi-deux.

- **Pearson.** Cette méthode fournit des calculs plus rapides mais doit être utilisée avec précaution sur les petits échantillons.
- **Rapport de vraisemblance.** Cette méthode est plus puissante que la méthode Pearson, mais ses calculs s'avèrent plus longs. C'est la méthode la plus adaptée aux petits échantillons. Cette méthode est toujours utilisée pour les cibles continues.

Modification minimale dans les prévisions de fréquence de cellule Lors de l'estimation des fréquences de cellule (pour le modèle nominal et le modèle ordinal d'effets de ligne), une procédure itérative (epsilon) permet d'effectuer une convergence sur l'estimation optimale utilisée pendant le test du Chi-deux d'une division. Epsilon détermine le nombre de changements devant survenir pour que les itérations puissent se poursuivre. Si le nombre de changements produits par la dernière itération est inférieur à la valeur fournie, ces itérations prennent fin. Si vous ne parvenez pas à faire converger l'algorithme, vous pouvez incrémenter cette valeur ou le nombre maximal d'itérations jusqu'à ce que la convergence ait lieu.

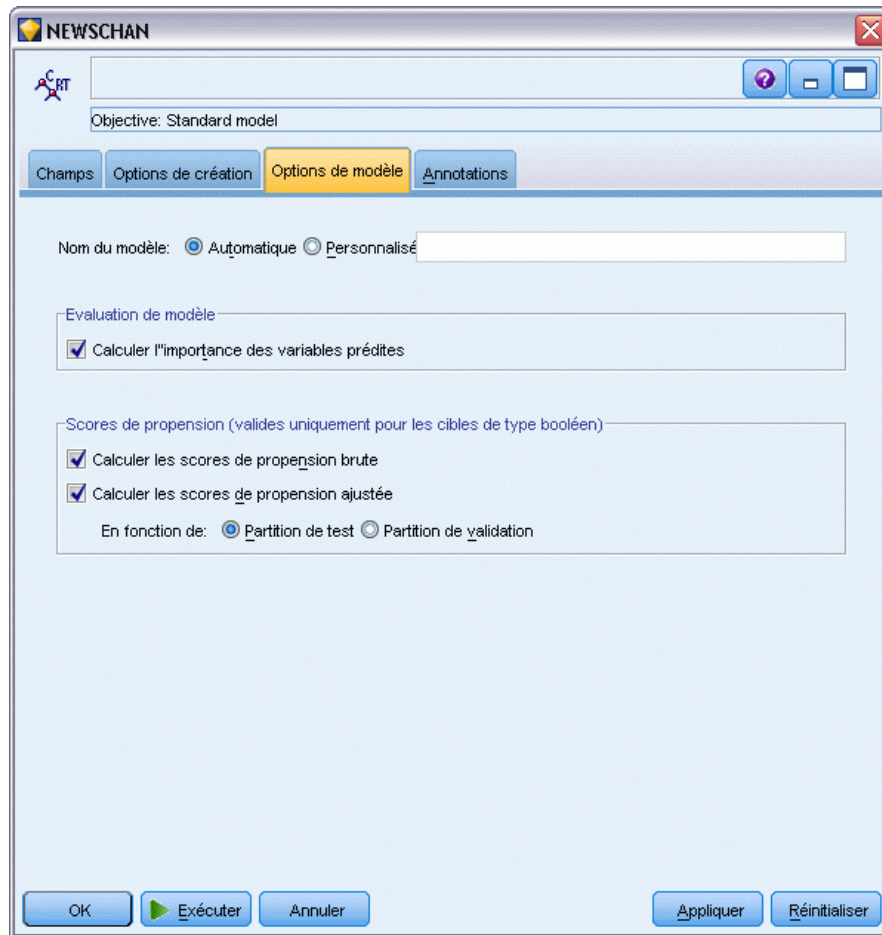
Itérations maximales pour convergence. Indique le nombre maximal d'itérations se produisant avant de s'interrompre, que la convergence ait lieu ou non.

Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus.

Noeud Arbre décision - 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 aussi choisir d'obtenir les informations d'importance des variables indépendantes, ainsi que les scores de propension brute et ajustée pour les cibles boléennes.

Figure 6-34
Définition des options de modèle du noeud 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.

Evaluation de modèle

Calculer l'importance des valeurs prédites. Pour des modèles qui produisent une mesure appropriée d'importance, vous pouvez afficher un graphique qui indique l'importance relative de chaque valeur prédite dans l'estimation du modèle. En général, vous souhaitez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Notez que l'importance des valeurs prédites peut prendre davantage de temps à être calculée pour certains modèles, en particulier si vous travaillez sur de plus grands ensembles de données, et qu'elle est par conséquent désactivée par défaut pour certains modèles. L'importance des valeurs prédites n'est pas disponible pour des modèles de liste de décision. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Scores de propension

Des scores de propension peuvent être établis dans le noeud de modélisation et dans l'onglet Paramètres du nugget de modèle. Cette fonctionnalité n'est disponible que lorsque la cible sélectionnée est un champ booléen. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Calculer les scores de propension brute. Les scores de propension brute sont calculés à partir du modèle uniquement en fonction des données d'apprentissage. Si le modèle prédit la valeur *true (vrai)* (que cette valeur va être la réponse), alors la propension est identique à P, où P est la probabilité de la prédiction. Si le modèle prédit la valeur *false (faux)*, alors la propension est calculée comme $(1 - P)$.

- Si vous choisissez cette option lors de la construction du modèle, des scores de propension sont activés par défaut dans le nugget du modèle. Cependant, vous pouvez toujours choisir d'activer des scores de propension brute dans le nugget du modèle que vous les sélectionnez ou non dans le noeud de modélisation.
- Lors du scoring du modèle, des scores de propension brute seront ajoutés dans un champ avec les lettres *RP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RRP-churn*.

Calculer les scores de propension ajustée. Les propensions brutes sont uniquement basées sur des estimations données par le modèle, qui peut être surajusté, ce qui entraîne des estimations trop optimistes de la propension. Les propensions ajustées tentent de compenser en vérifiant comment le modèle se comporte sur des partitions de test ou de validation et en ajustant les propensions pour donner une meilleure estimation en conséquence.

- Ce paramètre nécessite qu'un champ de partition valide soit présent dans le flux. [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*.](#)
- A la différence des scores de confiance brute, les scores de propension ajustée doivent être calculés lors de la construction du modèle ; sinon, ils ne seront pas disponibles lors du scoring du nugget du modèle.
- Lors du scoring du modèle, des scores de propension ajustée seront ajoutés dans un champ avec les lettres *AP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RAP-churn*. Les scores de propension ajustée ne sont pas disponibles pour des modèles de régression logistique.
- Lors du calcul des scores de propension ajustée, la partition de test ou de validation utilisée pour le calcul ne doit pas avoir été équilibrée. Pour éviter cela, assurez-vous que l'option Equilibrer uniquement les données d'apprentissage est sélectionnée dans tous les noeuds Equilibrer en amont. [Pour plus d'informations, reportez-vous à la section Définition des options du noeud Equilibrer dans le chapitre 3 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#) De plus, si un échantillon complexe a été pris en amont, cela invalide les scores de propension ajustée.
- Les scores de propension ajustée ne sont pas disponibles pour les modèles d'arbre "amélioré" et les modèles d'ensemble de règles. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés sur p. 193.](#)

En fonction de. Pour que les scores de propension ajustée soient calculés, un champ de partition doit être présent dans le flux. Vous pouvez spécifier s'il faut utiliser la partition de test ou de validation pour ce calcul. Pour de meilleurs résultats, la partition de test ou de validation doit comprendre au moins autant d'enregistrements que la partition utilisée pour l'apprentissage du modèle original. [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*.](#)

Noeud C5.0

Remarque : cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Ce nœud utilise l'algorithme C5.0 pour générer un **arbre décision** ou un **ensemble de règles**. Le fonctionnement d'un modèle C5.0 repose sur un découpage de l'échantillon basé sur le champ qui fournit le **gain d'informations** le plus important. Chaque sous-échantillon issu du premier découpage est de nouveau découpé (le modèle utilise généralement un autre champ). Ce processus se répète jusqu'à ce que les sous-échantillons ne puissent plus être découpés. Finalement, les sous-échantillons finaux sont réexaminés : ceux qui n'influent pas de manière significative sur la valeur du modèle sont supprimés ou **élagués**.

Remarque : Le noeud C5.0 ne peut prévoir qu'une cible catégorielle. Lors de l'analyse de données à l'aide de champs catégoriels (ordinaux ou nominaux), le noeud regroupera généralement des catégories différentes de celles des versions C5.0 antérieures à la version 11.0.

Un nœud C5.0 peut produire deux types de modèle. Un **arbre décision** est une description simple des découpages trouvés par l'algorithme. Chaque noeud terminal (ou « feuille » décrit un sous-ensemble particulier des données d'apprentissage ; chacune des observations contenues dans les données d'apprentissage correspond à un seul noeud terminal de l'arbre. Autrement dit, chacun des enregistrements présentés à l'arbre décision ne peut donner lieu qu'à une seule prévision.

En revanche, un **ensemble de règles** tente de générer plusieurs prévisions pour chaque enregistrement. Les ensembles de règles, dérivés des arbres décision, représentent d'une certaine manière une version simplifiée des informations contenues dans l'arbre. Les ensembles de règles sont capables de conserver la plupart des informations importantes d'un arbre décision, selon un modèle moins complexe cependant. Les ensembles de règles n'ont pas les mêmes propriétés que les arbres décision. La principale différence est que dans un ensemble de règles, un enregistrement spécifique peut faire l'objet de plusieurs règles ou bien d'aucune. Si plusieurs règles peuvent s'appliquer à un enregistrement, chacune de ces règles fait l'objet d'un « vote » pondéré basé sur le degré de confiance associée à cette règle. La prévision finale est alors calculée en combinant les votes pondérés de toutes les règles qui s'appliquent à l'enregistrement en question. Si aucune règle ne s'applique à un enregistrement, une prévision par défaut lui est alors attribuée.

Exemple : Un chercheur en médecine a rassemblé des données sur un ensemble de patients, tous souffrant de la même maladie. Lors du traitement, chaque patient a réagi à l'un des cinq médicaments. Vous pouvez utiliser un modèle C5.0, en association avec d'autres noeuds, afin de découvrir le médicament le plus approprié pour un patient futur porteur de la même maladie. [Pour plus d'informations, reportez-vous à la section Traitements par médicaments \(Graphiques exploratoires/C5.0\) dans le chapitre 8 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Pour former un modèle C5.0, il doit exister un champ catégoriel (c'est-à-dire nominal ou ordinal) *Cible* et un ou plusieurs champs *Entrée* de tout type. Les champs paramétrés sur *Les deux* ou *Aucun* sont ignorés. Les types des champs utilisés dans ce modèle doivent être complètement instanciés. Un champ poids peut aussi être spécifié.

Puissance. Les modèles C5.0 s'avèrent relativement solides en présence de problèmes tels que des données manquantes ou un grand nombre de champs. Leur temps d'apprentissage est généralement court. En outre, les modèles C5.0 sont généralement plus faciles à comprendre que d'autres types de modèle dans la mesure où les règles extraites de ces modèles ne sont pas difficiles à interpréter. C5.0 propose également une **méthode** d'amélioration qui permet d'accroître la précision de la classification.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de création des modèles C5.0. [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*.](#)

Noeud C5.0 - Options du modèle

Figure 6-35
Noeud C5.0 - Options du modèle

The screenshot shows the 'C5.0' dialog box with the following settings:

- Nom du modèle:** Radio buttons for 'Auto' (selected) and 'Personnalisé'. An empty text field is to the right.
- Utiliser les données partitionnées
- Créer un modèle pour chaque division
- Type de sortie:** Radio buttons for 'Arbre décision' (selected) and 'Ensemble de règles'.
 - Grouper les valeurs symboliques
 - Utiliser l'amélioration. Nombre d'essais: 10
 - Effectuer la validation croisée. Nombre de validations croisées: 10
- Mode:** Radio buttons for 'Simple' (selected) and 'Expert'.
 - Préférence:** Radio buttons for 'Précision' (selected) and 'Généralité'.
 - Bruit théorique (%):** Spin box set to 0.

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

Nom du modèle. Indiquez le nom du modèle à générer.

- **Auto** : Lorsque cette option est sélectionnée, le nom du modèle est généré automatiquement, sur la base des noms du champ cible. Il s'agit de la valeur par défaut.
- **Personnalisée**. Sélectionnez cette option pour indiquer votre nom pour le nugget de modèle qui sera créé par ce noeud.

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*.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Type de sortie. Spécifiez le type de nugget de modèle : Arbre décision ou Ensemble de règles.

Grouper les valeurs symboliques. Lorsque cette option est sélectionnée, le nœud C5.0 tente de combiner les valeurs symboliques présentant des tendances similaires en ce qui concerne le champ de sortie. Si elle n'est pas sélectionnée, le nœud C5.0 crée un nœud enfant pour chacune des valeurs du champ symbolique utilisé pour le découpage du nœud parent. Par exemple, si le nœud C5.0 découpe un champ *Couleur* (comprenant les valeurs *ROUGE*, *VERT* et *BLEU*), il crée un découpage en trois groupes par défaut. Cependant, si cette option est sélectionnée et que les enregistrements dans lesquels *Couleur* = *ROUGE* sont très similaires à ceux dans lesquels *Couleur* = *BLEU*, le nœud effectue un découpage en deux groupes, avec les valeurs *VERT* dans un groupe, et les valeurs *BLEU* et *ROUGE* dans un autre.

Utiliser l'amélioration. Pour améliorer sa précision, l'algorithme dispose d'une méthode spéciale appelée **amélioration**. Cette méthode repose sur la création séquentielle de plusieurs modèles. Le premier modèle est généré de manière classique. Un deuxième modèle est généré ; il contient essentiellement les enregistrements qui n'ont pas été correctement classés par le premier modèle. Un troisième modèle contenant les erreurs du deuxième arbre est ensuite généré, et ainsi de suite. Au final, l'ensemble des modèles est appliqué aux observations à l'aide d'une procédure de vote pondéré qui permet de combiner les différentes prévisions en une prévision globale. Si le processus d'amélioration peut accroître de façon significative la précision des modèles C5.0, son utilisation requiert également un temps d'apprentissage plus long. L'option Nombre d'essais permet de contrôler le nombre de modèles utilisés pour générer le modèle amélioré. Cette fonction est basée sur les recherches de Freund & Schapire, associées à des améliorations propriétaires permettant de mieux gérer les données parasites.

Effectuer la validation croisée. Si cette option est sélectionnée, le nœud C5.0 utilise un ensemble de modèles générés à partir de sous-ensembles des données d'apprentissage pour évaluer le degré de précision d'un modèle généré à partir de la totalité des données. Cette option peut se révéler utile lorsque votre ensemble de données n'est pas assez important pour effectuer un découpage traditionnel en ensembles d'apprentissage et de test. Les modèles de validation croisée sont supprimés une fois l'estimation de la précision calculée. Vous pouvez spécifier le **nombre de validations croisées** ou le nombre de modèles utilisés pour la validation croisée. Notez que dans les précédentes versions de IBM® SPSS® Modeler, la construction et la validation croisée du modèle étaient deux opérations distinctes. Dans la version actuelle, aucune étape de construction

de modèle distincte n'est nécessaire. La construction et la validation croisée du modèle sont effectuées en même temps.

Mode. Pour un apprentissage simple, la plupart des paramètres du nœud C5.0 sont automatiquement configurés. L'apprentissage expert vous procure un contrôle plus direct des paramètres de l'apprentissage.

Options du mode simple

Préférence. Par défaut, le nœud C5.0 tente de générer l'arbre le plus précis possible. Cela peut parfois provoquer des problèmes de surajustement, ce qui peut se traduire par des performances médiocres lorsque vous appliquez le modèle à de nouvelles données. Sélectionnez l'option Généralité pour utiliser les paramètres de l'algorithme qui sont les moins susceptibles de provoquer ce problème.

Remarque : les modèles générés avec l'option Généralité sélectionnée ne généralisent pas nécessairement mieux que d'autres modèles. Pour éviter ce problème, évaluez toujours votre modèle à l'aide d'un échantillon.

Bruit théorique (%). Spécifiez le pourcentage de données superflues ou erronées prévu dans l'ensemble des données d'apprentissage.

Options du mode expert

Taux d'élagage. Permet de déterminer l'importance de l'élagage qui sera effectué sur l'arbre de décision ou l'ensemble de règles généré. Plus vous augmentez cette valeur, plus votre arbre sera concis. Plus vous la réduisez, plus votre arbre sera précis. Ce paramètre a une incidence sur l'élagage local uniquement (reportez-vous à la rubrique « Utiliser l'élagage global » ci-dessous).

Nombre minimal d'enregistrements par branche enfant. Vous pouvez modifier la taille des sous-groupes afin de limiter le nombre de séparations que comportent les branches de votre arbre. Une branche sera divisée uniquement si au moins deux des sous-branches obtenues contiennent le nombre minimum d'enregistrements de données d'apprentissage spécifié. La valeur par défaut est 2. Vous pouvez augmenter cette valeur pour éviter le phénomène de **surentraînement** en cas de données parasites.

Utiliser l'élagage global. Deux étapes sont nécessaires pour élaguer un arbre : L'étape d'élagage local examine d'abord les sous-arbres et réduit les branches pour augmenter la précision du modèle. Au cours de l'élagage global, l'arbre est considéré dans son ensemble et les sous-arbres faibles peuvent être réduits. L'élagage global est effectué par défaut. Pour passer outre à l'élagage global, désélectionnez l'option correspondante.

Attributs Winnow. Si cette option est sélectionnée, C5.0 étudie l'utilité des variables indépendantes avant de commencer la construction du modèle. Les variables indépendantes considérées comme non pertinentes sont ensuite exclues du processus de construction du modèle. Cette option peut être utile pour les modèles contenant de nombreux champs variables indépendantes et peut permettre d'éviter le surajustement.

Remarque : L'activation du traitement parallèle peut profiter à la vitesse de création des modèles C5.0. [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*.](#)

Nuggets de modèle Arbre de décision

Les nuggets de modèle d'arbre décision représentent les arborescences utilisées pour effectuer une prévision sur un champ de sortie spécifique détecté par l'un des noeuds de modélisation d'arbre décision (Arbre C&RT, CHAID, QUEST, ou C5.0). Vous pouvez générer les modèles d'arbres directement à partir du noeud de génération d'arbre ou de manière indirecte à partir du générateur d'arbres interactif. [Pour plus d'informations, reportez-vous à la section Générateur d'arbres interactifs sur p. 131.](#)

Scoring des modèles d'arbre

Lorsque vous exécutez un flux contenant un nugget de modèle d'arbre, le résultat obtenu dépend du type d'arbre.

- Pour les arbres de classification supervisée (cible catégorielle), deux nouveaux champs, l'un contenant la valeur prédite et l'autre la confiance de chaque enregistrement, sont ajoutés aux données. La prévision est basée sur la catégorie la plus fréquente du noeud terminal auquel l'enregistrement est affecté ; si une majorité des personnes sondées dans un noeud donné présentent la valeur « oui », la prévision de tous les enregistrements affectés à ce noeud est « oui ».
- Dans le cas des arbres de régression, seules les variables indépendantes sont générées ; aucune confiance n'est affectée.
- Vous pouvez également ajouter, pour les modèles CHAID, QUEST et C&RT, un champ supplémentaire qui indique l'ID du noeud auquel chaque enregistrement est affecté.

Les noms des nouveaux champs sont constitués du nom du modèle auquel des préfixes sont ajoutés. Pour les arbres C&RT, CHAID et QUEST, le préfixe \$R- est utilisé pour désigner le champ de prévision, le préfixe \$RC- pour désigner le champ contenant le degré de confiance et le préfixe \$RI- pour désigner le champ contenant l'identificateur de noeud. Pour les arbres C5.0, le préfixe \$C- est utilisé pour désigner le champ de prévision et le préfixe \$CC- est utilisé pour désigner le champ contenant le degré de confiance. S'il existe plusieurs noeuds de modèle d'arbre, le nom des nouveaux champs comporte un *préfixe* numérique permettant de les distinguer en cas de besoin (par exemple, \$RI- et \$RC1-, \$R2-).

Travailler avec les nuggets de modèle d'arbre

Vous pouvez enregistrer ou exporter des informations liées au modèle de plusieurs manières.

Remarque : Ces options sont, pour la plupart, également disponibles à partir de la fenêtre du générateur d'arbres.

A partir du générateur d'arbres ou d'un nugget de modèle d'arbre, vous pouvez effectuer les opérations suivantes :

- Générer un noeud Filtrer ou Sélectionner à partir de l'arbre actuel. [Pour plus d'informations, reportez-vous à la section Génération de noeuds Filtrer et Sélectionner sur p. 154.](#)
- Générer un noeud Ensemble de règles représentant l'arborescence sous la forme d'un ensemble de règles définissant les branches terminales de l'arbre. [Pour plus d'informations, reportez-vous à la section Génération d'un noeud Ensemble de règles à partir d'un noeud Arbre décision sur p. 154.](#)
- En outre, vous pouvez exporter le modèle au format PMML pour les nuggets de modèle d'arbre uniquement. [Pour plus d'informations, reportez-vous à la section Palette Modèles dans le chapitre 3 sur p. 52.](#) Notez que si le modèle comporte des divisions personnalisées, ces informations ne sont pas conservées dans le fichier PMML exporté. (Les divisions sont conservées, mais il n'est plus possible de les distinguer de celles choisies par l'algorithme.)
- Générer un graphique à partir d'une partie sélectionnée de l'arbre actuel. *Remarque* : ceci fonctionne pour un nugget uniquement lorsqu'il est relié à d'autres noeuds dans un flux. [Pour plus d'informations, reportez-vous à la section Génération de graphiques sur p. 195.](#)
- Pour les modèles C5.0 améliorés uniquement, vous pouvez sélectionner Arbre décision unique (espace de travail) ou Arbre décision unique (onglet Modèles) pour créer un ensemble de règles unique à partir de la règle sélectionnée. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés sur p. 193.](#)

Remarque : Bien que le noeud Bâtir règles ait été remplacé par le noeud Arbre C&RT, les noeuds Arbre décision figurant dans des flux existants et ayant été créés à partir d'un noeud Bâtir règles continueront de fonctionner correctement.

Nuggets de modèle d'arbre unique

Si vous sélectionnez Créer un seul arbre comme objectif principal sur le noeud de modélisation, le nugget de modèle généré contient les onglets suivants.

Tabulation	Description	Informations supplémentaires
Modèle	Affiche les règles définissant le modèle.	Pour plus d'informations, reportez-vous à la section Règles du modèle Arbre décision sur p. 186.
Afficheur	Affiche l'arborescence du modèle.	Pour plus d'informations, reportez-vous à la section Afficheur du modèle Arbre décision sur p. 190.
Récapitulatif	Affiche des informations sur les champs, les paramètres de création et le processus d'estimation du modèle.	Pour plus d'informations, reportez-vous à la section Récapitulatif /Informations sur les nuggets de modèle dans le chapitre 3 sur p. 56.

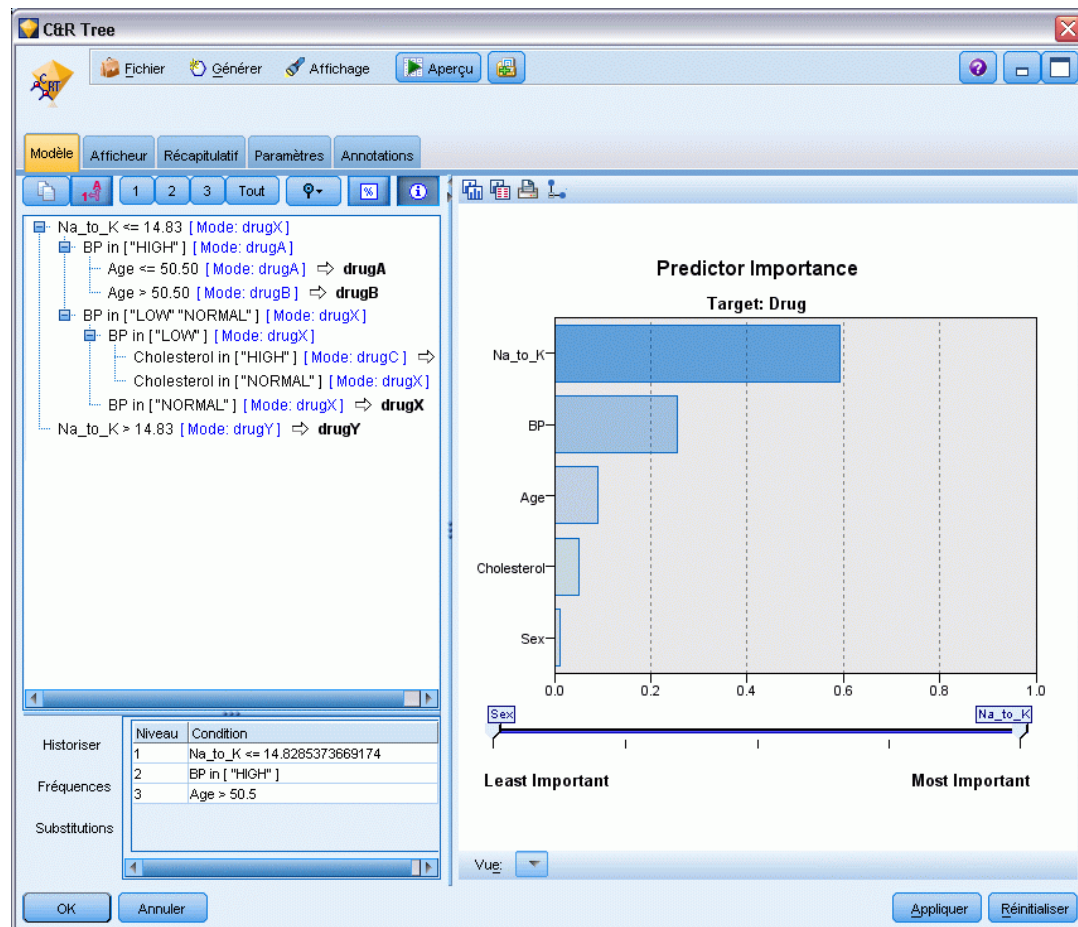
Tabulation	Description	Informations supplémentaires
Paramètres	Vous permet de spécifier des options de confiance et de génération SQL lors de l'évaluation du modèle.	Pour plus d'informations, reportez-vous à la section Paramètres du modèle Arbre décision/Nugget de modèle d'ensemble de règles sur p. 191.
Annotation	Vous permet d'ajouter des annotations descriptives, de spécifier un nom personnalisé, d'ajouter du texte aux info-bulles et de définir des mots-clés de recherche pour le modèle.	Pour plus d'informations, reportez-vous à la section Annotations dans le chapitre 5 dans <i>Guide de l'utilisateur de IBM SPSS Modeler 15</i>.

Règles du modèle Arbre décision

L'onglet Modèle d'un nugget d'arbre décision affiche les règles qui définissent le modèle. Si nécessaire, un graphique d'importance des variables indépendantes et un troisième panneau comportant des informations sur l'historique, les effectifs et les substitutions peuvent également s'afficher.

Remarque : Lorsque vous sélectionnez l'option Créer un modèle pour des ensembles de données très volumineux dans l'onglet Options de création du noeud CHAID (volet Objectif), l'onglet Modèle affiche uniquement les détails de règles de l'arbre.

Figure 6-36
Nugget du modèle d'arbre décision



Règles d'arbre

Dans le volet gauche s'affiche la liste des conditions définissant la technique de partitionnement des données découvertes par l'algorithme. Il s'agit essentiellement d'un ensemble de règles pouvant être utilisées pour affecter les enregistrements à des noeuds enfant, sur la base des différentes variables indépendantes.

Les arbres décision séparent les données de manière récursive en fonction des valeurs des champs d'entrée. Ces séparations de données sont appelées des **branches**. La branche initiale (également appelée **racine**) englobe tous les enregistrements de données. La racine est divisée en sous-ensembles ou **branches enfant**, en fonction de la valeur d'un champ d'entrée spécifique. Chaque branche enfant peut elle-même être divisée en sous-branches qui, à leur tour, peuvent être de nouveau divisées, et ainsi de suite. Les branches situées au dernier niveau de l'arbre ne peuvent pas être divisées. De telles branches sont appelées des **branches terminales** (ou des **feuilles**).

Le navigateur de règles affiche les valeurs d'entrée définissant chaque séparation ou branche, ainsi qu'un récapitulatif des valeurs de champ de sortie des enregistrements situés au niveau de cette séparation. Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle](#).

Pour les séparations basées sur des champs numériques, les branches sont représentées par une ligne ayant la forme suivante :

```
fieldname relation value [summary]
```

où *relation* représente une relation numérique. Par exemple, une branche définie par des valeurs du champ *revenue* (recette) supérieures à 100 apparaîtra de la manière suivante

```
revenue > 100 [summary]
```

Pour les séparations basées sur des champs symboliques, les branches sont représentées par une ligne de la forme :

```
fieldname = value [summary] ou fieldname in [values] [summary]
```

où *values* correspond aux valeurs de champ définissant la branche. Par exemple, une branche comportant des enregistrements pour lesquels la valeur du champ *région* peut être *North*, *West* ou *South* sera représentée comme suit :

```
region ["North" "West" "South"] [summary]
```

Une prévision est également fournie pour les branches terminales, une flèche et la valeur prédite étant ajoutées à la fin de la condition de la règle. Par exemple, une branche terminale définie par *revenue > 100* et prédisant la valeur *high* (élevé) pour le champ de sortie apparaît comme suit :

```
revenue >> 100 [Mode: high] • high
```

Le **résumé** de la branche est différent pour les champs de sortie symboliques et numériques. Pour les arbres contenant des champs de sortie numériques, la valeur **moyenne** et l'**effet** de la branche (différence entre la moyenne de la branche et la moyenne de sa branche parent) sont affichés. Pour les arbres contenant des champs de sortie symboliques, la valeur **modale** (la valeur la plus fréquente) des enregistrements de la branche est affichée.

Pour décrire complètement une branche, vous devez non seulement inclure la condition définissant la branche, mais également les conditions figurant en amont dans l'arbre qui définissent les séparations. Par exemple, dans l'arbre :

```
revenue > 100
region = "North"
region in ["South" "East" "West"]
revenue <= 200
```

la branche représentée par la deuxième ligne est définie par les conditions *revenue > 100* et *région = « Nord »*.

Si vous cliquez sur **Afficher instances/confiance** dans la barre d'outils, chaque règle affiche également des informations sur le nombre d'enregistrements auxquels la règle s'applique (**Instances**) et la proportion de ces enregistrements pour lesquels la règle est vraie (**Confiance**).

Importance des valeurs prédites

Un graphique illustrant l'importance relative de chaque valeur prédite dans l'estimation du modèle peut également être affiché dans l'onglet Modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique n'est disponible que si Calculer l'importance des valeurs prédites a été sélectionné dans l'onglet Analyse avant la génération du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Informations supplémentaires sur les modèles

Si vous cliquez sur Afficher un panneau d'informations supplémentaire dans la barre d'outils, un panneau apparaît, au bas de la fenêtre, contenant des informations détaillées sur la règle sélectionnée. Le panneau d'informations contient trois onglets.

Figure 6-37
Substitutions affichées dans le panneau d'informations

Règle	Cholesterol in ["NORMAL"]
1	Age > 30
2	Na_to_K > 10.6886346572169

Historique. Cet onglet indique les conditions de division depuis le noeud racine jusqu'au noeud sélectionné. Vous obtenez ainsi la liste des conditions qui déterminent le moment où un enregistrement est affecté au noeud sélectionné. Les enregistrements pour lesquels toutes les conditions sont vraies (true) sont affectés à ce noeud.

Effectifs. Pour les modèles avec des champs cible symboliques, cet onglet indique, pour chaque valeur cible possible, le nombre d'enregistrements affectés à ce noeud (dans les données d'apprentissage) qui ont cette valeur cible. La valeur d'effectif, exprimée en pourcentage (au maximum trois chiffres après la virgule), est également affichée. Pour les modèles ayant des cibles numériques, cet onglet est vide.

Substitutions. Cet onglet affiche toutes les substitutions du champ de division principal pour le noeud sélectionné. Les substitutions sont des champs de remplacement utilisés si la valeur prédite principale d'un enregistrement déterminé est manquante. Le nombre de substitutions maximal autorisé pour une division donnée est indiqué dans le noeud de création de modèle, mais dépend en fait des données d'apprentissage. En règle générale, plus il manque de données, plus les substitutions ont de chances d'être utilisées. Pour les autres modèles d'arbres décision, cet onglet est vide.

Remarque : Pour faire partie du modèle, les substitutions doivent être identifiées au cours de la phase d'apprentissage. Si l'échantillon d'apprentissage ne comporte aucune valeur manquante, aucune substitution n'est identifiée. Les enregistrements contenant des valeurs manquantes repérés au cours du test ou du scoring sont alors automatiquement placés dans le noeud enfant comptant le plus grand nombre d'enregistrements. Si vous prévoyez la détection de valeurs manquantes lors

du test ou du scoring, veillez à ce que ces valeurs soient également manquantes dans l'échantillon d'apprentissage. Les substitutions ne sont pas disponibles avec les arbres CHAID.

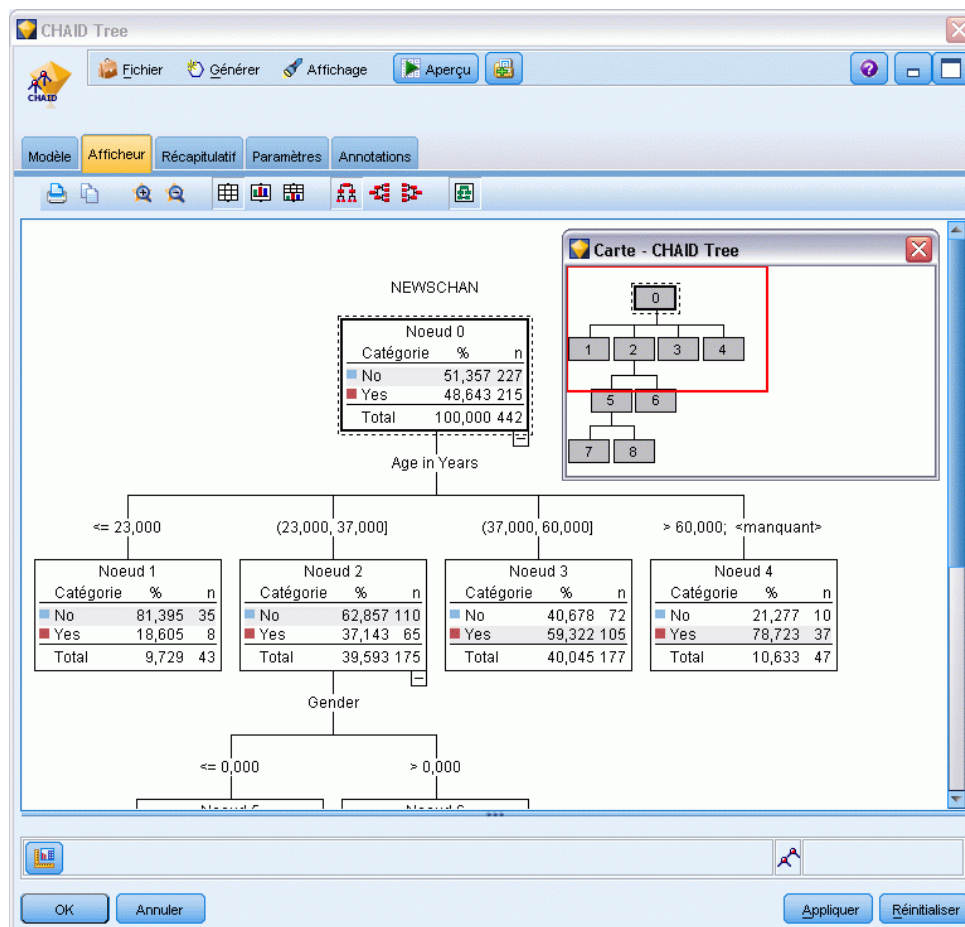
Afficheur du modèle Arbre décision

L'onglet Afficheur d'un nugget de modèle d'arbre décision ressemble au contenu du générateur d'arbres. La principale différence réside dans le fait que vous ne pouvez pas développer ou modifier l'arbre lorsque vous parcourez le nugget de modèle. Les autres options de visualisation et de personnalisation de l'affichage sont identiques pour ces deux composants. [Pour plus d'informations, reportez-vous à la section Personnalisation de l'affichage d'un arbre sur p. 138.](#)

Remarque : L'onglet Afficheur n'est pas affiché pour la création de nuggets de modèle CHAID si vous sélectionnez l'option Créer un modèle pour des ensembles de données très volumineux dans le panneau Objectif de l'onglet Options de création.

Figure 6-38

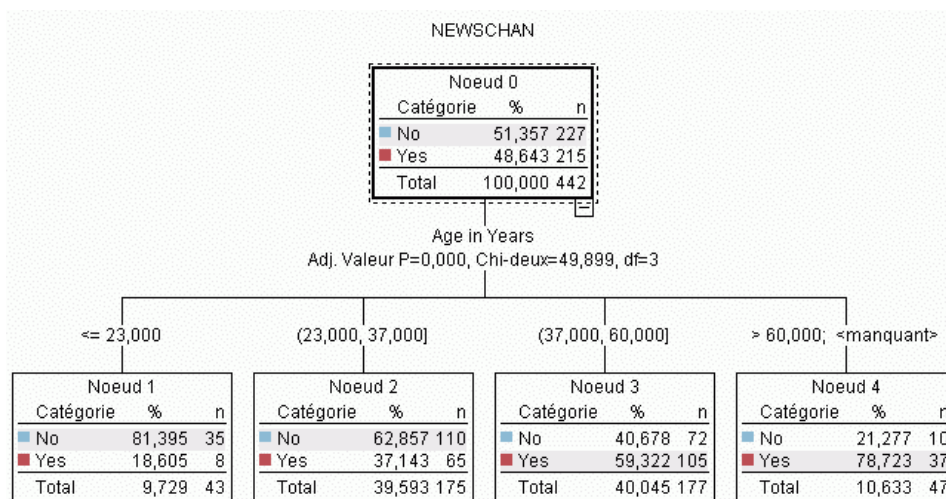
Nœud Arbre décision - Onglet Afficheur avec fenêtre de la carte de l'arbre



Lorsque vous affichez des règles de division dans l'onglet Afficheur, les crochets signifient que la valeur voisine est comprise dans l'intervalle alors que les parenthèses indiquent que la valeur voisine est exclue de l'intervalle. L'expression (23,37] signifie par conséquent de 23 non compris à 37 compris ; autrement dit, de la valeur immédiatement supérieure à 23 à 37. Dans l'onglet Modèle, la même condition serait affichée ainsi :

Age > 23 and Age <= 37

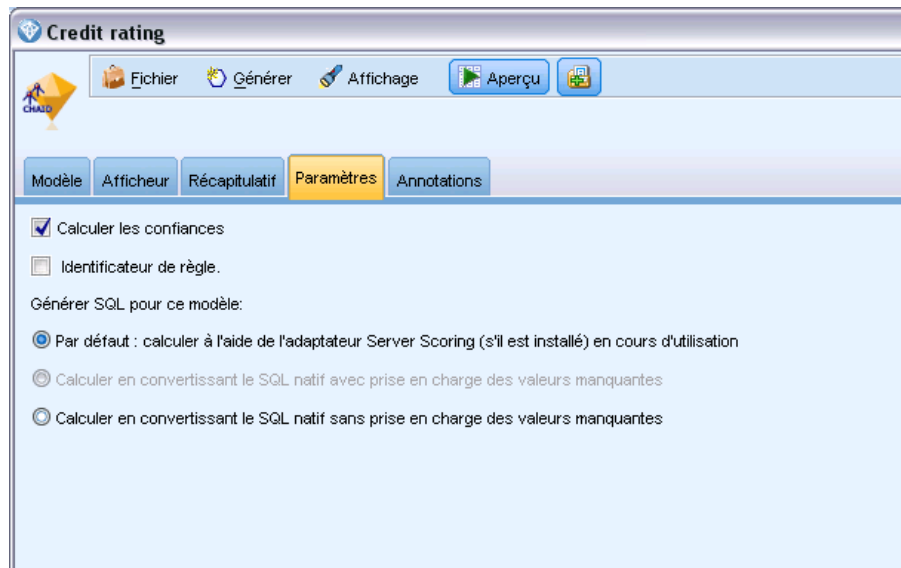
Figure 6-39
Règles de division affichées dans l'onglet Afficheur



Paramètres du modèle Arbre décision/Nugget de modèle d'ensemble de règles

L'onglet Paramètres d'un modèle Arbre décision ou d'un nugget de modèle Ensemble de règles généré vous permet d'indiquer les options de confiance et de génération SQL au cours du scoring du modèle. Cet onglet n'est disponible qu'une fois le nugget de modèle ajouté à un flux.

Figure 6-40
Exemple de paramètres de nugget de modèle d'Arbre décision



Calculer les confiances. Sélectionnez cette option pour inclure des confiances dans les opérations de scoring. L'exclusion des confiances lors du scoring des modèles de la base de données vous permet de générer des instructions SQL plus efficaces. Dans les arbres de régression, les confiances ne sont pas affectées.

Remarque : Lorsque vous sélectionnez l'option Créer un modèle pour des ensembles de données très volumineux sur le panneau Méthode de l'onglet Options de création pour les modèles CHAID, cette case n'est disponible que dans les nuggets du modèle pour les cibles catégorielles de champ booléen ou nominal.

Calculer les scores de propension brute. Pour les modèles disposant d'une variable cible booléenne (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

Remarque : Lorsque vous sélectionnez l'option Créer un modèle pour des ensembles de données très volumineux sur le panneau Méthode de l'onglet Options de création pour les modèles CHAID, cette case n'est disponible que dans les nuggets du modèle pour les cibles catégorielles de champ booléen.

Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le nœud de modélisation avant la génération du modèle.

Remarque : Les scores de propension ajustés ne sont pas disponibles pour les modèles arbre amélioré et ensemble de règles. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés sur p. 193.](#)

Identificateur de règle. Pour les modèles CHAID, QUEST et C&RT, cette option permet d'ajouter à la sortie du scoring un champ qui indique l'ID du noeud terminal auquel chaque enregistrement est affecté.

Remarque : Lorsque cette option est sélectionnée, la génération SQL n'est pas disponible.

Générer SQL pour ce modèle. Lorsque vous utilisez des données provenant d'une base de données, le code SQL peut être renvoyé à la base de données pour exécution, ce qui assure des performances supérieures pour de nombreuses opérations. [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.*](#)

Sélectionnez une des options suivantes pour spécifier comment la génération de SQL est effectuée.

- **Par défaut : évaluer à l'aide de l'adaptateur de scoring de serveur (si installé) en cours d'exécution.** Si vous êtes connecté à une base de données qui contient un adaptateur de scoring, génère le SQL à l'aide de l'adaptateur de scoring, sinon génère le SQL de manière native dans SPSS Modeler.
- **Générer sans prise en charge des valeurs manquantes.** Sélectionnez cette option pour activer la génération SQL sans vous préoccuper des valeurs manquantes. Cette option détermine simplement la prévision sur null (\$null\$) lorsqu'une valeur manquante apparaît au cours de la détermination des scores d'une observation.

Remarque : Vous ne pouvez pas utiliser cette option pour les modèles CHAID. Pour les autres types de modèles, elle n'est disponible que pour les arbres décision (pas pour les ensembles de règles).

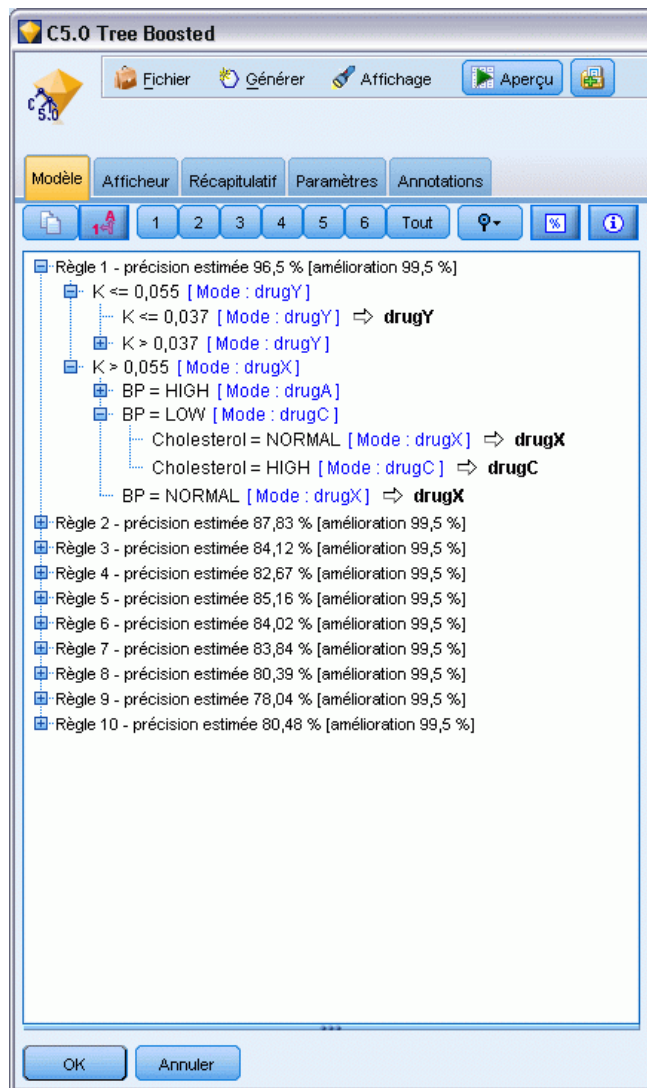
- **Générer avec prise en charge des valeurs manquantes.** Pour les modèles d'arbre CHAID, QUEST et C&RT, vous pouvez activer la génération SQL avec prise en charge complète des valeurs manquantes. Le langage SQL est généré afin que les valeurs manquantes soient gérées conformément aux spécifications du modèle. Par exemple, les arbres C&RT utilisent des règles de substitution et le noeud enfant le plus élevé.

Remarque : Pour les modèles C5.0, elle n'est disponible que pour les ensembles de règles (pas pour les arbres décision).

Modèles C5.0 améliorés

Remarque : cette fonction est disponible dans SPSS Modeler Professional et SPSS Modeler Premium.

Figure 6-41
Nugget de modèle C5.0 amélioré - Onglet Modèle



Lorsque vous créez un modèle C5.0 amélioré (que ce soit un ensemble de règles ou un arbre de décision), vous créez en fait un ensemble de modèles connexes. Le navigateur de règles de modèle pour un modèle C5.0 amélioré affiche la liste des modèles figurant au niveau supérieur de la hiérarchie, ainsi que la précision estimée de chaque modèle et la précision globale de l'ensemble des modèles améliorés. Pour examiner les règles ou les divisions d'un modèle spécifique, sélectionnez le modèle et développez-le de la même manière que vous développez une règle ou une branche d'un modèle.

Vous pouvez également extraire un modèle spécifique de l'ensemble de modèles améliorés et créer un nouveau nugget de modèle Ensemble de règles généré contenant uniquement ce modèle. Pour créer un ensemble de règles à partir d'un modèle C5.0 amélioré, sélectionnez l'ensemble de règles ou l'arbre qui vous intéresse, et choisissez Arbre décision unique (palette GM) ou Arbre décision unique (espace de travail) dans le menu Générer.

Génération de graphiques

Les noeuds Arbre fournissent de nombreuses informations ; toutefois celles-ci ne sont pas toujours dans un format facilement accessible pour les utilisateurs professionnels. Pour fournir les données d'une manière facilement incorporable aux rapports d'activité, aux présentations, etc., vous pouvez réaliser des diagrammes à partir des données sélectionnées. Par exemple, à partir de l'un des onglets Modèle ou Afficheur d'un nugget de modèle, ou à partir de l'onglet Afficheur d'un arbre interactif, vous pouvez générer un graphique pour une partie sélectionnée de l'arbre, et ainsi ne générer un graphique que pour les observations du noeud de la branche ou de l'arbre sélectionné.

Remarque : Vous ne pouvez générer un graphique qu'à partir d'un nugget attaché à d'autres noeuds dans un flux.

Générer un diagramme

La première étape consiste à sélectionner les informations à afficher dans le graphique :

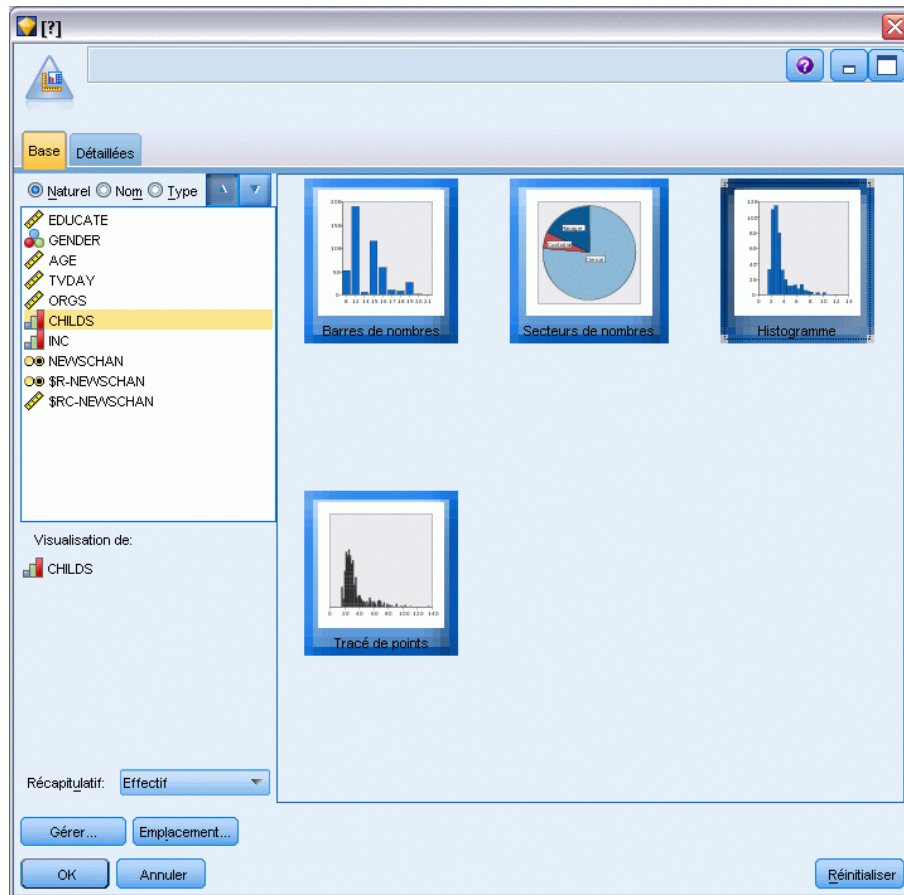
- Sur l'onglet Modèle d'un nugget, développez la liste de conditions et de règles dans le panneau gauche et sélectionnez celle qui vous intéresse.
- Sur l'onglet Afficheur d'un nugget, développez la liste de branches et sélectionnez le noeud qui vous intéresse.
- Sur l'onglet Afficheur d'un arbre interactif, développez la liste de branches et sélectionnez le noeud qui vous intéresse.

Remarque : Vous ne pouvez pas sélectionner le premier noeud de l'un des onglets Afficheur.

La façon de créer un graphique est la même, quelle que soit la façon dont vous sélectionnez les données à afficher :

- ▶ Dans le menu Générer, sélectionnez Graphique (dans la sélection) ; vous pouvez également cliquer sur le bouton Graphique (dans la sélection), dans l'onglet Afficheur, dans le coin inférieur gauche. L'onglet Base de Représentation graphique apparaît.

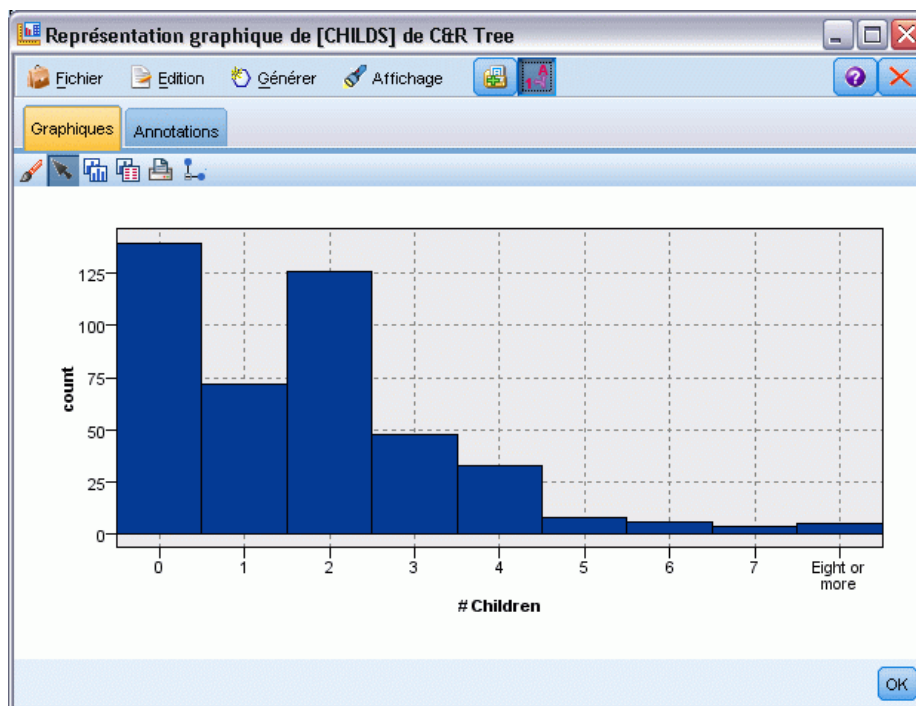
Figure 6-42
Boîte de dialogue Noeud Représentation graphique, onglet de Base



Remarque : Seuls les onglets de Base et Détaillé seront disponibles lorsque vous affichez le graphique de cette manière. [Pour plus d'informations, reportez-vous à la section Noeud Représentation Graphique dans le chapitre 5 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

- ▶ À l'aide des paramètres de l'onglet de Base ou Détaillé, spécifiez les détails à afficher sur le diagramme.
- ▶ Cliquez sur OK pour générer le diagramme.

Figure 6-43
Histogramme généré à partir de l'onglet de base de la représentation graphique



Le titre du graphique identifie les noeuds ou les règles que vous avez choisis d'inclure.

Nuggets de modèle pour l'amélioration, le regroupement et les ensembles très volumineux

Si vous sélectionnez Améliorer la précision du modèle (amélioration), Améliorer la stabilité du modèle (agrégation), ou Créer un modèle pour des ensembles de données très volumineux comme objectif principal sur le noeud de modélisation, IBM® SPSS® Modeler crée un ensemble de modèles multiples. [Pour plus d'informations, reportez-vous à la section Modèles pour des ensembles dans le chapitre 3 sur p. 60.](#)

Le nugget de modèles généré contient les onglets suivants. L'onglet Modèle fournit plusieurs vues différentes du modèle.

Tabulation	Affichage	Description	Informations supplémentaires
Modèle	Récapitulatif des modèles	Affiche un récapitulatif de la qualité de l'ensemble et de sa diversité (sauf pour les modèles améliorés et les cibles continues), une mesure de la variation des	Pour plus d'informations, reportez-vous à la section Récapitulatif de modèle dans le chapitre 3 sur p. 62.

Tabulation	Affichage	Description	Informations supplémentaires
		prédictions sur les différents modèles.	
	Importance des valeurs prédites	Affiche un graphique qui indique l'importance relative de chaque variable indépendante (champ d'entrée) dans l'évaluation du modèle.	Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 63.
	Fréquence des variables indépendantes	Affiche un graphique qui indique la fréquence relative avec laquelle chaque variable indépendante est utilisée dans l'ensemble de modèles.	Pour plus d'informations, reportez-vous à la section Fréquence des valeurs prédites dans le chapitre 3 sur p. 64.
	Précision des modèles de composant	Graphique affichant la précision prédictive de chacun des modèles différents de l'ensemble.	
	Détails des modèles de composant	Affiche des informations sur chacun des différents modèles de l'ensemble.	Pour plus d'informations, reportez-vous à la section Détails d'un modèle de composant dans le chapitre 3 sur p. 67.
	Informations	Affiche des informations sur les champs, les paramètres de création et le processus d'estimation du modèle.	Pour plus d'informations, reportez-vous à la section Récapitulatif /Informations sur les nuggets de modèle dans le chapitre 3 sur p. 56.
Paramètres		Vous permet d'inclure des confiances dans les opération de scoring.	Pour plus d'informations, reportez-vous à la section Paramètres du modèle Arbre décision/Nugget de modèle d'ensemble de règles sur p. 191.
Annotation		Vous permet d'ajouter des annotations descriptives, de spécifier un nom personnalisé, d'ajouter du texte aux info-bulles et de définir des mots-clés de recherche pour le modèle.	Pour plus d'informations, reportez-vous à la section Annotations dans le chapitre 5 dans <i>Guide de l'utilisateur de IBM SPSS Modeler 15</i> .

Nugget Modèle d'ensemble de règles

Un nugget de modèle Ensemble de règles représente les règles servant à la prédiction d'un champ de sortie particulier découvert par des noeuds de modélisation de règle d'association (Apriori) ou par l'un des noeuds de génération d'arbre (C&RT, CHAID, QUEST, ou C5.0). Dans le cas des règles d'association, l'ensemble de règles doit être généré à partir d'un nugget de règle non affiné. Pour les arbres, vous pouvez générer un ensemble de règles à partir du générateur d'arbres, d'un noeud de création de modèle C5.0 ou de tout nugget de modèle d'arbre. Contrairement aux noeuds Règle non affinée, les nuggets Ensemble de règles générés peuvent être utilisés dans des flux pour générer des prévisions.

Lorsque vous exécutez un flux contenant un nugget Ensemble de règles, deux nouveaux champs sont ajoutés au flux, contenant la variable indépendante et le degré de confiance de chaque enregistrement. Les noms des nouveaux champs sont constitués du nom du modèle auquel des préfixes sont ajoutés. Pour les ensembles de règles d'association, le préfixe \$A- est utilisé pour désigner le champ de prévision et le préfixe \$AC- est utilisé pour désigner le champ contenant le degré de confiance. Pour les ensembles de règles C5.0, le préfixe \$C- est utilisé pour désigner le champ de prévision et le préfixe \$CC- est utilisé pour désigner le champ contenant le degré de confiance. Pour les ensembles de règles Arbre C&RT, le préfixe \$R- est utilisé pour désigner le champ de prévision et le préfixe \$RC- est utilisé pour désigner le champ contenant le degré de confiance. Dans un flux comportant plusieurs nuggets Ensemble de règles connectés en série et effectuant une prévision sur les mêmes champs de sortie, un nombre séquentiel sera ajouté au préfixe du nom des nouveaux champs afin de pouvoir les distinguer. Le premier nugget Ensemble de règles d'association du flux utilise les noms d'origine, le deuxième noeud les noms commençant par \$A1- et \$AC1-, le troisième noeud les noms commençant par \$A2- et \$AC2-, et ainsi de suite.

Comment les règles sont-elles appliquées ? Les ensembles de règles générés à partir de règles d'association sont différents des autres nuggets de modèle dans la mesure où plusieurs prévisions peuvent être générées pour un enregistrement et que ces prévisions peuvent être différentes. Il existe deux méthodes vous permettant de générer des prévisions à partir d'ensembles de règles.

Remarque : les ensembles de règles générés à partir d'arbres décision renvoient le même résultat, quelle que soit la méthode utilisée ; en effet, les règles calculées à partir d'un arbre décision s'excluent mutuellement.

- **Vote.** Cette méthode tente de combiner les prévisions de toutes les règles qui s'appliquent à l'enregistrement. Pour chaque enregistrement, toutes les règles sont examinées et chaque règle qui s'applique à l'enregistrement est utilisée pour générer une prévision et un degré de confiance associé. La somme des degrés de confiance de chaque valeur de sortie est calculée et la valeur pour laquelle cette somme est la plus élevée sera retenue comme prévision finale. Le degré de confiance de la prévision finale correspond à la somme des degrés de confiance de cette valeur divisée par le nombre de règles qui se sont déclenchées pour cet enregistrement.
- **Premier résultat.** Cette méthode teste les règles les unes après les autres et la première règle qui s'applique à l'enregistrement est celle utilisée pour générer la prévision.

Vous pouvez sélectionner la méthode à utiliser dans la boîte de dialogue des options de flux. [Pour plus d'informations, reportez-vous à la section Définition d'options générales pour les flux dans le chapitre 5 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Génération de noeuds. Le menu Générer vous permet de créer des noeuds sur la base d'un ensemble de règles.

- **Noeud Filtre.** Crée un nouveau noeud Filtrer pour filtrer les champs qui ne sont pas utilisés par les règles de l'ensemble de règles.
- **Noeud Sélectionner.** Crée un noeud Sélectionner pour sélectionner les enregistrements auxquels la règle sélectionnée s'applique. Le noeud généré sélectionne les enregistrements pour lesquels tous les antécédents de la règle se sont avérés vrais (True). Vous devez d'abord sélectionner une règle avant d'utiliser cette option.
- **Noeud Tracer règle.** Crée un super noeud pour créer un champ indiquant quelle règle a été utilisée pour calculer la prévision de chaque enregistrement. Lorsqu'un ensemble de règles est évalué à l'aide de la méthode Premier résultat, le champ créé contient simplement un

symbole indiquant quelle est la première règle qui se déclenchera. En revanche, lorsqu'un ensemble de règles est évalué à l'aide de la méthode Vote, le champ créé contient une chaîne plus complexe indiquant l'entrée du mécanisme de vote.

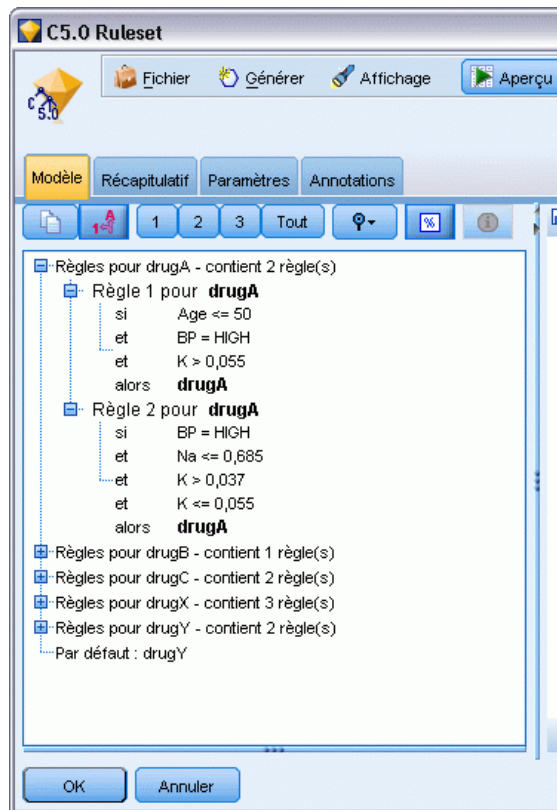
- **Arbre décision unique (espace de travail)/Arbre décision unique (palette GM).** Crée un nugget d'ensemble de règles à partir de la règle actuellement sélectionnée. Cette option est disponible uniquement pour les modèles C5.0 améliorés. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés sur p. 193.](#)
- **Modèle vers palette.** Renvoie le modèle à la palette de modèles. Ceci est utile dans des situations où un collègue peut vous avoir envoyé un flux contenant le modèle mais pas le modèle même.

Remarque : les onglets Paramètres et Récapitulatif du nugget Ensemble de règles sont identiques à ceux des modèles d'arbre décision.

Onglet Modèle d'ensemble de règles

L'onglet Modèle des nugget Ensemble de règles affiche la liste des règles extraites des données par l'algorithme.

Figure 6-44
Nugget de modèle Ensemble de règles, onglet Modèle



Les règles sont ventilées en fonction de la conséquence (catégorie prédite) et sont représentées dans le format suivant :

```
si antecedent_1
and antecedent_2
...
and antecedent_n
then predicted value
```

où *consequent* et *antecedent_1* en passant par *antecedent_n* sont toutes des conditions. La règle est pour “les enregistrements où *antecedent_1* en passant par *antecedent_n* sont tous vrais, *consequent* est également susceptible d’être vrai.” Si vous cliquez sur le bouton Afficher instances/confiance dans la barre d’outils, chaque règle affiche également des informations sur le nombre d’enregistrements auxquels la règle s’applique, c’est-à-dire pour lesquels les antécédents sont vrais (True) (**Instances**) et la proportion de ces enregistrements pour lesquels la totalité de la règle est vraie (True) (**Confiance**).

Notez que le degré de confiance est calculé différemment pour les ensembles de règles C5.0. C5.0 utilise la formule suivante pour calculer le degré de confiance d’une règle :

$$(1 + \text{number of records where rule is correct}) / (2 + \text{number of records for which the rule's antecedents are true})$$

Ce calcul du degré de confiance s’ajuste au processus de généralisation de règles à partir d’un arbre décision (C5.0 effectue également cette opération lorsqu’il crée un ensemble de règles).

Importation de projets à partir d’AnswerTree 3.0

IBM® SPSS® Modeler peut importer des projets enregistrés dans AnswerTree 3.0 ou 3.1 via la boîte de dialogue Fichier > Ouvrir standard, comme suit :

- ▶ Dans les menus SPSS Modeler, sélectionnez :
Fichier > Ouvrir un flux
- ▶ Dans la liste déroulante Types de fichier, sélectionnez Fichiers de projet AT (*.atp ; *.ats).

Chaque projet importé est converti en flux SPSS Modeler avec les noeuds suivants :

- Un noeud source définissant la source de données utilisée (par exemple, un fichier de données ou une source de base de données IBM® SPSS® Statistics).
 - Pour chaque arbre du projet (il peut en contenir plusieurs), le noeud Typer créé définit les propriétés de chaque champ (variable), y compris le type, le rôle (champ d’entrée ou variable indépendante, et champ de sortie ou prédit), les valeurs manquantes, etc.
 - Pour chaque arbre du projet, un noeud Partitionner partitionnant les données d’un échantillon d’apprentissage ou de test est créé, ainsi qu’un noeud générateur d’arbre définissant les paramètres de génération d’arbre (un noeud Arbre C&R, QUEST ou CHAID).
- ▶ Pour afficher les arbres créés, exécutez le flux.

Commentaires

- Les arbres décision générés dans SPSS Modeler ne peuvent pas être exportés vers AnswerTree. L'importation des projets depuis AnswerTree vers SPSS Modeler est unilatérale.
- Les profits définis dans AnswerTree ne sont pas conservés si le projet concerné est importé dans SPSS Modeler.

Modèles de réseau Bayésien

Noeud Réseau Bayésien

Le noeud **Réseau Bayésien** permet de 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. Le noeud est axé sur le Tree Augmented Naïve Bayes (TAN) et sur les réseaux Couverture de Markov qui servent principalement à la classification.

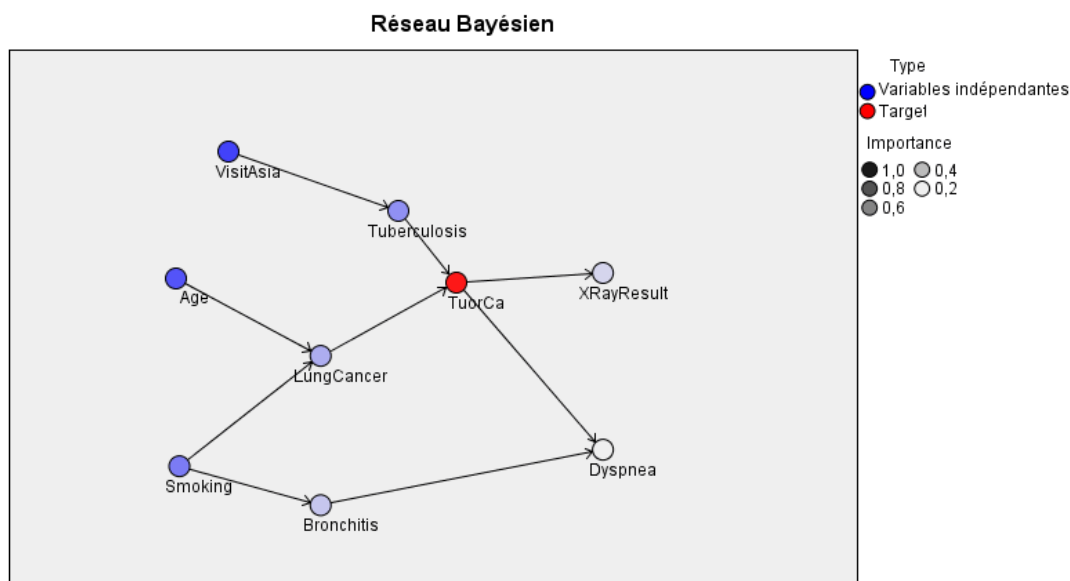
Les réseaux Bayésiens permettent d’effectuer des prédictions dans de nombreuses situations, par exemple :

- La sélection des opportunités de prêt avec des risques par défaut peu élevés.
- L’évaluation du moment auquel le matériel aura besoin d’être vérifié, remplacé ou nécessitera de nouvelles pièces, en fonction des données du capteur et des enregistrements existants.
- La résolution des problèmes client à l’aide des outils de dépannage en ligne.
- Le diagnostic et le dépannage des réseaux de téléphones cellulaires en temps réel.
- L’évaluation des risques et des bénéfices potentiels de projets de recherche et de développement afin de concentrer les ressources sur les meilleures opportunités.

Un réseau Bayésien est un modèle graphique qui présente des variables (souvent appelées **noeuds**) dans un ensemble de données et les indépendances probabilistes ou conditionnelles qui les relie. La relation causale entre les noeuds peut être représentée par un réseau Bayésien ; cependant, les liens dans le réseau (également appelés **arcs**) ne représentent pas nécessairement un lien direct de cause à effet. Par exemple, un réseau Bayésien peut être utilisé pour calculer la probabilité d’une maladie spécifique chez un patient, en fonction de la présence ou de l’absence de certains symptômes et d’autres données importantes, si les indépendances probabilistes entre les symptômes et la maladie apparaissent comme valides dans le graphique. Ces réseaux restent très solides lorsque des informations sont manquantes et font les meilleures prédictions possibles à l’aide des informations disponibles.

Lauritzen et Spiegelhalter (1988) ont créé un exemple général et basique d’un réseau Bayésien. Il est généralement appelé le modèle “asiatique” et est une version simplifiée d’un réseau qui peut être utilisé pour diagnostiquer les nouveaux patients d’un médecin ; la direction des liens correspond grossièrement à la causalité. Chaque noeud représente une facette qui peut être liée à l’état du patient ; par exemple, “Fumeur” indique que le patient est un fumeur confirmé et “AVisitél’Asie” indique qu’il est récemment allé en Asie. La relation de probabilité est indiquée par les liens entre les noeuds ; par exemple, fumer augmente les risques que le patient développe à la fois des bronchites et un cancer des poumons, alors que l’âge semble être uniquement associé à la possibilité de développer un cancer des poumons. De la même façon, les anomalies sur une radiographie des poumons peuvent provenir de la tuberculose ou d’un cancer des poumons, alors que les risques qu’un patient ait des problèmes respiratoires (dyspnée) augmentent s’il souffre également soit d’une bronchite ou d’un cancer des poumons.

Figure 7-1
Exemple du réseau Asiatique de Lauritzen et Spiegelhalter



Il existe plusieurs raisons pour lesquelles vous pourriez décider d'utiliser un réseau Bayésien :

- Cela vous permet de connaître les relations causales. Ainsi, vous pouvez comprendre le contexte de la problématique et prédire les conséquences de toute intervention.
- Le réseau offre une approche efficace permettant d'éviter le surajustement des données.
- Une visualisation claire des relations impliquées est facilement accessible.

Conditions requises. Les champs cible doivent être de type catégoriel et ils peuvent comporter un niveau de mesure *Nominal*, *Ordinal*, ou *Booléen*. Les données peuvent être des champs de tout type. Les champs de données continues (intervalle numérique) seront automatiquement discrétisés ; mais si la proportion est asymétrique, il est possible d'obtenir de meilleurs résultats en discrétisant manuellement ces champs à l'aide du noeud Discrétiser avant d'utiliser le noeud Réseau Bayésien. Par exemple, utilisez la Création d'intervalles optimale où le champ de superviseur est le même que le champ Cible du noeud Réseau Bayésien. [Pour plus d'informations, reportez-vous à la section Noeud Discrétiser dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

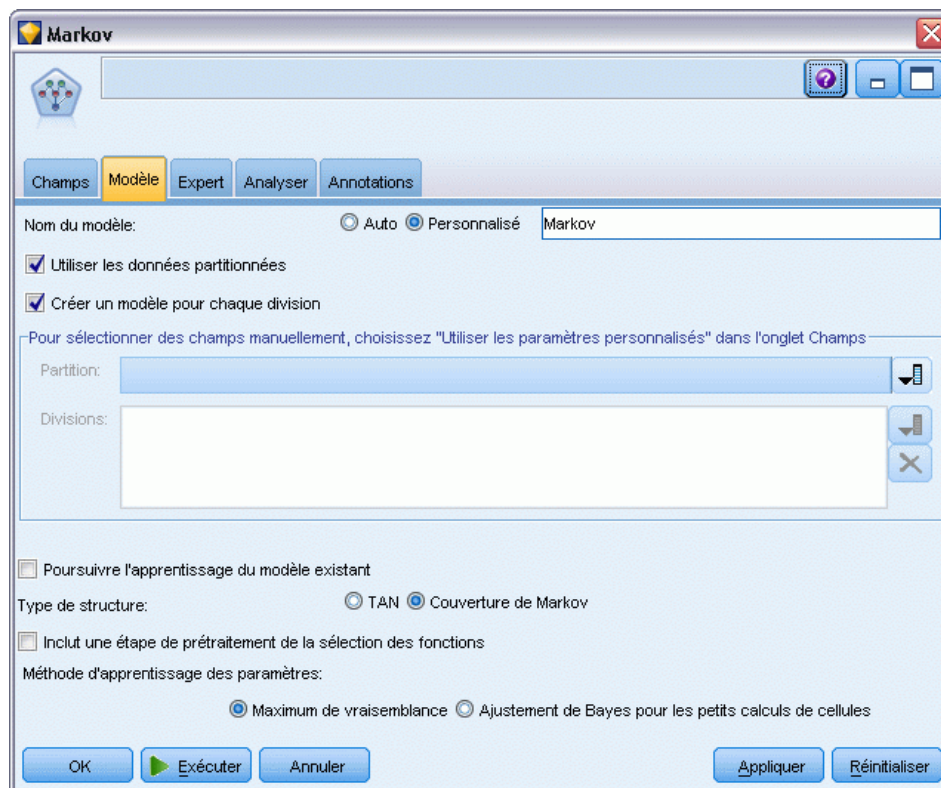
Exemple : L'analyste d'une banque souhaite pouvoir prédire quels clients, ou clients potentiels, risquent de ne pas être capables de rembourser leur prêt. Vous pouvez utiliser un modèle de réseau Bayésien pour identifier les caractéristiques des clients les plus susceptibles de ne pas pouvoir rembourser leur prêt, et créer plusieurs types de modèles différents afin de choisir le meilleur modèle pour ce type de prédiction. [Pour plus d'informations, reportez-vous à la section Préviation des défauts de paiement \(Réseau Bayésien\) dans le chapitre 17 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Exemple. Un opérateur de télécommunications souhaite réduire le nombre de clients qui le quittent (appelé " attrition ") et mettre à jour le modèle chaque mois en utilisant les données du mois précédent. Vous pouvez utiliser un modèle de réseau Bayésien pour identifier les

caractéristiques des clients les plus susceptibles de partir et continuer à renseigner chaque mois le modèle avec de nouvelles données. [Pour plus d'informations, reportez-vous à la section Recyclage d'un modèle chaque mois \(Réseau Bayésien\) dans le chapitre 18 dans Guide d'applications de IBM SPSS Modeler 15.](#)

Options du modèle de noeud Réseau Bayésien

Figure 7-2
Noeud Réseau Bayésien : 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.](#)

Créer un modèle pour chaque.division. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

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 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.)

Découpages. Pour des modèles découpés, sélectionnez le ou les champs de découpage. Cela revient à définir le rôle du champ sur la valeur *Division* dans un noeud Typer. Vous ne pouvez désigner que des champs ayant un niveau de mesure Booléen, Nominal, Ordinal ou Continu comme champs de découpage. Les champs sélectionnés en tant que le champ de découpage ne peuvent pas être utilisés comme champs de cible, d'entrée, de partition, de fréquence ou de pondération. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Poursuivre le modèle d'apprentissage existant. Si vous sélectionnez cette option, les résultats affichés dans l'onglet Modèle du nugget de modèle sont régénérés et mis à jour à chaque exécution du modèle. Par exemple, c'est ce que vous feriez si vous aviez ajouté une source de données nouvelle ou mise à jour à un modèle existant.

Remarque : Seul le réseau existant est mis à jour ; il ne peut ni ajouter ni supprimer des noeuds ou des connexions. Chaque fois que vous recyclez le modèle, le réseau conserve la même forme et seules les probabilités conditionnelles et l'importance des valeurs prédites changent. Si vos nouvelles données sont dans l'ensemble semblables à vos anciennes données, cela importe peu car vous vous attendez à ce que les mêmes choses aient la même importance ; cependant, si vous souhaitez vérifier ou mettre à jour *ce qui* est important (en opposition à " combien " c'est important), vous devrez créer un nouveau modèle, c'est-à-dire créer un nouveau réseau

Type de structure. Sélectionnez la structure à utiliser lors de la création d'un réseau Bayésien :

- **TAN.** Le modèle Tree Augmented Naïve Bayes (TAN) crée un modèle de réseau Bayésien simple qui est un modèle Naïve Bayes standard amélioré. Il permet que chaque valeur prédite dépende d'une autre valeur prédite en plus de la variable cible, et d'augmenter ainsi la précision de la classification.
- **Couverture de Markov.** Cela permet de sélectionner l'ensemble de noeuds de l'ensemble de données qui contient les parents de la variable cible, ses enfants et les parents de ses enfants. Une couverture de Markov identifie principalement toutes les variables du réseau qui sont nécessaires à la prédiction de la variable cible. Cette méthode de construction de réseau est considérée comme étant la plus précise ; cependant, avec de grands ensembles de données, la durée du traitement peut être plus importante en raison du grand nombre de variables concerné. Pour réduire la durée du traitement, vous pouvez utiliser les options Sélection de fonctions de l'onglet Expert pour sélectionner les variables qui sont nettement associées à la variable cible.

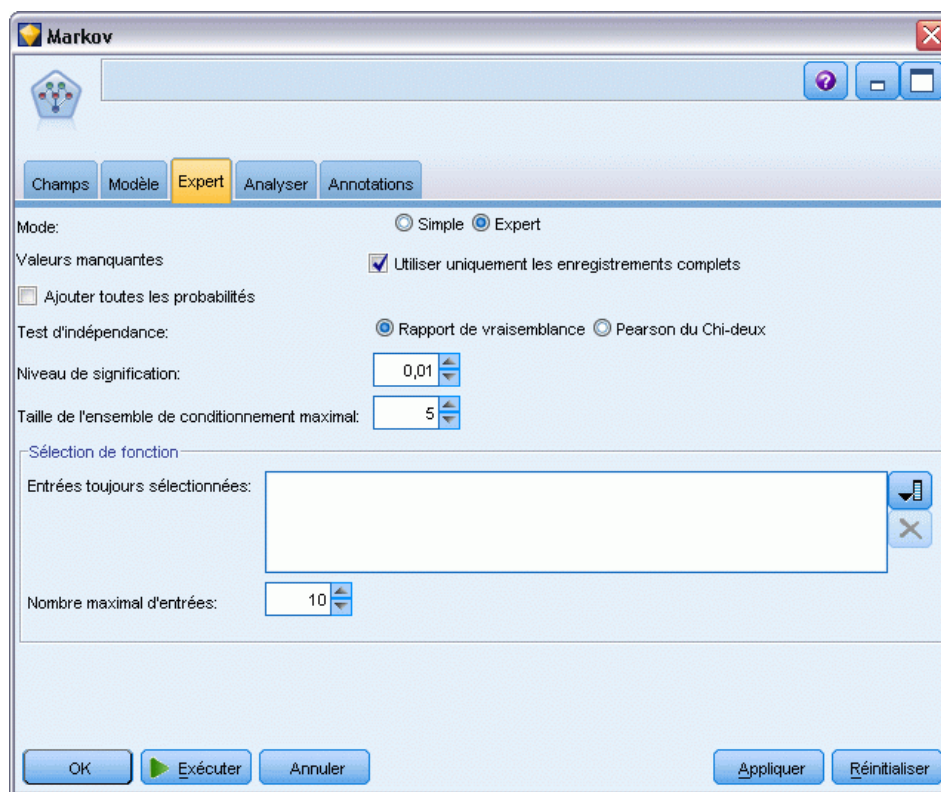
Inclure une étape de prétraitement de la sélection des fonctions. Sélectionner cette case permet d'utiliser les options de Sélection de fonctions de l'onglet Expert.

Méthode d'apprentissage des paramètres. Les paramètres de réseau Bayésien font référence aux probabilités conditionnelles de chaque noeud en fonction des valeurs de ses parents. Il existe deux sélections possibles permettant de contrôler la tâche d'estimation des tableaux de probabilités conditionnelles entre les noeuds où les valeurs des parents sont connues :

- **Maximum de vraisemblance.** Sélectionnez cette case lorsque vous utilisez un grand ensemble de données. Il s'agit de la sélection par défaut.
- **Ajustement de Bayes pour les petits calculs de cellules.** Pour les petits ensembles de données, il existe un risque de surajustement du modèle, ainsi qu'une possibilité d'un grand nombre de zéros. Sélectionnez cette option pour éviter ces problèmes en utilisant le lissage et réduire ainsi l'effet de tout zéro et de toute estimation non fiable.

Noeud Réseau Bayésien - Options expert

Figure 7-3
Noeud Réseau Bayésien : Onglet Expert



Les options expert du noeud vous permettent d'affiner le processus de création de modèle. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Valeurs manquantes : Par défaut, IBM® SPSS® Modeler utilise exclusivement les enregistrements dont les valeurs sont valides pour tous les champs du modèle. (cette stratégie est parfois appelée **suppression des observations incomplètes** des valeurs manquantes). S'il manque un grand nombre de données, cette stratégie risque d'éliminer trop d'enregistrements ; par conséquent, la quantité de données disponibles peut être insuffisante pour générer un modèle pertinent. Dans ces cas, vous pouvez désélectionner l'option N'utiliser que les enregistrements complets, SPSS Modeler puis essayez d'utiliser autant d'informations que possible pour estimer le modèle, y compris les enregistrements où certains champs ont des valeurs manquantes. (Cette stratégie est parfois appelée **suppression des valeurs manquantes appariées**.) Toutefois dans certains cas, l'utilisation d'enregistrements incomplets peut entraîner des difficultés pour le calcul de l'estimation du modèle.

Ajouter toutes les probabilités. Spécifie l'ajout ou le non-ajout des probabilités de chacune des catégories du champ de sortie à chacun des enregistrements traités par le noeud. Lorsqu'elle ne l'est pas, seule la probabilité de la catégorie prédite est ajoutée.

Test d'indépendance. Un test d'indépendance évalue si les observations par paire sur deux variables sont indépendantes. Sélectionnez le type de test à utiliser, les options disponibles sont :

- **Rapport de vraisemblance.** Teste l'indépendance des valeurs prédites cible en calculant un rapport entre la probabilité maximum d'un résultat selon deux hypothèses différentes.
- **Pearson du Chi-deux.** Teste l'indépendance des valeurs prédites cible en utilisant une hypothèse nulle que les fréquences relatives d'occurrence des événements observés suivent une distribution de fréquence spécifiée.

Les modèles de réseau Bayésien effectuent des tests conditionnels d'indépendance où des variables supplémentaires sont utilisées au-delà des paires testées. De plus, ces modèles n'explorent pas uniquement les relations entre la cible et les valeurs prédites, mais également les relations entre les valeurs prédites elle-mêmes.

Remarque : Les options du test d'indépendance sont uniquement disponibles si vous sélectionnez Inclure une étape de prétraitement de la sélection des fonctions ou un Type de structure de couverture de Markov sur l'onglet Modèle.

Niveau de signification : Utilisé en conjonction avec les paramètres du test d'indépendance, il permet de définir une valeur-seuil à utiliser pour effectuer les tests. Plus la valeur est basse, moins il reste de liens dans le réseau ; le niveau par défaut est 0,01.

Remarque : Cette option est uniquement disponible si vous sélectionnez Inclure une étape de prétraitement de la sélection des fonctions ou un Type de structure de couverture de Markov sur l'onglet Modèle.

Taille de l'ensemble de conditionnement maximal. L'algorithme permettant de créer une structure de couverture de Markov utilise des ensembles de conditionnement de taille augmentée pour effectuer des tests d'indépendance et supprimer les liens inutiles du réseau. Parce que les tests traitant un grand nombre de variables de conditionnement nécessitent plus de temps et de mémoire pour le traitement, vous pouvez limiter le nombre de variables à inclure. Cela est particulièrement utile lors du traitement de données ayant de fortes dépendances sur plusieurs variables. Cependant, il est important de noter que le réseau en résultant peut contenir des liens superflus.

Spécifiez le nombre maximal de variables de conditionnement à utiliser pour les tests d'indépendance. Le paramètre par défaut est de 5.

Remarque : Cette option est uniquement disponible si vous sélectionnez Inclure une étape de prétraitement de la sélection des fonctions ou un Type de structure de couverture de Markov sur l'onglet Modèle.

Sélection de fonctions. Ces options permettent de réduire le nombre d'entrées utilisées lors du traitement du modèle afin de réduire la durée du processus de création de modèle. Cela est particulièrement utile lors de la création d'une structure de couverture de Markov, en raison du nombre important d'entrées potentielles ; vous pouvez ainsi sélectionner des entrées qui sont reliées à la variable cible de manière significative.

Remarque : Les options de sélection de fonctions sont uniquement disponibles si vous sélectionnez Inclure une étape de prétraitement de la sélection des fonctions sur l'onglet Modèle.

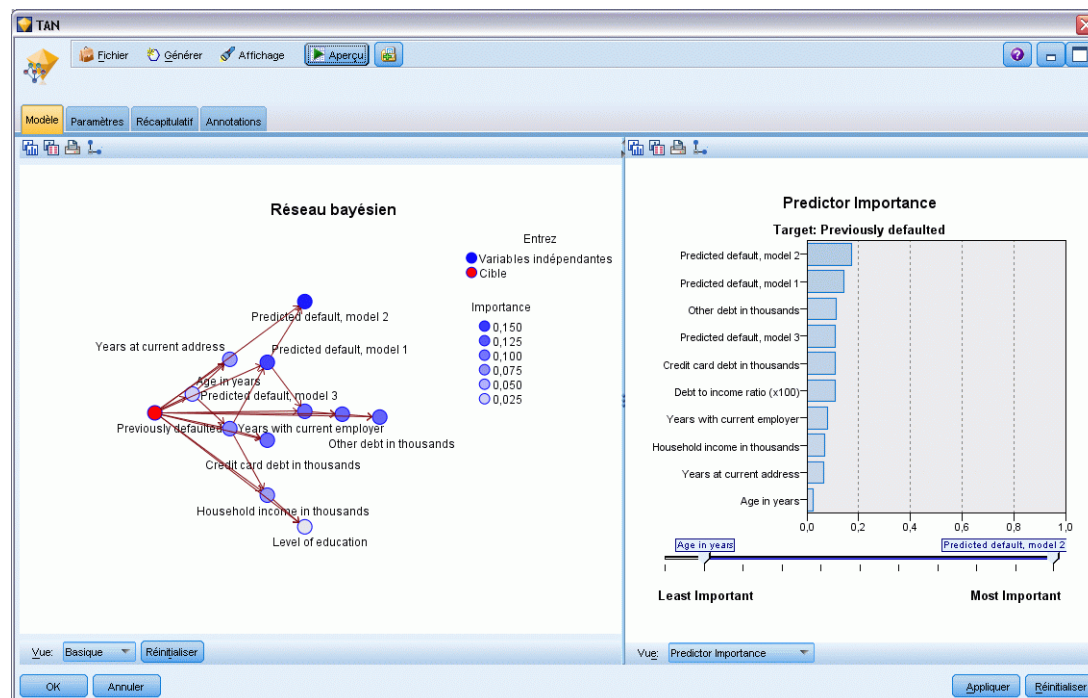
- **Données toujours sélectionnées** A l'aide du sélecteur de champs (bouton à droite du champ de texte), sélectionnez les champs de l'ensemble de données qui doivent toujours être utilisés lors de la création d'un modèle de réseau Bayésien. Veuillez noter que le champ cible est toujours sélectionné.
- **Nombre maximal d'entrées.** Spécifiez le nombre total d'entrées de l'ensemble de données à utiliser lors de la création d'un modèle de réseau Bayésien. Le nombre maximum que vous pouvez saisir est le nombre total d'entrées de l'ensemble de données.

Remarque : Si le nombre de champs sélectionnés dans Données toujours sélectionnées dépasse la valeur du Nombre maximal d'entrées, un message d'erreur apparaît.

Nuggets de modèle de Réseau Bayésien

Figure 7-4

Détails du modèle de Réseau Bayésien et de l'importance des variables prédites associées



Remarque : si vous avez sélectionné Poursuivre les paramètres d'apprentissage existants dans l'onglet Modèle du noeud de modélisation, les informations affichées dans cet onglet Modèle du nugget de modèle sont mises à jour chaque fois que vous régénérez le modèle.

L'onglet Modèle du nugget de modèle est divisé en deux panneaux :

Panneau de gauche

Basique. Cette vue contient un graphique de noeuds en réseau qui affiche la relation entre la cible et ses valeurs prédites les plus importantes, ainsi que la relation entre les valeurs prédites. L'importance de chaque valeur prédite est représentée par la densité de sa couleur : une couleur foncée représente une valeur prédite importante et vice versa.

Les valeurs d'intervalle des noeuds qui représentent un intervalle sont affichées dans une info-bulle contextuelle lorsque le curseur de la souris passe sur le noeud.

Vous pouvez utiliser les outils de graphiques de IBM® SPSS® Modeler pour interagir, modifier et enregistrer le graphique. Par exemple, pour une utilisation dans d'autres applications comme MS Word.

Astuce : Si le réseau contient de nombreux noeuds, vous pouvez cliquer sur l'un d'eux et le faire glisser pour que le graphique soit plus lisible.

Distribution : Cette vue affiche les probabilités conditionnelles de chaque noeud dans le réseau sous la forme d'un mini-graphique. Passez le curseur de la souris sur un graphique pour que ses valeurs apparaissent dans une info-bulle contextuelle.

Panneau de droite

Importance des valeurs prédites. Affiche un graphique qui indique l'importance relative de chaque variable prédite à évaluer le modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

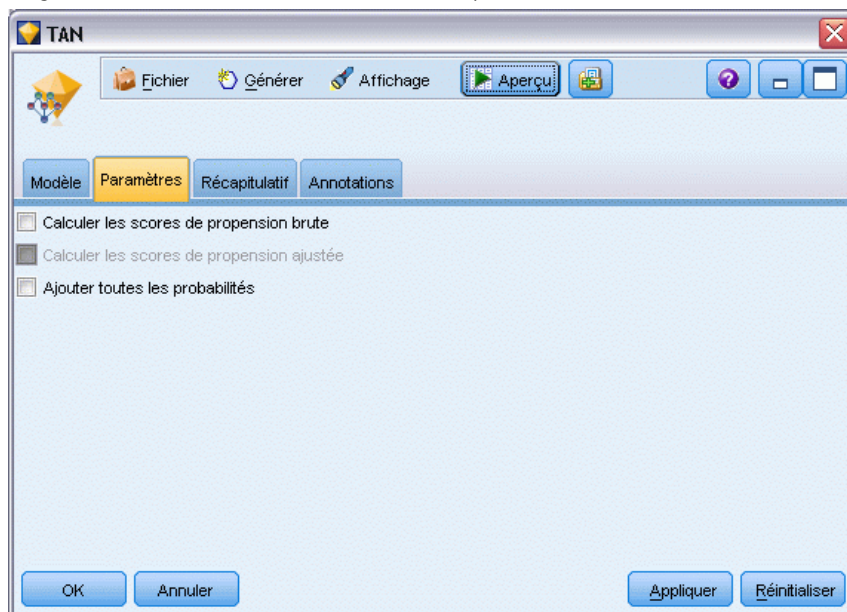
Probabilités conditionnelles. Lorsque vous sélectionnez un noeud ou un mini-graphique Proportion dans le panneau de gauche, le tableau des probabilités conditionnelles associées apparaît dans le panneau de droite. Ce tableau contient la valeur de probabilité conditionnelle de chaque valeur de noeud et chaque combinaison de valeurs dans ses noeuds parent. De plus, elle contient le nombre d'enregistrements observés pour chaque valeur d'enregistrement et pour chaque combinaison de valeurs dans les noeuds parent.

Paramètres de modèle de Réseau Bayésien

L'onglet Paramètres d'un nugget de modèle de Réseau Bayésien affiche les options de modification du modèle créé. Par exemple, vous pouvez utiliser le noeud de Réseau Bayésien pour créer plusieurs modèles différents en utilisant les mêmes données et paramètres, puis utiliser cet onglet dans chaque modèle pour modifier légèrement les paramètres et voir en quoi cela influe sur les résultats.

Remarque : Cet onglet n'est disponible qu'une fois le nugget de modèle ajouté à un flux.

Figure 7-5
Onglet Paramètres d'un modèle Réseau Bayésien



Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

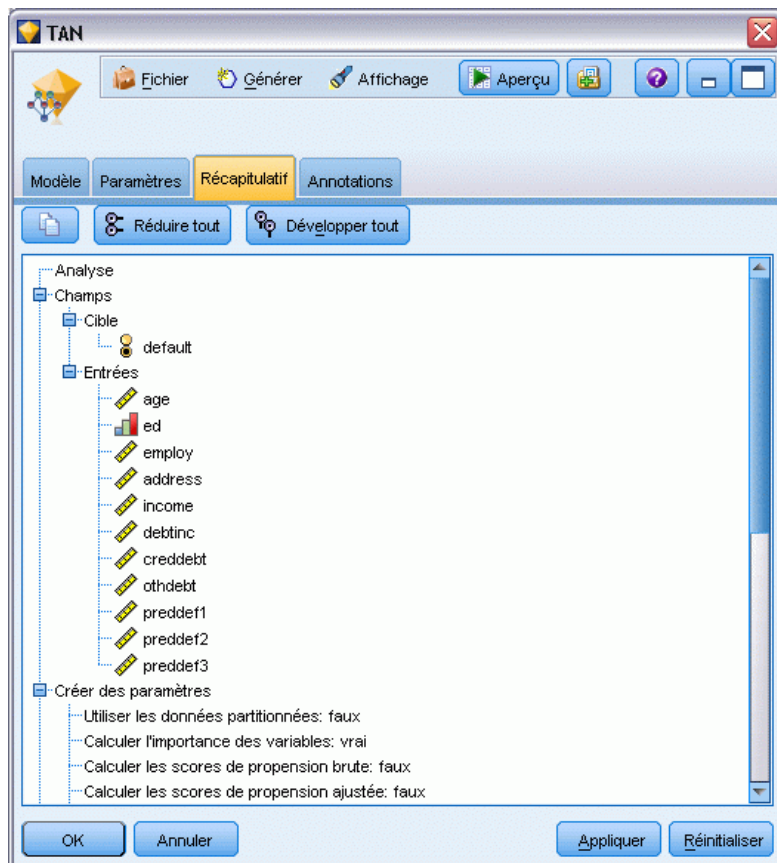
Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Ajouter toutes les probabilités. Spécifie l'ajout ou le non-ajout des probabilités de chacune des catégories du champ de sortie à chacun des enregistrements traités par le noeud. Lorsqu'elle ne l'est pas, seule la probabilité de la catégorie prédite est ajoutée.

Le paramètre par défaut de cette case à cocher est déterminé par la case correspondante sur l'onglet Expert du noeud de modélisation. [Pour plus d'informations, reportez-vous à la section Noeud Réseau Bayésien - Options expert sur p. 207.](#)

Récapitulatif du modèle de Réseau Bayésien

Figure 7-6
Onglet Récapitulatif d'un modèle de Réseau Bayésien



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.

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.

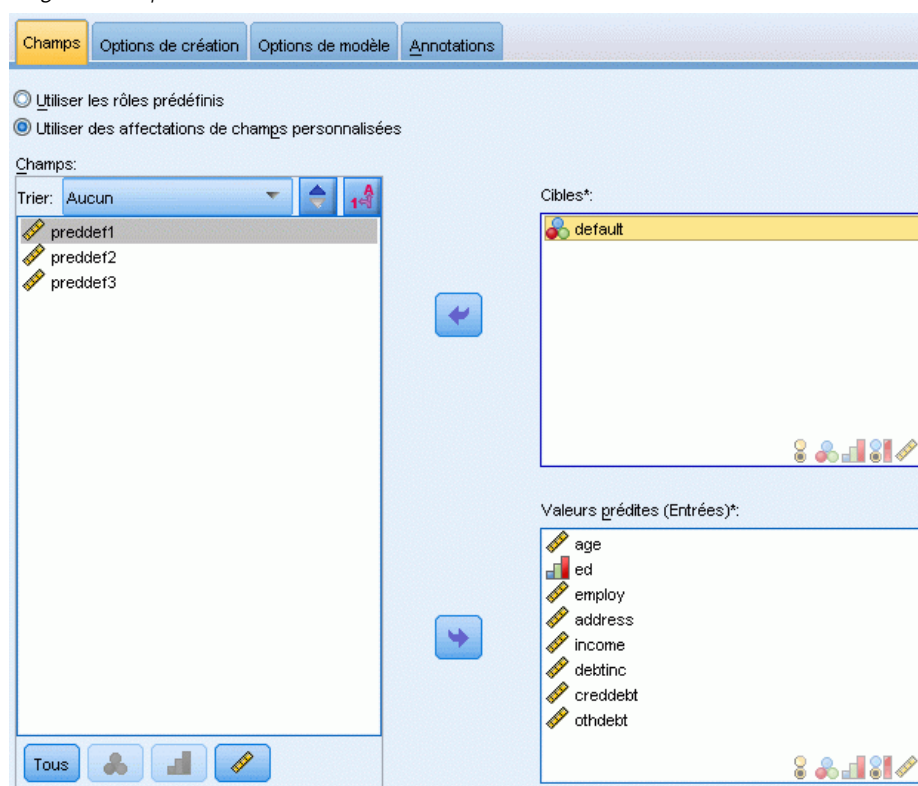
Récapitulatif de l'apprentissage. Indique le type du modèle, le flux utilisé pour le créer, l'utilisateur qui l'a créé, ainsi que sa date et sa durée de création.

Réseaux de neurones

Un **réseau de neurones** peut approcher une large gamme de modèles prédictifs avec un minimum de demandes sur la structure et les hypothèses du modèle. La forme des relations est déterminée pendant le processus d'apprentissage. Si une relation linéaire entre la cible et les valeurs prédites est appropriée, les résultats du réseau de neurones doivent approcher étroitement ceux d'un modèle linéaire traditionnel. Si une relation non linéaire est plus adéquate, le réseau de neurones approche automatiquement la structure de modèle « appropriée ».

Le compromis concernant cette flexibilité réside dans le fait que le réseau de neurones n'est pas facilement interprétable. Si vous essayez d'expliquer un processus sous-jacent à l'origine des relations entre la cible et les valeurs prédites, il serait préférable d'utiliser un modèle statistique plus traditionnel. Cependant, si l'interprétabilité du modèle n'est pas importante, vous pouvez obtenir de bonnes prédictions à l'aide d'un réseau de neurones.

Figure 8-1
Onglet Champs



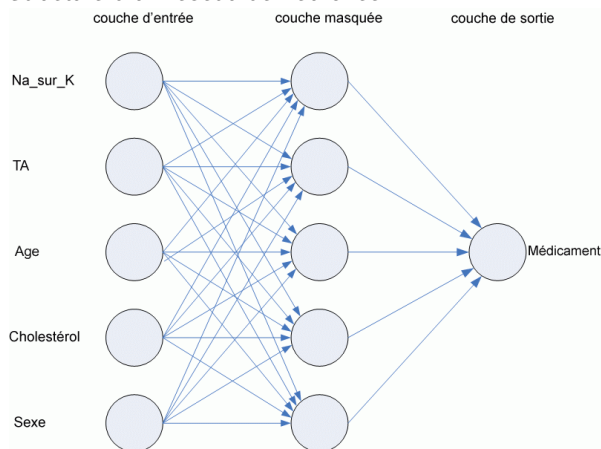
Exigences concernant les champs. Il doit y avoir au moins une cible et une entrée. Les champs définis sur Les deux ou Aucun sont ignorés. Il n'existe pas de restrictions concernant le niveau de mesure des cibles ou des valeurs prédites (entrées). [Pour plus d'informations, reportez-vous à la section Options de champs des noeuds de modélisation dans le chapitre 3 sur p. 39.](#)

Le modèle des réseaux de neurones

Les réseaux de neurones sont des modèles simples représentant le fonctionnement du système nerveux. Les unités de base sont les **neurones**. Ils sont généralement organisés en **couches**, comme l'illustre la figure ci-dessous.

Figure 8-2

Structure d'un réseau de neurones



Le **noeud R. neurones** est un modèle simplifié de la manière dont le cerveau humain traite les informations. Le fonctionnement de ce modèle repose sur la simulation d'un grand nombre d'unités de traitement interconnectées, qui sont en quelque sorte des versions abstraites de nos neurones.

Ces unités de traitement sont organisées en couches. Il existe généralement trois types de couche dans un réseau de neurones : une **couche d'entrée** dans laquelle les unités représentent les champs d'entrée, une ou plusieurs **couches cachées**, ainsi qu'une **couche de sortie** dans laquelle des unités représentent les champs cibles. Les unités sont reliées entre elles par des connexions de puissance (ou de **pondération**) différentes. Les données d'entrée sont présentées dans la première couche et les valeurs sont transmises entre les neurones d'une couche à l'autre. Le résultat final est obtenu à partir de la couche de sortie.

Lors de son apprentissage, le réseau procède à l'examen de tous les enregistrements afin de générer des prévisions et modifie les pondérations lorsque l'une de ses prévisions s'avère incorrecte. Ce processus se répète plusieurs fois et le réseau continue d'améliorer ses prévisions jusqu'à ce que l'un des critères d'arrêt soit atteint.

Au début, tous les coefficients de pondération sont aléatoires et les réponses en provenance du réseau risquent de ne pas avoir de sens. Le réseau apprend à travers l'**apprentissage**. Les exemples dont le résultat est connu sont présentés à plusieurs reprises au réseau et les réponses qu'il donne sont comparées aux résultats connus. Les informations de cette comparaison sont réacheminées via le réseau, modifiant progressivement les coefficients de pondération. Au fur et

à mesure de l'apprentissage, les résultats connus répliqués par le réseau sont à chaque fois plus précis. Lorsque l'apprentissage est terminé, le réseau peut être appliqué à d'autres observations pour lesquelles le résultat est inconnu.

Utilisation des réseaux de neurones avec les flux hérités

La version 14 de IBM® SPSS® Modeler a présenté un nouveau noeud Réseau de neurones qui prend en charge les techniques de boosting et de bagging pour les ensembles de données très volumineux. Les flux existants contenant l'ancien noeud créent et évaluent encore des modèles dans cette version. Cependant, cette prise en charge sera supprimée dans une version ultérieure, aussi, nous vous recommandons d'utiliser la nouvelle version dès maintenant.

A partir de la version 13, les champs contenant des valeurs inconnues (c'est-à-dire des valeurs absentes des données d'apprentissage) ne sont plus automatiquement traités comme comportant des valeurs manquantes et sont évalués avec la valeur \$null\$. Par conséquent, si vous souhaitez évaluer les champs avec des valeurs inconnues comme des champs non nuls à l'aide d'un modèle de Réseau de neurones plus ancien (antérieur à la version 13) de la version 13 ou d'une version ultérieure, vous devez signaler les valeurs inconnues comme des valeurs manquantes (par exemple, à l'aide du noeud Typex.)

Objectifs

Figure 8-3
Paramètres des Objectifs

Champs Options de création Options de modèle Annotations

Sélectionnez un élément :

Objectifs
Bases
Règles d'arrêt
Ensembles
Options avancées

Que souhaitez-vous faire ?

Créer un nouveau modèle

Poursuivre le modèle d'apprentissage existant

Quel est votre objectif principal ?

Créer un modèle standard

Améliorer la précision du modèle (amélioration)

Améliorer la stabilité du modèle (bagging)

Créer un modèle pour des ensembles de données volumineux (nécessite un serveur)

Description

Crée un modèle unique standard expliquant les relations entre les champs. Les modèles standard sont plus faciles à interpréter et peuvent être plus rapidement évalués que les ensembles de données volumineux, améliorés ou agrégés. Un modèle standard est toujours utilisé pour les cibles multiples.

Que souhaitez-vous faire ?

- **Créer un nouveau modèle.** Créer un modèle entièrement nouveau. Il s'agit de l'opération ordinaire du noeud.
- **Poursuivre l'apprentissage d'un modèle existant.** L'apprentissage se poursuit sur le dernier modèle généré par le noeud. Cela permet de mettre à jour ou d'actualiser un modèle existant sans avoir à accéder aux données d'origine et peut permettre d'obtenir des performances nettement plus rapides car seuls les enregistrements nouveaux ou mis à jour sont acheminés dans le flux. Les détails relatifs au modèle précédent sont stockés avec le noeud de modélisation, ce qui permet d'utiliser cette option même si le nugget de modèle précédent n'est plus disponible dans le flux ou dans la palette de modèles.

Remarque : Lorsque cette option est activée, toutes les autres commandes des onglets Champ et Options de création sont désactivés.

Quel est votre objectif principal ?

- **Créer un modèle standard.** La méthode crée un modèle unique afin de prédire la cible à l'aide des variables indépendantes. En général, les modèles standard sont plus faciles à interpréter et peuvent être plus rapides à évaluer que des ensembles de données améliorés, agrégés ou volumineux.
- **Améliorer la précision du modèle (amélioration).** La méthode crée un modèle d'ensemble à l'aide d'une amélioration, qui génère une séquence de modèles pour obtenir des prédictions plus précises. La création et l'évaluation des ensembles peut prendre davantage de temps qu'un modèle standard.

Le boosting produit une succession de « modèles de composant », chacun généré à partir de l'ensemble de données entier. Avant la création de chaque modèle de composant successif, les enregistrements sont pondérés en fonction des résidus du modèle de composant précédent. Les observations possédant des résidus élevés reçoivent des pondérations d'analyse relativement supérieure de manière à ce que le modèle de composant suivant donne une meilleure prédiction de ces enregistrements. Ensemble ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.

- **Améliorer la stabilité du modèle (agrégation).** La méthode crée un modèle d'ensemble à l'aide du bagging (agrégation de bootstrap), qui génère plusieurs modèles pour obtenir des prédictions plus fiables. La création et l'évaluation des ensembles peut prendre davantage de temps qu'un modèle standard.

L'agrégation de bootstrap (bagging) produit des doubles de l'ensemble de données d'apprentissage en réalisant un échantillonnage avec remplacement à partir de l'ensemble de données d'origine. Ceci permet de créer des échantillons de bootstrap de taille égale à l'ensemble de données d'origine. Ensuite, un « modèle de composant » est créé à partir de chaque double. Ensemble ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.

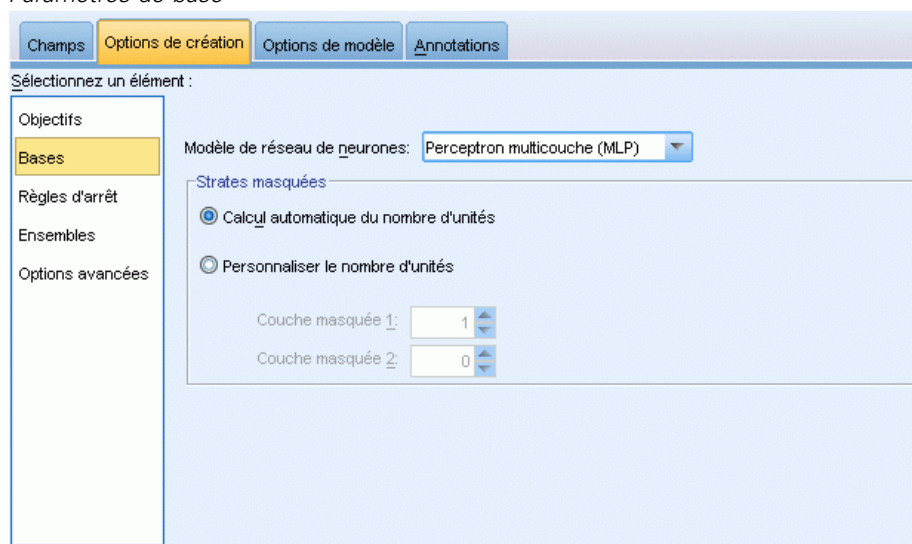
- **Créer un modèle pour de très grands ensembles de données (requiert IBM® SPSS® Modeler Server).** La méthode crée un modèle d'ensemble en divisant l'ensemble de données en des blocs de données distincts. Sélectionnez cette option si votre ensemble de données est trop volumineux pour créer l'un des modèles ci-dessus, ou pour créer un modèle incrémentiel.

Cette option peut prendre moins de temps à créer, mais davantage à évaluer qu'un modèle standard. Cette option nécessite une connexion SPSS Modeler Server.

Lorsqu'il y a plusieurs cibles, cette méthode ne crée qu'un modèle standard, quel que soit l'objectif sélectionné.

Bases

Figure 8-4
Paramètres de base



Modèle de réseau de neurones. Le type du modèle détermine la façon dont le réseau connecte les valeurs prédites aux cibles par le biais des couches masquées. Le **perceptron multicouche (MLP)** autorise des relations plus complexes au prix d'un accroissement de la durée d'apprentissage et de scoring. La **fonction radiale de base (RBF)** peut avoir une durée d'apprentissage et de scoring inférieure au prix d'une puissance de prédiction réduite comparée au MLP.

Couches masquées. La ou les couches masquées d'un réseau de neurones contiennent des unités inobservables. La valeur de chaque unité masquée est une fonction des valeurs prédites ; la forme exacte de la fonction dépend partiellement du type de réseau. Un perceptron multicouche peut avoir une ou deux couches masquées ; un réseau de fonction radiale de base peut avoir une couche masquée.

- **Calculer automatiquement le nombre d'unités.** Cette option crée un réseau comportant une couche masquée et calcule le « meilleur » nombre d'unités dans celle-ci.
- **Personnaliser le nombre d'unités.** Cette option vous permet de spécifier le nombre d'unités de chaque couche masquée. La première couche masquée doit comporter au moins une unité. Spécifier 0 unité pour la seconde couche masquée crée un perceptron multicouche comportant une seule couche masquée.

Remarque : Vous devez sélectionner des valeurs de sorte que le nombre de noeuds n'excède pas le nombre de valeurs prédites continues plus le nombre total de modalités parmi toutes les valeurs prédites catégorielles (booléennes, nominales et ordinales).

Règles d'arrêt

Figure 8-5
Paramètres des règles d'arrêt

Champs Options de création Options de modèle Annotations

Sélectionnez un élément :

- Objectifs
- Bases
- Règles d'arrêt**
- Ensembles
- Options avancées

i Les règles d'arrêt sont disponibles pour le modèle Perceptron multicouche et s'appliquent au modèle standard d'apprentissage (ou au modèles de composant si les ensembles sont effectifs). D'autres règles d'arrêt s'appliquent.

Utiliser le temps maximal d'apprentissage (par modèle de composant).

Minutes:

Personnaliser le nombre maximum de cycles d'apprentissage

Nombre maximal de cycles:

Utiliser le niveau de précision minimal

Précision (%):

Il s'agit des règles qui déterminent le moment d'arrêt de l'apprentissage de réseaux d'un perceptron multicouche ; ces paramètres sont ignorés lorsque l'algorithme de fonction radiale de base est utilisé. L'apprentissage exécute au moins un cycle (transmission de données) et peut alors être arrêté en fonction des critères suivants.

Utiliser la durée d'apprentissage maximale (par modèle de composant). Indiquez s'il faut spécifier un nombre maximum de minutes pour l'algorithme à exécuter. Spécifiez un nombre supérieur à 0. Lorsqu'un modèle d'ensemble est créé, il s'agit de la durée d'apprentissage autorisée pour chaque modèle de composant de l'ensemble. Veuillez noter que l'apprentissage peut se poursuivre un peu au-delà de la limite imposée pour permettre au cycle d'apprentissage en cours d'être achevé.

Personnaliser le nombre de cycles d'apprentissage maximum. Le nombre maximum de cycles d'apprentissage autorisé. Si le nombre maximum de cycles est dépassé, l'apprentissage s'arrête. Spécifiez un entier supérieur à 0.

Utiliser la précision minimum. Avec cette option, l'apprentissage se poursuit jusqu'à ce que la précision spécifiée soit atteinte. Il est possible que la précision spécifiée ne soit jamais atteinte. Cependant, vous pouvez à tout moment interrompre l'apprentissage et enregistrer le réseau en conservant la meilleure précision atteinte jusque-là.

L'algorithme d'apprentissage s'arrête aussi si l'erreur de l'ensemble de prévention de surajustement ne décroît pas après chaque cycle, si la modification relative à l'erreur d'apprentissage est petite ou si le taux d'erreur d'apprentissage actuel est petit comparé à l'erreur initiale.

Ensembles

Figure 8-6
Paramètres des ensembles

Sélectionnez un élément :

- Objectif
- Bases
- Règles d'arrêt
- Coûts et Probabilités a priori
- Ensembles**
- Options avancées

i Ces paramètres déterminent le comportement de création des ensembles lorsque l'amélioration, l'agrégation ou de très volumineux ensembles de données sont requis dans les objectifs. Les options ne s'appliquant pas sont ignorées.

Règles de combinaison

Règle de combinaison par défaut pour les cibles catégorielles:

Règle de combinaison par défaut pour les cibles continues:

Amélioration et agrégation

Nombre des modèles de composant pour l'amélioration et l'agrégation:

Ces paramètres déterminent le comportement d'assemblage qui se produit lors du boosting, du bagging ou lorsque que des ensembles volumineux de données sont requis dans les objectifs. Les options qui ne s'appliquent pas à l'objectif sélectionné sont ignorées.

Bagging and Very Large Datasets. When scoring an ensemble, this is the rule used to combine the predicted values from the base models to compute the ensemble score value.

- **Règles de combinaison par défaut pour les cibles catégorielles.** Des valeurs prédites d'ensemble pour ces cibles catégorielles peuvent être combinées à l'aide du vote, de la probabilité la plus élevée ou de la probabilité de moyenne la plus élevée. Le **vote** sélectionne la catégorie qui a la plus forte probabilité le plus souvent sur les mêmes modèles de base. La **probabilité la plus élevée** sélectionne la catégorie qui atteint la probabilité la plus élevée sur tous les modèles de base. La **probabilité de moyenne la plus élevée** sélectionne la catégorie dont la valeur est la plus élevée lorsqu'est effectué la moyenne des probabilités de catégorie sur les modèles de base.
- **Règles de combinaison par défaut pour les cibles continues.** Des valeurs prédites d'ensemble pour des cibles continues peuvent être combinées à l'aide de la moyenne ou de la médiane des valeurs prédites à partir des modèles de base.

Veillez noter que lorsque l'objectif consiste à améliorer la précision du modèle, les sélections de règles de combinaisons sont ignorées. Le boosting utilise toujours un vote majoritaire pondéré pour évaluer des cibles catégorielles et une médiane pondérée pour évaluer des cibles continues.

Boosting et bagging. Spécifiez le nombre de modèles de base à créer lorsque l'objectif est d'améliorer la précision ou la stabilité du modèle ; pour le bagging, il s'agit du nombre d'échantillons de bootstrap. Il doit s'agir d'un entier positif.

Avancés

Figure 8-7
Paramètres avancés

Sélectionnez un élément :

- Objectifs
- Bases
- Règles d'arrêt
- Ensembles
- Options avancées

i La modélisation de réseau de neurones sépare les enregistrements de manière interne en un ensemble de création de modèle et en un ensemble de prévention du surajustement. Spécifiez un pourcentage d'enregistrements pour l'ensemble de prévention du surajustement.

Ensemble de prévention de surajustement (%):

Dupliquer les résultats

Graine aléatoire:

Valeurs manquantes dans les valeurs prédites :

- Supprimer les observations incomplètes
- Attribuer les valeurs manquantes

Les paramètres avancés offrent un contrôle sur des options qui ne sont pas correctement adaptées à d'autres groupes de paramètres.

Ensemble de prévention de surajustement. La méthode du réseau de neurones sépare des enregistrements de manière interne en un ensemble de création de modèle et un ensemble de prévention de surajustement, qui est un ensemble d'enregistrements de données servant à tracer les erreurs pendant l'apprentissage afin d'éviter que la méthode n'effectue une modélisation d'une variation aléatoire dans les données. Spécifier un pourcentage d'enregistrements. La valeur par défaut est 30.

Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus. Par défaut, les analyses sont dupliquées avec une valeur de départ de 229176228.

Valeurs manquantes dans les valeurs prédites. Ceci spécifie la façon de traiter les valeurs manquantes. **Supprimer par liste** supprime de la création du modèle des enregistrements comportant des valeurs manquantes dans des valeurs prédites. **Inclure les valeurs manquantes** remplace les valeurs manquantes dans les valeurs prédites et utilise ces enregistrements dans l'analyse. Les champs continus attribuent la moyenne des valeurs minimum et maximum observées ; les champs catégoriels attribuent la modalité la plus fréquente. Veuillez noter que les enregistrements comportant des valeurs manquantes sur tous les autres champs spécifiés dans l'onglet Champ, sont toujours supprimés de la création du modèle.

Options du modèle

Figure 8-8
Onglet Options de modèle

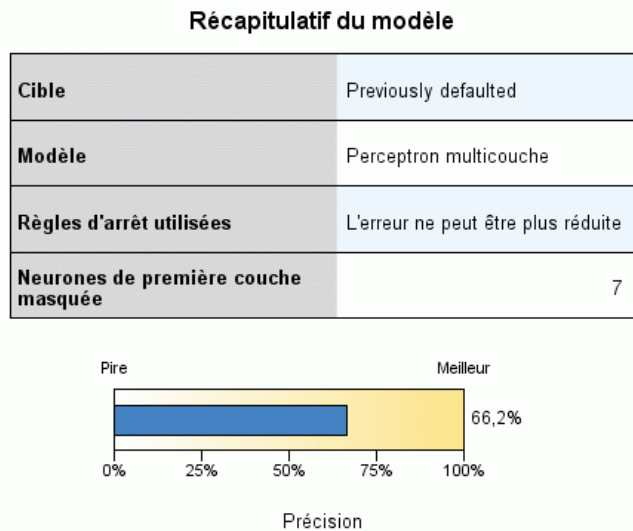
Nom de modèle. Vous pouvez générer automatiquement le nom du modèle en fonction des champs cible ou spécifier un nom personnalisé. Le nom généré automatiquement est le nom du champ cible. S'il existe plusieurs cibles, le nom du modèle correspond aux noms des champs reliés par des esperluettes. Par exemple, si *champ1* *champ2* *champ3* sont des cibles, alors le nom du modèle est : *champ1 & champ2 & champ3*.

Rendre disponible pour le scoring. Lorsque le modèle est évalué, les éléments sélectionnés dans ce groupe doivent être générés. La valeur prédite (pour toutes les cibles) et la confiance (pour les cibles catégorielles) sont toujours calculées lorsque le modèle est évalué. La confiance calculée peut être basée sur la probabilité de la valeur prédite (la probabilité prédite la plus élevée) ou sur la différence entre la probabilité prédite la plus élevée et la deuxième probabilité prédite la plus élevée.

- **Probabilité prédite pour les cibles catégorielles.** Ceci génère les probabilités prédites pour les cibles catégorielles. Un champ est créé pour chaque modalité.
- **Scores de propension pour des variables cibles booléennes.** Pour les modèles disposant d'une variable cible booléenne (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Le modèle produit des scores de propension brute. Si les partitions sont activées, le modèle produit également des scores de propension ajustée basés sur la partition de test. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Récapitulatif du modèle

Figure 8-9
Vue Récapitulatif du modèle réseaux de neurones



La vue récapitulative du modèle est un instantané, un récapitulatif accessible d'un coup d'œil du réseau de neurones prédictif ou de la précision du classement.

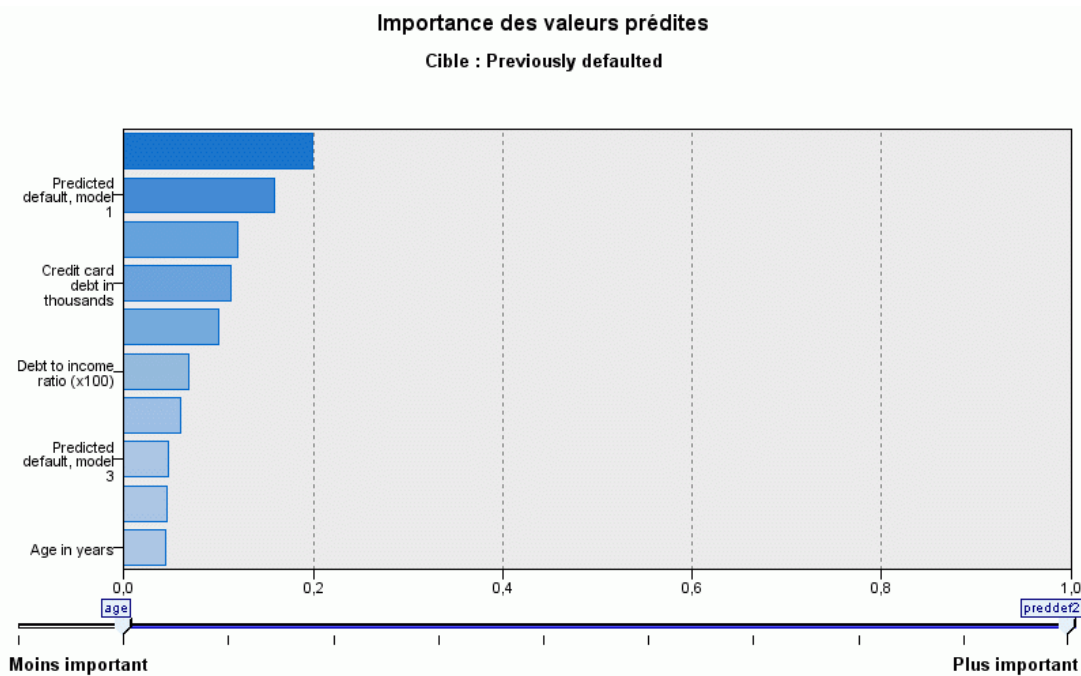
Récapitulatif du modèle. Le tableau identifie la cible, le type de réseau de neurones formé, la règle d'arrêt qui a interrompu l'apprentissage (affichée si un réseau de perceptron multicouche a été formé) et le nombre de neurones de chaque couche masquée du réseau.

Qualité du réseau de neurones. Le diagramme affiche la précision du modèle final, qui est présenté en plus grand, disposant d'un meilleur format. Pour les cibles catégorielles, il s'agit du pourcentage des enregistrements pour lesquels la valeur prédite correspond à la valeur observée. Pour une cible continue, il s'agit de 1 moins le taux d'erreur moyenne absolue de la prédiction (la moyenne des valeurs absolues des valeurs prédites moins les valeurs observées) divisé par la plage des valeurs prédites (la valeur maximale prédite moins la valeur minimale prédite).

Cibles multiples. S'il y a plusieurs cibles, chacune d'elle est affichée dans la ligne Cible du tableau. La précision affichée dans le graphique correspond à la moyenne des précisions des cibles individuelles.

Importance des valeurs prédites

Figure 8-10
Vue Importance des variables indépendantes

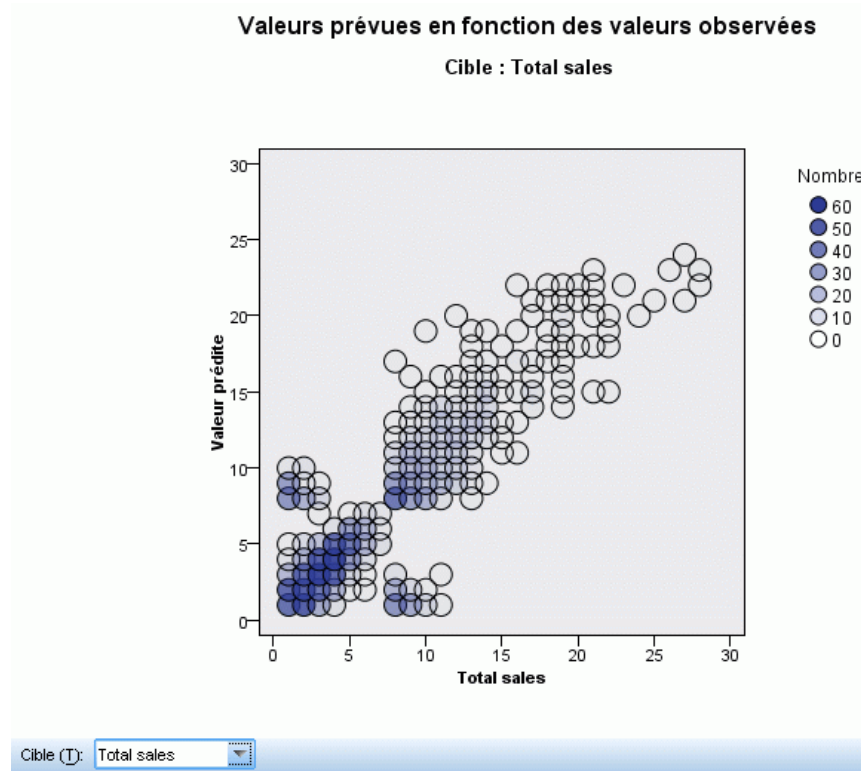


En général, vous préférerez concentrer vos efforts de modélisation sur les champs prédictifs les plus importants et abandonner ou ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des valeurs prédites affichée est 1,0. L'importance des des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Cibles multiples. S'il existe plusieurs cibles, chacune est affichée dans un graphique séparé et une liste déroulante Cible contrôle la cible à afficher.

Valeurs prédites en fonction des valeurs observées

Figure 8-11
Vue Valeurs prédites en fonction des valeurs observées

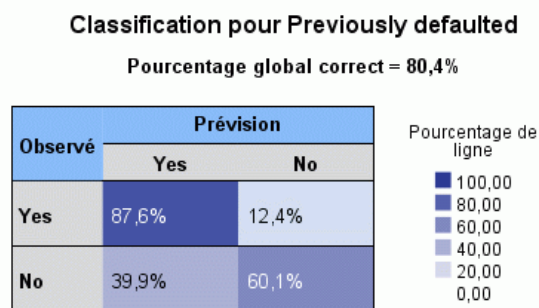


Pour des cibles continues, ceci affiche un diagramme de dispersion mis en intervalles des valeurs prédites sur l'axe vertical par les valeurs observées sur l'axe horizontal.

Cibles multiples. S'il existe plusieurs cibles continues, chacune est affichée dans un graphique séparé et une liste déroulante Cible contrôle la cible à afficher.

Classification

Figure 8-12
Vue Classification, style du pourcentage des lignes



Pour les cibles catégorielles, cette option affiche la classification croisée des valeurs observées par rapport aux valeurs prédites sur une carte des zones de chaleur, ainsi que le pourcentage global correct.

Styles de tableaux. Il existe différents styles d’affichage, accessibles depuis la liste déroulante Style .

- **Pourcentages de ligne.** Affiche les pourcentages de ligne (les effectifs de cellules exprimés sous forme de pourcentage du nombre total de lignes) dans les cellules. Il s’agit de la valeur par défaut.
- **Effectifs de cellules.** Affiche les effectifs de cellules dans les cellules. L’ombrage de la carte des zones de chaleur reste basé sur les pourcentages de ligne.
- **Zones de chaleur.** N’affiche aucune valeur dans les cellules, uniquement l’ombrage.
- **Compressé.** N’affiche aucun en-tête de ligne ou de colonne ni de valeur dans les cellules. Ceci peut être utile lorsque la cible possède de nombreuses catégories.

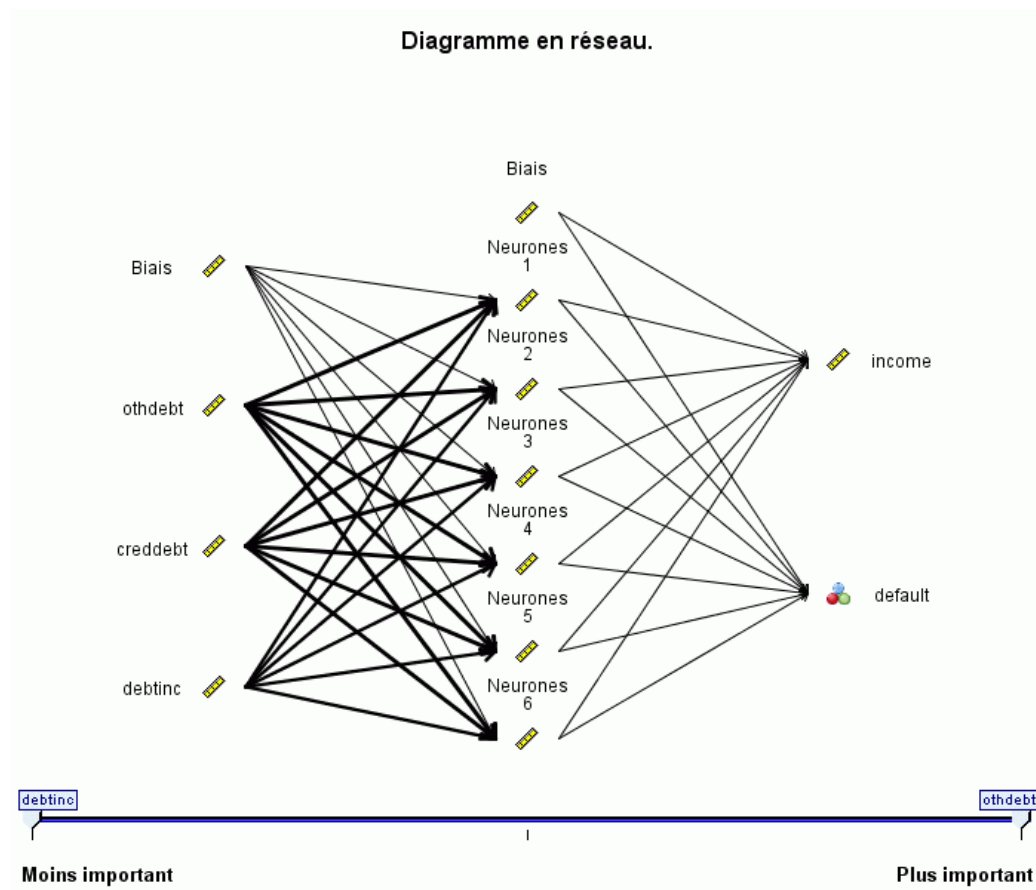
Manquantes. Si des enregistrements ont des valeurs manquantes sur la cible, elles sont affichées sur une ligne (Manquantes) en dessous des lignes valides. Les enregistrements avec des valeurs manquantes ne contribuent pas au pourcentage général correct.

Cibles multiples. S’il existe plusieurs cibles catégorielles, chacune est affichée dans un tableau distinct et une liste déroulante Cible contrôle la cible à afficher.

Grands tableaux. Si la cible affichée comporte plus de 100 catégories, aucun tableau n’est affiché.

Réseau

Figure 8-13
Vue Réseau, entrées à gauche, style d'effets



Affiche une représentation graphique du réseau de neurones.

Styles de diagramme. Il existe deux styles différents d'affichage accessibles depuis la liste déroulante Style .

- **Effets.** Ceci affiche chaque valeur prédite et chaque cible sous la forme d'un noeud dans le diagramme, que l'échelle de mesure soit continue ou catégorielle. Il s'agit de la valeur par défaut.
- **Coefficients.** Affiche des noeuds à indicateurs multiples pour des valeurs prédites et des cibles catégorielles. Les lignes de connexion dans le diagramme de style coefficient sont colorées en fonction de la valeur estimée de la pondération synaptique.

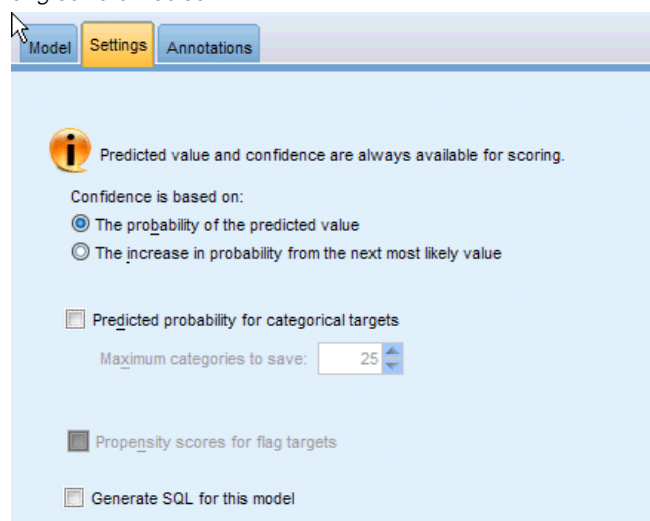
Orientation du diagramme. Par défaut, le diagramme du réseau est organisé avec les entrées à gauche et les cibles à droite. En utilisant les commandes de la barre d'outils, vous pouvez modifier l'orientation afin que les entrées se trouvent en haut et les cibles en bas ou que les entrées se trouvent en bas et les cibles en haut.

Importance des variables indépendantes. Les lignes de connexion du diagramme sont pondérées en fonction de l'importance des valeurs prédites, une largeur de ligne plus importante correspondant à une plus grande importance. Il existe un curseur de l'importance des valeurs prédites dans la barre d'outils qui contrôle celles qui sont affichées dans le diagramme du réseau. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les variables indépendantes les plus importantes.

Cibles multiples. S'il existe plusieurs cibles, toutes sont affichées dans le graphique.

Paramètres

Figure 8-14
onglet Paramètres



Lorsque le modèle est évalué, les éléments sélectionnés dans cet onglet doivent être générés. La valeur prédite (pour toutes les cibles) et la confiance (pour les cibles catégorielles) sont toujours calculées lorsque le modèle est évalué. La confiance calculée peut être basée sur la probabilité de la valeur prédite (la probabilité prédite la plus élevée) ou sur la différence entre la probabilité prédite la plus élevée et la deuxième probabilité prédite la plus élevée.

- **Probabilité prédite pour les cibles catégorielles.** Ceci génère les probabilités prédites pour les cibles catégorielles. Un champ est créé pour chaque modalité.
- **Scores de propension pour des variables cibles booléennes.** Pour les modèles disposant d'une variable cible booléenne (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Le modèle produit des scores de propension brute. Si les partitions sont activées, le modèle produit également des scores de propension ajustée basés sur la partition de test. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Générer SQL pour ce modèle. Lorsque vous utilisez des données provenant d'une base de données, le code SQL peut être renvoyé à la base de données pour exécution, ce qui assure des performances supérieures pour de nombreuses opérations. [Pour plus d'informations, reportez-vous à la section](#)

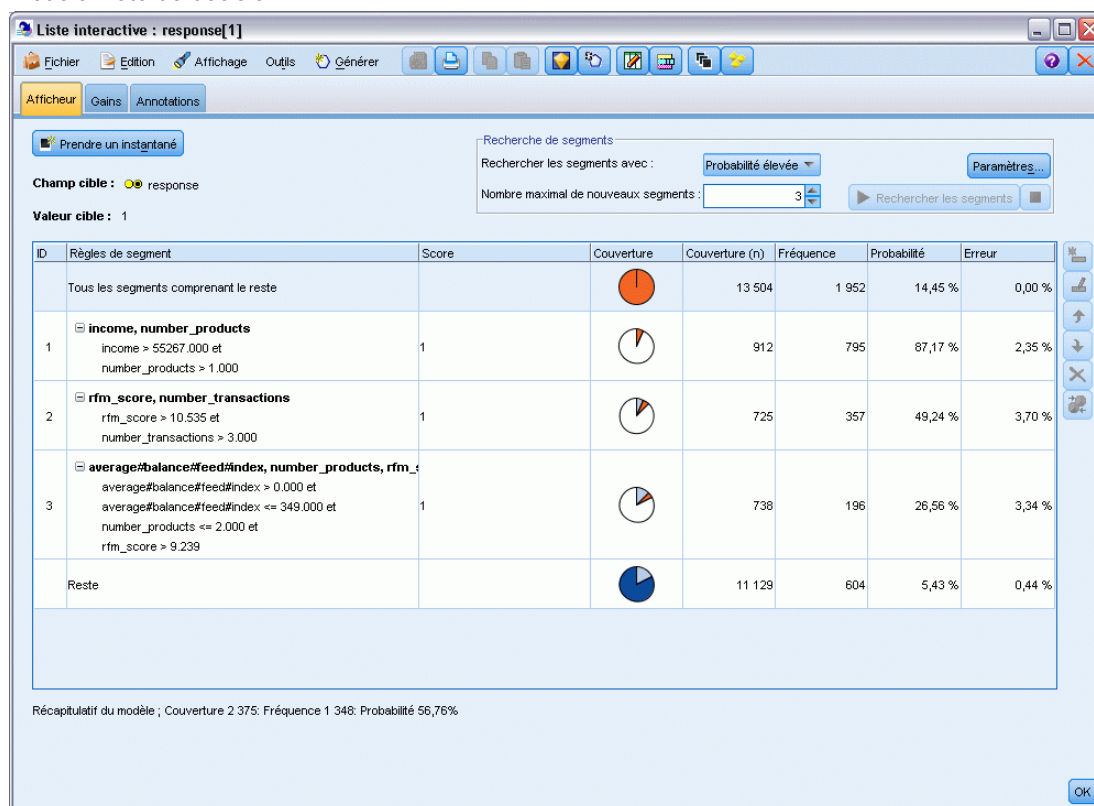
Optimisation SQL dans le chapitre 6 dans *Guide d'administration et des performances de IBM SPSS Modeler Server 15*.

Evaluer en convertissant en SQL natif. Si cette option est sélectionnée, le SQL est généré pour évaluer le modèle de manière native dans l'application.

Liste de décision

Les modèles de la Decision List identifient les sous-groupes ou les **segments**, qui présentent une probabilité plus élevée ou plus faible d'un résultat binaire (oui ou non) par rapport à l'échantillon global. Par exemple, vous pouvez rechercher les clients qui présentent la plus faible probabilité d'attrition ou qui sont les plus susceptibles de répondre favorablement à une offre ou une campagne spécifique. L'Decision List Viewer offre un contrôle complet sur le modèle : vous avez ainsi la possibilité d'éditer des segments, d'ajouter vos propres règles de marché, de spécifier les modalités de détermination du score de chaque segment et de personnaliser le modèle de plusieurs façons afin d'optimiser la proportion des correspondances dans tous les segments. Ainsi, il est particulièrement adapté à la génération de fichiers d'adresses ou à l'identification d'enregistrements à cibler pour une campagne spécifique. Vous pouvez également utiliser plusieurs **tâches d'exploration** pour combiner diverses approches de modélisation—par exemple en identifiant les segments performants et moins performants au sein du même modèle, et en les incluant ou en les excluant à l'étape de scoring selon les besoins.

Figure 9-1
Modèle Liste de décision



Segments, règles et conditions

Un modèle se compose d'une liste de segments, chacun étant défini par une règle qui sélectionne les enregistrements avec correspondance. Une règle donnée peut être constituée de plusieurs conditions, par exemple :

```
RFM_SCORE > 10 et  
MONTHS_CURRENT <= 9
```

Les règles sont appliquées dans l'ordre indiqué, la première règle correspondante déterminant le résultat pour un enregistrement donné. Prises individuellement, les règles ou les conditions peuvent se recouper, mais l'ordre des règles résout toute ambiguïté. Si aucune règle ne correspond, l'enregistrement est attribué à la règle restante.

Maîtrise totale du scoring

L'Decision List Viewer permet de visualiser, de modifier et de réorganiser les segments, mais également de choisir les segments à inclure ou à exclure à des fins de scoring. Vous pouvez par exemple choisir d'exclure un groupe de clients des offres à venir, en inclure d'autres et visualiser immédiatement la façon dont cette modification affecte votre taux de correspondance global. Les modèles de la Decision List renvoient un score *Yes* (oui) pour les segments inclus et *\$null\$* pour tous les autres, y compris le reste. Ce contrôle direct sur le scoring fait des modèles de la Decision List une solution idéale pour la génération de fichiers d'adresses ; ils sont largement utilisés dans la gestion de la relation client, notamment dans les applications marketing ou de centre d'appels.

Figure 9-2
Modèle Liste de décision

Prendre un instantané

Champ cible : ● réponse
Valeur cible : 1

Recherche de segments
Rechercher les segments avec : Probabilité élevée
Nombre maximal de nouveaux segments : 3
Rechercher les segments

ID	Règles de segment	Score	Couverture	Couverture (n)	Fréquence	Probabilité	Erreur
	Tous les segments comprenant le reste			13 504	1 952	14,45 %	0,00 %
1	income > 55267.000 et number_products > 1.000	1				87,17 %	2,35 %
2	rfm_score > 10.535 et number_transactions > 3.000	1				49,24 %	3,70 %
3	average#balance#feed#index, number_products, rfm_s average#balance#feed#index > 0.000 et average#balance#feed#index <= 349.000 et number_products <= 2.000 et rfm_score > 9.239	1				26,56 %	3,34 %
	Reste					5,43 %	0,44 %

Récapitulatif du modèle ; Couverture 2 375; Fréquence 1 348; Probabilité 56,76%

Tâches d'exploration, mesures et sélections

Le processus de modélisation est géré par les **tâches d'exploration**. Chaque tâche d'exploration lance efficacement une nouvelle exécution de modélisation et renvoie un nouvel ensemble d'autres modèles parmi lesquels vous pouvez effectuer votre choix. La tâche par défaut est fondée sur vos spécifications initiales dans le noeud Decision List, mais vous pouvez définir autant de tâches personnalisées que vous le souhaitez. Vous pouvez appliquer des tâches à plusieurs reprises. Par exemple, vous pouvez lancer une recherche de forte probabilité sur la totalité de l'ensemble d'apprentissage, puis une recherche de faible probabilité sur le reste pour éliminer les segments les moins performants.

Figure 9-3
Création d'une tâche exploratoire

Créer/Editer une tâche d'exploration : {0}

Chargement des paramètres : Nouveau... X

Cible

Champ cible : ● response Valeur cible : 1

Paramètres simples

Rechercher les segments avec : Probabilité élevée

Nombre maximal de nouveaux segments : 3

Taille minimale de segment

Comme pourcentage du segment précédent (%) : 5,0

Comme valeur absolue (N) : 50

Nombre maximal d'alternatives : 3

Nombre maximal d'attributs par segment : 5

Autoriser la réutilisation d'attribut dans un segment

Intervalle de confiance pour les nouvelles conditions (%) : 85,0

Paramètres Expert

Méthode de création d'intervalles : Décompte égal Nombre d'intervalles : 10

Largeur de recherche de modèle : 5 Largeur de recherche de règle : 5

Facteur de fusion d'intervalles : 2.00

Autoriser les valeurs manquantes dans les conditions : Vrai Supprimer les résultats intermédiaires : Vrai

Editer...

Données

Sélection de création : Toutes les données

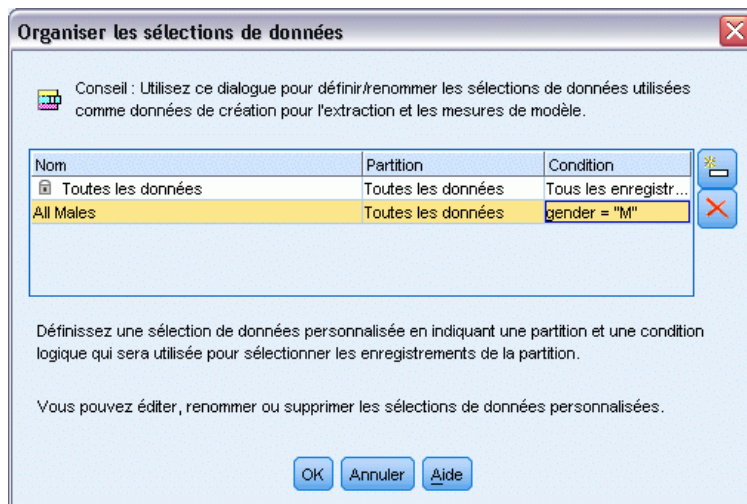
Champs disponibles : Tous les champs Personnalisé Editer...

OK Annuler Aide

Sélections de données

Vous pouvez définir les sélections de données et les mesures de modèle personnalisées pour la génération et l'évaluation de modèles. Vous pouvez par exemple indiquer une sélection de données dans une tâche exploratoire pour adapter le modèle à une région spécifique et créer une mesure personnalisée pour évaluer les performances du modèle sur l'ensemble du pays. Contrairement aux tâches d'exploration, les mesures ne modifient pas le modèle sous-jacent mais offrent une autre vision qui permet d'évaluer ses performances.

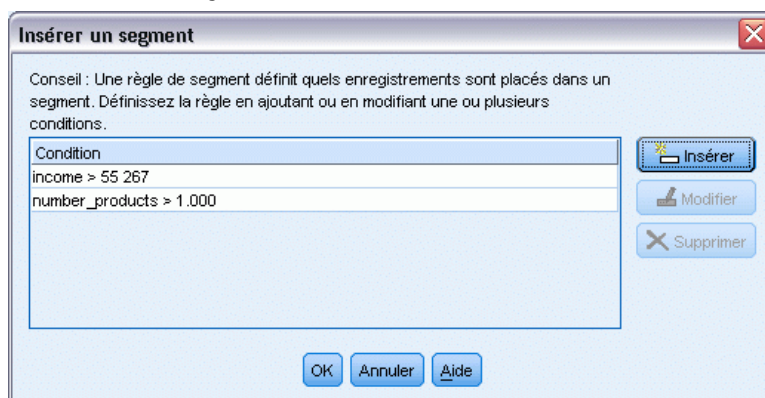
Figure 9-4
Création d'une sélection de données



Ajout de vos connaissances du marché

En affinant ou en étendant les segments identifiés par l'algorithme, l'Decision List Viewer vous permet d'incorporer votre connaissance du marché dans le modèle. Vous pouvez éditer les segments générés par le modèle ou ajouter des segments en fonction de règles que vous indiquez. Vous pouvez ensuite appliquer les modifications et obtenir un aperçu des résultats.

Figure 9-5
Définition d'une règle



Pour vous faire une idée plus précise, un lien dynamique à Excel vous permet d'exporter des données vers Excel, où vous pouvez les utiliser pour créer des graphiques de présentation et calculer des mesures personnalisées (par exemple, des mesures de profit et de retour sur investissement complexes). Vous pouvez visualiser ces éléments dans l'Decision List Viewer pendant la création du modèle.

Exemple : Le service marketing d'une institution financière souhaite obtenir des résultats plus rentables au cours des prochaines campagnes en présentant à chaque client une offre adaptée. Vous pouvez utiliser un modèle Liste de décision pour identifier les caractéristiques des clients

les plus à même de répondre favorablement sur la base des promotions précédentes et pour générer un fichier d'adresses en fonction des résultats. [Pour plus d'informations, reportez-vous à la section Modélisation de la réponse client \(Liste de décision\) dans le chapitre 11 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Un champ cible catégoriel unique avec un niveau de mesure de type *Booléen* ou *Nominal* qui indique le résultat binaire que vous souhaitez prévoir (oui/non) et au moins un champ d'entrée. Lorsque le type du champ cible est *Nominal*, vous devez choisir manuellement une valeur unique à traiter en tant que **correspondance** ou **réponse** ; toutes les autres valeurs sont regroupées en tant qu'**absence de correspondance**. Il est également possible de spécifier un champ d'effectif facultatif. Les champs date/heure continus sont ignorés. Les entrées d'intervalle numériques continues sont automatiquement mises en intervalles par l'algorithme, comme indiqué dans l'onglet Expert du noeud de modélisation. Pour un contrôle plus précis de la création d'intervalles, ajoutez un noeud Discrétiser en amont et utilisez le champ mis en intervalles en tant qu'entrée avec un niveau de mesure *Ordinal*.

Options du modèle Liste de décision

Figure 9-6
Noeud Liste de décision : Onglet Modèle

The screenshot shows a software dialog box titled 'response[1]' with a 'Modèle' tab selected. The dialog is organized into several sections:

- Nom du modèle:** Radio buttons for 'Auto' (selected) and 'Personnalisé', followed by an empty text field.
- Utiliser les données partitionnées:** A checked checkbox.
- Créer un modèle pour chaque division:** A checked checkbox.
- Mode:** Radio buttons for 'Générer un modèle' and 'Lancer une session interactive' (selected). Below it is a checked checkbox for 'Utiliser les informations de session interactive enregistrées'.
- Valeur cible:** A text field containing '1' and a small icon.
- Rechercher les segments avec:** A dropdown menu showing 'Probabilité élevée'.
- Nombre maximal de segments:** A spinner box set to '3'.
- Taille de segment minimale:** Two spinner boxes: 'En tant que pourcentage de segment précédent (%)' set to '5,0' and 'En tant que valeur absolue (N)' set to '50'.
- Règles de segment:** A spinner box for 'Nombre maximal d'attributs' set to '5', a checked checkbox for 'Autoriser la réutilisation des attributs', and a spinner box for 'Intervalle de confiance des nouvelles conditions (%)' set to '85,0'.

At the bottom, there are buttons for 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser'.

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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Mode. Indique la méthode utilisée pour créer le modèle.

- **Générer un modèle.** Génère automatiquement un modèle sur la palette Modèles lorsque le noeud est exécuté. Le modèle obtenu peut être ajouté à des flux, à des fins de scoring, mais ne peut plus être modifié.
- **Lancer une session interactive.** Ouvre la fenêtre (de sortie) de modélisation interactive de l'Decision List Viewer, qui permet de faire une sélection parmi plusieurs choix et d'appliquer à plusieurs reprises l'algorithme avec différents paramètres, afin de développer ou de modifier progressivement le modèle. [Pour plus d'informations, reportez-vous à la section Decision List Viewer sur p. 240.](#)
- **Utiliser les informations de session interactive enregistrées.** Lance une session interactive à l'aide de paramètres précédemment enregistrés. Il est possible d'enregistrer les paramètres interactifs à partir de l'Decision List Viewer, à l'aide du menu Générer (pour créer un modèle ou un noeud de modélisation) ou du menu Fichier (pour mettre à jour le noeud à partir duquel la session a été lancée).

Valeur cible. Spécifie la valeur du champ cible qui indique le résultat à modéliser. Par exemple, si l'attrition du champ cible est codée 0 = no et 1 = yes, spécifiez 1 pour identifier les règles qui indiquent les enregistrements qui ont une probabilité d'attrition.

Rechercher les segments avec. Indique si la recherche de la variable cible doit rechercher une Probabilité élevée ou une Probabilité faible d'occurrence. Le fait de rechercher et d'exclure ces segments peut s'avérer utile pour améliorer votre modèle, en particulier lorsque la probabilité du reste est faible.

Nombre maximal de segments. Spécifie le nombre maximal de segments à renvoyer. Les N premiers segments sont créés, dans lesquels le meilleur segment est celui dont la probabilité est la plus forte ou, si plusieurs modèles ont la même probabilité, la couverture la plus élevée. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Taille minimale de segment. Les deux valeurs ci-dessous indiquent la taille minimale du segment. La valeur la plus importante des deux est prioritaire. Par exemple, si la valeur de pourcentage est supérieure à la valeur absolue, la valeur de pourcentage est prioritaire.

- **Comme pourcentage du segment précédent (%).** Spécifie la taille minimale du groupe en pourcentage d'enregistrements. Les valeurs minimale et maximale autorisées sont respectivement 0 et 99,9.
- **Comme valeur absolue (N).** Spécifie la taille minimale du groupe en nombre absolu d'enregistrements. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Règles de segment.

Nombre d'attributs maximal. Indique le nombre maximal de conditions par règle de segment. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

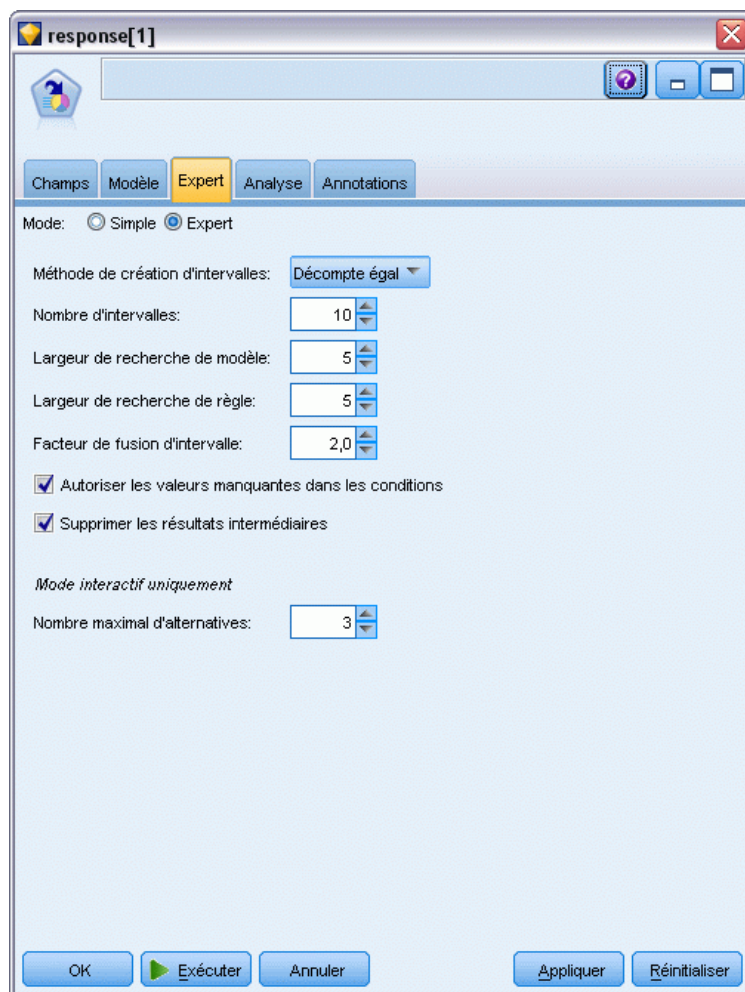
- **Autoriser la réutilisation d'attribut.** Si cette option est activée, chaque cycle peut prendre en compte tous les attributs, même ceux qui ont été utilisés dans des cycles précédents. Les conditions d'un segment se font sous forme de cycles, où chaque cycle ajoute une nouvelle condition. Le nombre de cycles est défini à l'aide du paramètre Nombre maximal d'attributs.

Intervalle de confiance des nouvelles conditions (%). Spécifie le niveau de confiance pour les tests de signification des segments. Ce paramètre joue un rôle significatif dans le nombre de segments (le cas échéant) renvoyés ainsi que dans le nombre de conditions par règle de segment. Plus la

valeur est élevée, plus l'ensemble de résultats renvoyé est petit. Les valeurs minimale et maximale autorisées sont respectivement 50 et 99,9.

Noeud Liste de décision - Options expert

Figure 9-7
Noeud Liste de décision : Onglet Expert



Les options Expert vous permettent d'affiner le processus de création de modèle.

Méthode de création d'intervalles. Méthode utilisée pour mettre en intervalle des champs continus (Nombre égal ou Largeur égale).

Nombre d'intervalles. Nombre d'intervalles à créer pour les champs continus. La valeur minimale autorisée est 2 ; aucune valeur maximale n'est définie.

Largeur de recherche de modèle. Nombre maximal de résultats de modèle par cycle pouvant être utilisés pour le cycle suivant. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Largeur de recherche de règle. Nombre maximal de résultats de règle par cycle pouvant être utilisés pour le cycle suivant. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Facteur de fusion d'intervalles. Augmentation minimale d'un segment lors de sa fusion avec son voisin. La valeur minimale autorisée est 1,01 ; aucune valeur maximale n'est définie.

- **Autoriser les valeurs manquantes dans les conditions.** True pour autoriser le test des IS MISSING dans les règles.
- **Supprimer les résultats intermédiaires.** Lorsque la valeur est True, seuls les résultats finaux du processus de recherche sont renvoyés. Un résultat final est un résultat ayant atteint le dernier stade d'affinage dans le processus de recherche. Lorsque la valeur est False, les résultats intermédiaires sont également renvoyés.

Nombre maximal d'alternatives. Spécifie le nombre maximal d'alternatives qui pourront être obtenues au moment de l'exécution de la tâche exploratoire. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Notez que la tâche d'exploration ne renvoie que le nombre réel d'alternatives, jusqu'au maximum spécifié. Par exemple, si le maximum est fixé à 100 et que seules 3 alternatives sont trouvées, seules ces 3 dernières sont affichées.

Nugget du modèle Liste de décision

Un modèle se compose d'une liste de **segments**, chacun étant défini par une **règle** qui sélectionne les enregistrements avec correspondance. Vous pouvez facilement visualiser ou modifier les segments avant de générer le modèle, et choisir les segments à inclure ou exclure. Lorsqu'ils sont utilisés au cours du scoring, les modèles Liste de décision renvoient *Yes* (oui) pour les segments inclus et *\$null\$* pour tous les autres éléments, y compris le reste. Ce contrôle direct sur le scoring fait des modèles Liste de décision une solution idéale pour la génération de fichiers d'adresses ; ils sont largement utilisés dans la gestion de la relation client, notamment dans les applications marketing ou de centre d'appels.

Figure 9-8
Nugget du modèle Liste de décision

ID	Règles de segment	Score	Couvertu...	Fréquence	Probabilité
	Tous les segments comprenant le r...		13 504	1 952	14,45 %
1	months_customer months_customer = "0"	Éléments ...	1 747	0	0,00 %
2	rfm_score rfm_score <= 0.000	Éléments ...	6 003	0	0,00 %
3	rfm_score, income rfm_score > 12.333 et income > 52213.000	1	555	456	82,16 %
4	income income > 55267.000	1	643	551	85,69 %
5	number_transactions, rfm_sc number_transactions > 2 et rfm_score > 12.333	1	533	206	38,65 %

Lorsque vous exécutez un flux contenant un modèle Liste de décision, le noeud ajoute trois nouveaux champs contenant le score *I* (signifiant *Yes* (oui)) pour les champs inclus ou *\$null* pour les champs exclus, la probabilité (taux de correspondance) du segment dans lequel se situe l'enregistrement et le numéro ID du segment. Le nom des nouveaux champs est formé à partir du nom du champ de sortie sur lequel porte la prévision, auquel est ajouté le préfixe *\$D-* pour désigner le score, *\$DP-* pour désigner la probabilité et *\$DI-* pour désigner l'ID du segment.

L'évaluation du modèle est basée sur la valeur cible spécifiée au moment de la création du modèle. Vous pouvez exclure manuellement des segments de façon à ce que leur score soit *\$null\$*. Par exemple, si vous exécutez une recherche à faible probabilité pour trouver les segments présentant des taux de correspondance inférieurs à la moyenne, le score de ces segments à "faible taux" sera *Yes* (oui) sauf si vous les excluez manuellement. Si nécessaire, les valeurs nulles peuvent être recodées sous la forme *No* (non) à l'aide d'un noeud Calculer ou Remplacer.

PMML

Un modèle Liste de décision peut être stocké sous la forme d'un modèle PMML RuleSetModel présentant un critère de sélection Premier résultat. Toutefois, il est prévu que toutes les règles aient le même score. Pour permettre des modifications du champ cible ou de la valeur cible, plusieurs modèles Ensemble de règles peuvent être stockés dans un fichier et être appliqués dans l'ordre, les observations sans correspondance avec le premier modèle étant transmises au deuxième, etc. Le nom d'algorithme *DecisionList* est utilisé pour indiquer ce comportement hors normes ; seuls les modèles Ensemble de règles portant ce nom sont reconnus en tant que modèles Liste de décision et évalués comme tels.

Paramètres du nugget du modèle Liste de décision

L'onglet Paramètres pour un nugget du modèle Liste de décision vous permet d'obtenir les scores de propension et d'activer ou de désactiver l'optimisation SQL. Cet onglet est disponible uniquement après l'ajout d'un nugget de modèle à un flux.

Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

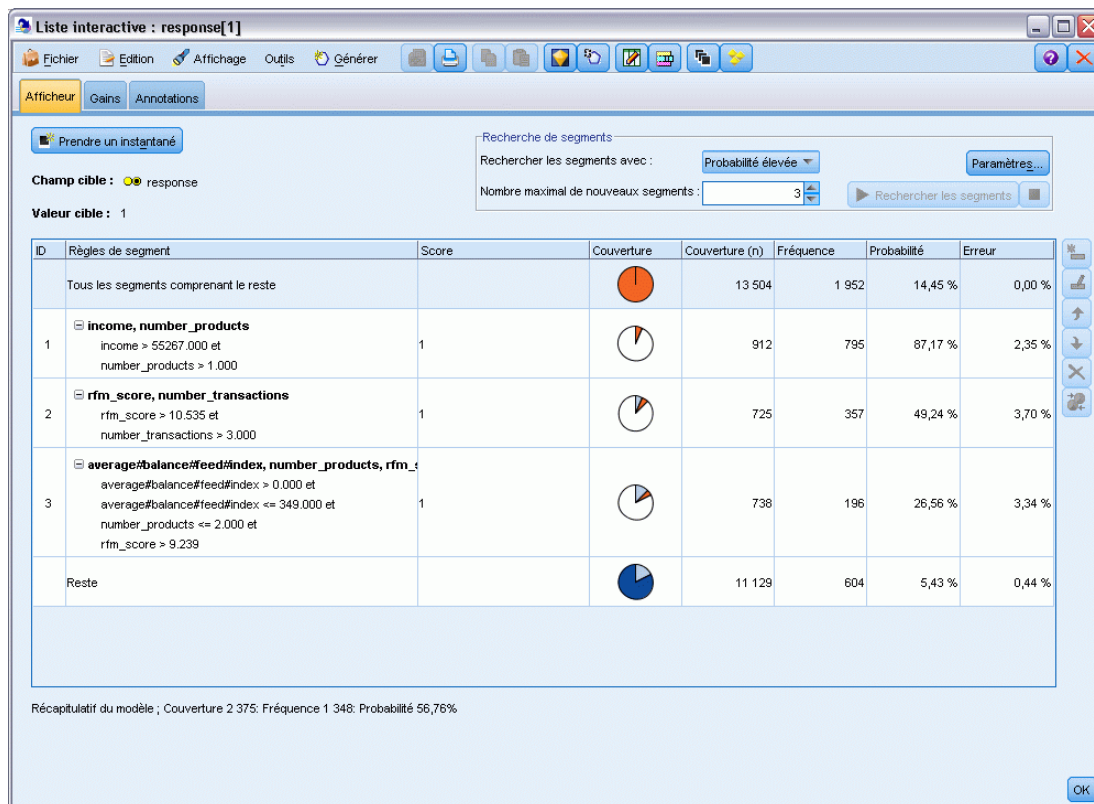
Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Evaluer en convertissant en SQL natif. Si cette option est sélectionnée, le SQL est généré pour évaluer le modèle de manière native dans l'application.

Decision List Viewer

L'interface graphique conviviale de Decision List Viewer, basée sur les tâches, simplifie le processus de création de modèles ; en effet, elle vous libère des détails de bas niveau des techniques de Data mining et vous permet de vous concentrer sur les parties de l'analyse qui nécessitent l'intervention de l'utilisateur, telles que la définition d'objectifs, le choix de groupes cible, l'analyse des résultats et la sélection du modèle optimal.

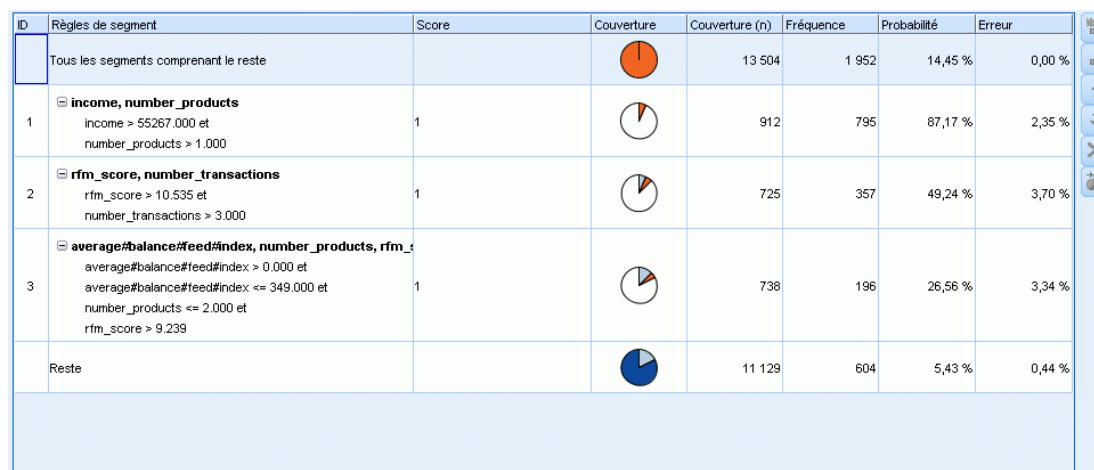
Figure 9-9
Afficheur interactif Liste de décision



Panneau du modèle de travail

Le panneau du modèle de travail affiche le modèle actuel, notamment les tâches d'exploration et d'autres actions qui s'appliquent au modèle de travail.

Figure 9-10
Panneau du modèle de travail



ID. Identifie l'ordre séquentiel des segments. Les segments du modèle sont calculés de façon séquentielle, conformément à leur numéro d'identification.

Règles de segment. Fournit le nom du segment et les conditions définies correspondantes. Par défaut, le nom de segment est le nom de champ ou les noms de champ concaténés utilisés dans les conditions, séparés par une virgule.

Score. Champ que vous voulez prévoir et dont la valeur est censée avoir un lien avec les valeurs des autres champs (les valeurs prédites).

Remarque : l'affichage de ces options peut être activé via la boîte de dialogue [Organisation des mesures de modèle](#).

Couverture. Le graphique sectoriel identifie de manière visuelle la couverture de chaque segment par rapport à la couverture globale.

Couverture (n). Répertoire la couverture de chaque segment par rapport à la couverture globale.

Effectif. Répertoire le nombre de correspondances reçues par rapport à la couverture. Par exemple, une couverture de 79 et un effectif de 50 indiquent l'obtention de 50 réponses sur 79 pour le segment sélectionné.

Probabilité. Indique la probabilité du segment. Par exemple, pour une couverture de 79 et un effectif de 50, la probabilité du segment est de 63,29 % (50 divisé par 79).

Erreur. Indique l'erreur du segment.

Les informations situées au bas du panneau indiquent la couverture, l'effectif et la probabilité du modèle global.





Barre d'outils du modèle de travail



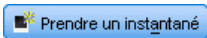






La barre d'outils du panneau du modèle de travail fournit les fonctions suivantes.

Remarque : Certaines fonctions sont également disponibles en cliquant avec le bouton droit de la souris sur un segment du modèle.

Table 9-1

Boutons de la barre d'outils du modèle de travail

	Lance la boîte de dialogue Générer un nouveau modèle ; celle-ci fournit des options permettant de créer un nouveau nugget de modèle.
	Enregistre l'état actuel de la session interactive. Le noeud de modélisation Liste de décision est mis à jour avec les paramètres en cours, y compris les tâches d'exploration, les instantanés de modèle, les sélections de données et les mesures personnalisées. Pour restaurer une session à cet état, cochez la case Utiliser les informations de session interactive enregistrées de l'onglet Modèle du noeud de modélisation et cliquez sur Exécuter.
	Affiche la boîte de dialogue Organiser les mesures de modèle. Pour plus d'informations, reportez-vous à la section Organisation des mesures de modèle sur p. 259.
	Affiche la boîte de dialogue Organiser les sélections de données Pour plus d'informations, reportez-vous à la section Organisation des sélections de données sur p. 252.

	Affiche l'onglet Instantanés. Pour plus d'informations, reportez-vous à la section Onglet Instantanés sur p. 245.
	Affiche l'onglet Alternatives. Pour plus d'informations, reportez-vous à la section Onglet Alternatives sur p. 243.
	Prend un instantané de la structure du modèle actuel. Les instantanés apparaissent dans l'onglet Instantanés, et sont couramment utilisés à des fins de comparaison de modèles.
	Lance la boîte de dialogue Insertion de segments ; celle-ci fournit des options permettant de créer des segments de modèle.
	Lance la boîte de dialogue Edition de règles de segment ; celle-ci fournit des options permettant d'ajouter des conditions aux segments de modèle ou de modifier des conditions préalablement définies.
	Déplace le segment sélectionné vers le haut de la hiérarchie du modèle.
	Déplace le segment sélectionné vers le bas de la hiérarchie du modèle.
	Supprime le segment sélectionné.
	Indique si le segment sélectionné est inclus dans le modèle. En cas d'exclusion, les résultats du segment sont ajoutés au reste. Cette fonction est différente de la suppression d'un segment car vous avez, ici, la possibilité de le réactiver.

Onglet Alternatives

Généré lorsque vous cliquez sur Rechercher les segments, l'onglet Alternatives répertorie tous les autres résultats de l'exploration pour le modèle ou le segment sélectionné dans le panneau du modèle de travail.

- Pour convertir une alternative en modèle de travail, surlignez l'alternative et cliquez sur Charger ; le modèle alternatif est affiché dans le panneau du modèle de travail.

Remarque : L'onglet Alternatives ne s'affiche que si vous avez défini le Nombre maximal d'alternatives dans l'onglet Expert du noeud de modélisation Liste de décision afin de créer plus d'une alternative.

Figure 9-11
onglet Alternatives

The screenshot shows a software window titled "Albums de modèles". It contains a table with the following data:

Nom	Cible	Nbre de segments	Couverture	Effect.	Prob.
Alternative 1	1	3	2 375	1 348	56,76 %
Alternative 2	1	3	2 368	1 326	56,00 %
Alternative 3	1	3	2 380	1 329	55,84 %

Below this table is a section titled "Aperçu d'alternative" with a detailed table:

ID	Règles de segment	Score	Couvertur...	Fréquence	Probabilité
	Tous les segments comprenant le reste		13 504	1 952	14,45 %
1	income, number_products income > 55267.000 et number_products > 1.000	1	912	795	87,17 %
2	rfm_score, number_transactions rfm_score > 12.333 et number_transactions > 2.000	1	737	360	48,85 %
3	number_transactions, income number_transactions > 0.000 et number_transactions <= 1.000 et income > 46072.000	1	731	174	23,80 %

At the bottom of the window, there are buttons for "Chargement", "Alternatives", "Instantanés", "OK", "Annuler", and "Aide".

Chaque alternative de modèle générée affiche des informations spécifiques sur le modèle :

Nom. Chaque alternative est numérotée de manière séquentielle. La première alternative comporte généralement les meilleurs résultats.

Cible. Indique la valeur cible. Par exemple : 1, qui est égale à "true" (vrai).

Nbre de segments. Le nombre de règles de segment utilisées dans le modèle alternatif.

Couverture. La couverture du modèle alternatif.

Effectif Le nombre de correspondances par rapport à la couverture.

Prob. Indique le pourcentage de probabilité du modèle alternatif.

Remarque : Les résultats alternatifs ne sont pas enregistrés avec le modèle ; ils ne sont valides que pendant la session active.

Onglet Instantanés

Un instantané constitue une vue d'un modèle à un moment spécifique dans le temps. Par exemple, vous pouvez prendre un instantané du modèle lorsque vous souhaitez charger un modèle alternatif différent dans le panneau de modèle de travail, mais ne souhaitez pas perdre le travail effectué sur le modèle actuel. L'onglet Instantanés répertorie tous les instantanés de modèle pris manuellement pour n'importe quel nombre d'états de modèle de travail.

Remarque : les instantanés sont enregistrés avec le modèle. Nous vous recommandons de prendre un instantané lorsque vous chargez le premier modèle. Cet instantané préservera alors la structure du modèle d'origine, vous pouvez toujours revenir à l'état du modèle initial. Le nom de l'instantané généré est affiché sous la forme d'un horodatage indiquant quand l'instantané a été généré.

Création d'un instantané de modèle

- ▶ Sélectionnez l'alternative/le modèle approprié à afficher dans le panneau du modèle de travail.
- ▶ Apportez les modifications nécessaires au modèle de travail.
- ▶ Cliquez sur Prendre un instantané. Un nouvel instantané apparaît dans l'onglet Instantanés.

Figure 9-12
Onglet Instantanés

Nom	Cible	Nbre de segments	Couverture	Effect.	Prob.
Instantané 1	1	3	2 375	1 348	56,76 %

ID	Règles de segment	Score	Couverture...	Fréquence	Probabilité
	Tous les segments comprenant le reste		13 504	1 952	14,45 %
1	income > 55267.000 et number_products > 1.000	1	912	795	87,17 %
2	rfm_score > 10.535 et number_transactions > 3.000	1	725	357	49,24 %
3	average#balance#feed#index > 0.000 et average#balance#feed#index <= 349.01 number_products <= 2.000 et rfm_score > 9.239		738	196	26,56 %
	Reste		11 129	604	5,43 %

Nom. Nom de l’instantané. Vous pouvez modifier le nom d’un instantané en double-cliquant sur son nom.

Cible. Indique la valeur cible. Par exemple : 1, qui est égale à “true” (vrai).

Nbre de segments. Le nombre de règles de segment utilisés dans le modèle.

Couverture. La couverture du modèle.

Effectif Le nombre de correspondances par rapport à la couverture.

Prob. Indique le pourcentage de probabilité du modèle.

- ▶ Pour convertir un instantané en modèle de travail, surlignez l’instantané et cliquez sur Charger ; le modèle de l’instantané est affiché dans le panneau du modèle de travail.
- ▶ Vous pouvez supprimer un instantané en cliquant sur Supprimer, ou en cliquant dessus avec le bouton droit de la souris, puis en choisissant Supprimer dans le menu.

Utilisation de l'Decision List Viewer

La création d'un modèle destiné à prévoir au mieux la réponse et le comportement du client s'effectue à diverses étapes. Lorsque l'Decision List Viewer démarre, le modèle de travail reçoit les segments et les mesures de modèle définis, afin que vous puissiez commencer une tâche d'exploration, modifier les segments/mesures selon les besoins, et générer un nouveau modèle ou un nouveau noeud de modélisation.

Vous pouvez ajouter une ou plusieurs règles de segment jusqu'à développer un modèle satisfaisant. Vous pouvez ajouter des règles de segment au modèle en exécutant des tâches d'exploration ou à l'aide de la fonction *Edit* la règle de segment.

Au cours du processus de création du modèle, vous pouvez évaluer la performance du modèle en validant le modèle par rapport aux données de mesure, en visualisant le modèle dans un graphique ou en générant des mesures Excel personnalisées.

Lorsque vous êtes certain de la qualité du modèle, vous pouvez générer un nouveau modèle et le placer dans l'espace de travail IBM® SPSS® Modeler ou dans la palette Modèle.

Tâches d'exploration

Une **tâche d'exploration** est un ensemble de paramètres qui détermine la manière dont les nouvelles règles sont générées. Certains de ces paramètres peuvent être sélectionnés, ce qui confère la flexibilité d'adaptation des modèles à de nouvelles situations. Une tâche est constituée d'un modèle de tâche (type), d'une cible et d'un build (ensemble de données d'exploration).

Les sections suivantes décrivent les diverses opérations de tâche d'exploration :

- [Exécution de tâches d'exploration](#)
- [Création et modification d'une tâche d'exploration](#)
- [Organisation des sélections de données](#)

Exécution de tâches d'exploration

L'Decision List Viewer permet d'ajouter manuellement des règles de segment à un modèle en exécutant des tâches d'exploration, ou en copiant et collant des règles de segment entre les modèles. Une tâche d'exploration contient des informations sur la manière de générer de nouvelles règles de segment (les paramètres de Data mining tels que la stratégie de recherche, les attributs source, la largeur de recherche, le niveau de confiance, etc.), le comportement du client à prévoir et les données à examiner. L'objectif d'une tâche d'exploration est de rechercher les meilleures règles de segment possibles.

Pour générer une règle de segment de modèle en exécutant une tâche d'exploration :

- ▶ Cliquez sur la ligne *Reste*. Si des segments sont déjà affichés dans le panneau du modèle de travail, vous pouvez également sélectionner l'un des segments pour rechercher d'autres règles en fonction de ce segment. Après avoir sélectionné le *reste* ou le segment, utilisez l'une des méthodes suivantes pour générer le modèle ou des modèles alternatifs :
 - Dans le menu *Outils*, sélectionnez *Rechercher les segments*.

- Cliquez avec le bouton droit de la souris sur la ligne/le segment Reste et choisissez Rechercher les segments.
- Cliquez sur le bouton Rechercher les segments dans le panneau du modèle de travail.

Pendant le traitement de la tâche, la progression du traitement est affichée en bas de l'espace de travail et vous êtes informé lorsque la tâche est terminée. La durée exacte d'exécution d'une tâche dépend de la complexité de la tâche d'exploration et de la taille de l'ensemble de données. S'il n'y a qu'un modèle unique dans les résultats, il est affiché dans le panneau de l'espace de travail aussitôt que la tâche est terminée ; cependant, si les résultats contiennent plus d'un modèle, ils sont affichés dans l'onglet Alternatives.

Remarque : les résultats d'une tâche comprennent un modèle ou n'en comprennent aucun, ou la tâche échoue.

Le processus de recherche de nouvelles règles de segment peut être répété jusqu'à ce qu'aucune nouvelle règle ne soit ajoutée au modèle. Cela signifie que tous les groupes de clients significatifs ont été trouvés.

Il est possible d'exécuter une tâche d'exploration sur n'importe quel segment de modèle existant. Si le résultat de la tâche n'est pas celui recherché, vous pouvez lancer une autre tâche d'exploration sur le même segment. Vous obtiendrez ainsi d'autres règles basées sur le segment sélectionné. Les segments « au-dessous » du segment sélectionné (à savoir, ajoutés au modèle plus tard que le segment sélectionné) sont remplacés par les nouveaux segments car chaque segment dépend de ses prédécesseurs.

Création et modification d'une tâche d'exploration

Une tâche d'exploration constitue le mécanisme qui recherche l'ensemble des règles constituant un modèle de données. Parallèlement aux critères de recherche définis dans le modèle sélectionné, une tâche définit également la cible (la question réelle ayant motivé l'analyse, telle que le nombre de clients susceptibles de répondre à un publipostage) et identifie les ensembles de données à utiliser. L'objectif d'une tâche d'exploration est de rechercher les meilleurs modèles possibles.

Créer une tâche d'exploration

Pour créer une tâche d'exploration :

- ▶ Sélectionnez le segment à partir duquel exploiter des conditions de segment supplémentaires.
- ▶ Cliquez sur Paramètres. La boîte de dialogue Créer/Editer une tâche d'exploration apparaît. Cette boîte de dialogue propose des options destinées à la définition de la tâche d'exploration.
- ▶ Procédez à toutes les modifications nécessaires puis cliquez sur OK pour revenir au panneau du modèle de travail. L'Decision List Viewer utilise les paramètres comme valeurs par défaut pour exécuter chaque tâche jusqu'à ce qu'une tâche ou des paramètres alternatifs soient sélectionnés.
- ▶ Cliquez sur Rechercher les segments pour lancer la tâche d'exploration sur le segment sélectionné.

Editer une tâche d'exploration

La boîte de dialogue Créer/Editer une tâche d'exploration fournit des options permettant de définir une nouvelle tâche d'exploration ou d'éditer une tâche déjà existante.

La plupart des paramètres disponibles pour les tâches d'exploration sont semblables à ceux présents dans le noeud Liste de décision. Les exceptions sont affichées en dessous. [Pour plus d'informations, reportez-vous à la section Options du modèle Liste de décision sur p. 235.](#)

Figure 9-13
Boîte de dialogue Créer/Editer une tâche d'exploration

Charge des paramètres : Si vous avez créé plus d'une tâche d'exploration, sélectionnez la tâche requise.

Nouveau... Cliquez pour créer une nouvelle tâche d'exploration basée sur les paramètres de la tâche actuellement affichée.

Cible

Champ cible : Champ que vous voulez prévoir et dont la valeur est censée avoir un lien avec les valeurs des autres champs (les valeurs prédites).

Valeur cible. Spécifie la valeur du champ cible qui indique le résultat à modéliser. Par exemple, si l'attrition du champ cible est codée 0 = no et 1 = yes, spécifiez 1 pour identifier les règles qui indiquent les enregistrements qui ont une probabilité d'attrition.

Paramètres simples

Nombre maximal d'alternatives. Spécifie le nombre d'alternatives qui seront affichées au moment de l'exécution de la tâche exploratoire. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Paramètres Expert

Modifier... Ouvre la boîte de dialogue Editer les paramètres avancés, qui permet de définir les paramètres avancés. [Pour plus d'informations, reportez-vous à la section Editer les paramètres avancés sur p. 250.](#)

Données

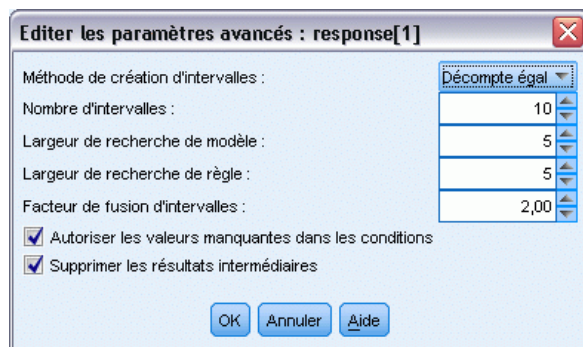
Sélection du build. Fournit des options permettant de spécifier la mesure d'évaluation que l'Decision List Viewer doit analyser pour trouver de nouvelles règles. Les mesures d'évaluation répertoriées sont créées/éditées dans la boîte de dialogue Organiser les sélections de données.

Champs disponibles. Fournit des options permettant d'afficher tous les champs ou de sélectionner manuellement les champs à afficher.

Modifier... Si l'option Personnalisé est sélectionnée, la boîte de dialogue Personnaliser les champs disponibles s'ouvre, ce qui permet de sélectionner les champs disponibles en tant qu'attributs de segment trouvés par la tâche d'exploration. [Pour plus d'informations, reportez-vous à la section Personnaliser les champs disponibles sur p. 251.](#)

Editer les paramètres avancés

Figure 9-14
Paramètres avancés



La boîte de dialogue Editer les paramètres avancés fournit les options de configuration suivantes.

Méthode de création d'intervalles. Méthode utilisée pour mettre en intervalle des champs continus (Nombre égal ou Largeur égale).

Nombre d'intervalles. Nombre d'intervalles à créer pour les champs continus. La valeur minimale autorisée est 2 ; aucune valeur maximale n'est définie.

Largeur de recherche de modèle. Nombre maximal de résultats de modèle par cycle pouvant être utilisés pour le cycle suivant. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

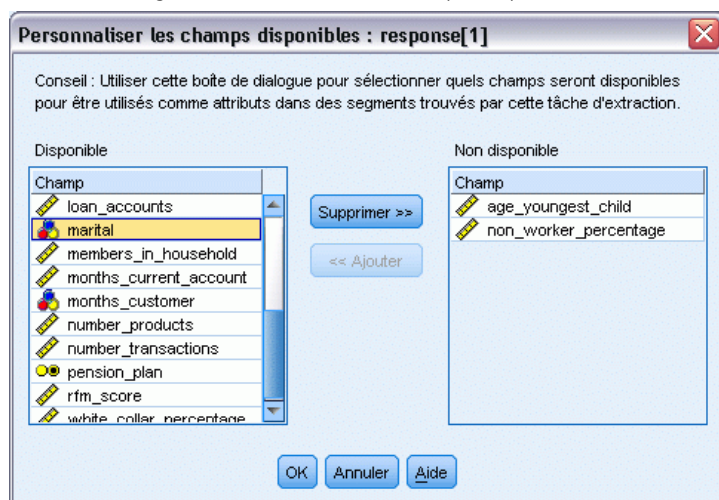
Largeur de recherche de règle. Nombre maximal de résultats de règle par cycle pouvant être utilisés pour le cycle suivant. La valeur minimale autorisée est 1 ; aucune valeur maximale n'est définie.

Facteur de fusion d'intervalles. Augmentation minimale d'un segment lors de sa fusion avec son voisin. La valeur minimale autorisée est 1,01 ; aucune valeur maximale n'est définie.

- **Autoriser les valeurs manquantes dans les conditions.** True pour autoriser le test des IS MISSING dans les règles.
- **Supprimer les résultats intermédiaires.** Lorsque la valeur est True, seuls les résultats finaux du processus de recherche sont renvoyés. Un résultat final est un résultat ayant atteint le dernier stade d'affinage dans le processus de recherche. Lorsque la valeur est False, les résultats intermédiaires sont également renvoyés.

Personnaliser les champs disponibles

Figure 9-15
Boîte de dialogue Personnaliser les champs disponibles



La boîte de dialogue Personnaliser les champs disponibles permet de sélectionner les champs disponibles en tant qu'attributs de segment trouvés par la tâche d'exploration.

Disponible. Répertorie les champs actuellement disponibles comme attributs de segment. Pour supprimer des champs de la liste, sélectionnez ces champs, et cliquez sur Supprimer >>. Les champs sélectionnés passent de la liste Disponible à la liste Non disponible.

Non disponible. Répertorie les champs non disponibles comme attributs de segment. Pour inclure des champs dans la liste Disponible, sélectionnez ces champs et cliquez sur << Ajouter. Les champs sélectionnés passent de la liste Non disponible à la liste Disponible.

Organisation des sélections de données

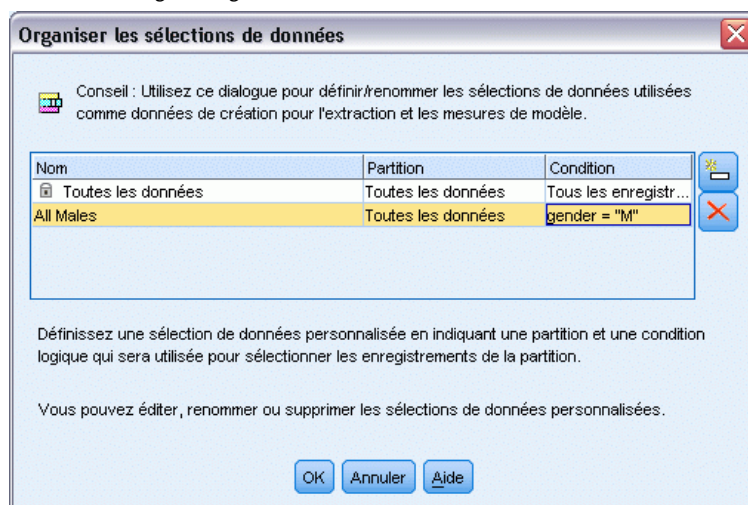
En organisant des sélections de données (un ensemble de données d'exploration), vous pouvez spécifier les mesures d'évaluation que l'Decision List Viewer doit analyser pour trouver de nouvelles règles et choisir les sélections de données utilisées comme base des mesures.

Pour organiser des sélections de données :

- Dans le menu Outils, sélectionnez Organiser les sélections de données, ou cliquez avec le bouton droit de la souris sur un segment et sélectionnez cette option. La boîte de dialogue Organiser les sélections de données apparaît.

Figure 9-16

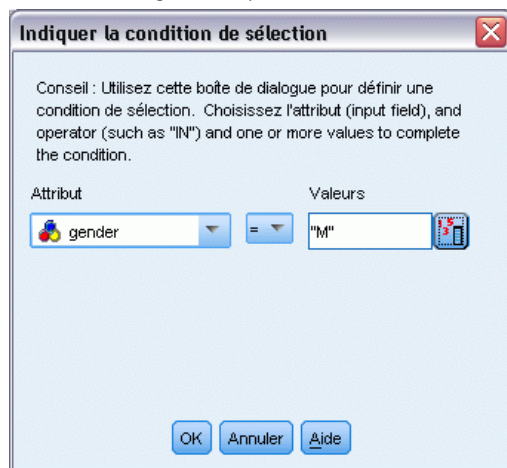
Boîte de dialogue Organiser les sélections de données



Remarque : La boîte de dialogue Organiser les sélections de données permet également d'éditer ou de supprimer des sélections de données existantes.

- Cliquez sur le bouton Ajouter une nouvelle sélection de données. Une nouvelle entrée de sélection de données est ajoutée à la table existante.
- Cliquez sur Nom et entrez un nom de sélection approprié.
- Cliquez sur Partitionner et sélectionnez un type de partition approprié.
- Cliquez sur Condition et sélectionnez une option de condition appropriée. Lorsque vous sélectionnez Spécifier, la boîte de dialogue Indiquer la condition de sélection apparaît, fournissant des options permettant de définir des conditions de champ spécifiques.

Figure 9-17
Boîte de dialogue Indiquer la condition de sélection



- Définissez la condition appropriée et cliquez sur OK.

Les sélections de données sont disponibles via la liste déroulante Sélection du build de la boîte de dialogue Créer/Editer une tâche d'exploration. Cette liste permet de sélectionner la mesure d'évaluation utilisée pour une tâche d'exploration spécifique.

Règles de segment

Pour trouver des règles de segment de modèle, exécutez une tâche d'exploration en fonction d'un modèle de tâche. Vous pouvez ajouter manuellement des règles de segment à un modèle à l'aide des fonctions Insérer un segment ou Editer la règle de segment.

Si vous choisissez d'explorer de nouvelles règles de segment, les résultats éventuels sont affichés dans l'onglet Afficheur de la boîte de dialogue Liste interactive. Vous pouvez rapidement affiner votre modèle en sélectionnant les résultats alternatifs à partir de la boîte de dialogue Albums de modèles et en cliquant sur Charger. Ainsi, vous pouvez tester différents résultats jusqu'à ce que vous soyez prêt à concevoir un modèle décrivant avec précision votre groupe cible optimal.

Insertion de segments

Vous pouvez ajouter manuellement des règles de segment à un modèle à l'aide de la fonction Insérer un segment.

Pour ajouter une condition de règle de segment à un modèle :

- Dans la boîte de dialogue Liste interactive, sélectionnez un emplacement de modèle où ajouter un nouveau segment. Le nouveau segment sera directement inséré au-dessus du segment sélectionné.

- ▶ Dans le menu Edition, choisissez Insérer un segment ou accédez à cette option en cliquant avec le bouton droit de la souris sur un segment.

La boîte de dialogue Insérer un segment apparaît, vous permettant d'insérer les conditions de règle de ce nouveau segment.

- ▶ Cliquez sur Insérer. La boîte de dialogue Insérer une condition apparaît, vous permettant de définir les attributs de la nouvelle condition de règle.
- ▶ Sélectionnez un champ et un opérateur dans les listes déroulantes.

Remarque : Si vous sélectionnez l'opérateur Absent, la condition sélectionnée fonctionnera comme une condition d'exclusion et apparaîtra en rouge dans la boîte de dialogue Insérer une règle. Par exemple, lorsque la condition region = 'TOWN' apparaît en rouge, TOWN est alors exclu de l'ensemble de résultats.

- ▶ Entrez une ou plusieurs valeurs, ou cliquez sur l'icône Insérer une valeur pour afficher la boîte de dialogue Insérer une valeur. La boîte de dialogue vous permet de sélectionner une valeur définie pour le champ sélectionné. Par exemple, le champ marié proposera les valeurs Oui et Non.
- ▶ Cliquez sur OK pour revenir à la boîte de dialogue Insérer un segment. Cliquez une nouvelle fois sur OK pour ajouter le segment créé au modèle.

Le nouveau segment apparaîtra à l'emplacement de modèle indiqué.

Edition de règles de segment

La fonction Editer la règle de segment permet d'ajouter, de changer ou de supprimer des conditions de règle de segment.

Pour modifier une condition de règle de segment :

- ▶ Sélectionnez le segment de modèle à éditer.
- ▶ Dans le menu Edition, choisissez Editer la règle de segment ou cliquez avec le bouton droit de la souris sur la règle pour accéder à cette sélection.

La boîte de dialogue Editer la règle de segment apparaît.

- ▶ Sélectionnez la condition appropriée et cliquez sur Edition.

La boîte de dialogue Editer la condition qui apparaît vous permet de définir les attributs pour la condition de règle sélectionnée.

- ▶ Sélectionnez un champ et un opérateur dans les listes déroulantes.

Remarque : Si vous sélectionnez l'opérateur Absent, la condition sélectionnée fonctionnera comme une condition d'exclusion et apparaîtra en rouge dans la boîte de dialogue Editer la règle

de segment. Par exemple, lorsque la condition region = 'TOWN' apparaît en rouge, TOWN est alors exclu de l'ensemble de résultats.

- ▶ Entrez une ou plusieurs valeurs, ou cliquez sur le bouton Insérer une valeur pour afficher la boîte de dialogue Insérer une valeur. La boîte de dialogue vous permet de sélectionner une valeur définie pour le champ sélectionné. Par exemple, le champ marié proposera les valeurs Oui et Non.
- ▶ Cliquez sur OK pour revenir à la boîte de dialogue Editer la règle de segment. Cliquez une seconde fois sur OK pour revenir au modèle de travail.

Le segment sélectionné apparaît, présentant les conditions de règle mises à jour.

Suppression de conditions de règle de segment

Pour supprimer une condition de règle de segment :

- ▶ Sélectionnez le segment de modèle qui contient les conditions de règle à supprimer.
- ▶ Dans le menu Edition, sélectionnez Editer la règle de segment, ou cliquez avec le bouton droit de la souris sur le segment pour accéder à cette sélection.

La boîte de dialogue Editer la règle de segment apparaît et vous permet de supprimer une ou plusieurs conditions de règle de segment.

- ▶ Sélectionnez la condition de règle appropriée et cliquez sur Supprimer.
- ▶ Cliquez sur OK.

Lorsque vous supprimez au moins une condition de règle de segment, le panneau du modèle de travail actualise ses mesures.

Copie de segments

L'Decision List Viewer propose une méthode pratique pour la copie de segments de modèle. Lorsque vous souhaitez appliquer un segment d'un modèle à un autre modèle, il vous suffit de copier (ou de couper) ce segment et de le coller dans l'autre modèle. Vous pouvez également copier un segment d'un modèle affiché dans le panneau Aperçu de l'alternative et le coller dans le modèle affiché dans le panneau du modèle de travail. Ces fonctions Couper, Copier et Coller utilisent le Presse-papiers pour stocker ou extraire des données temporaires. En d'autres termes, les conditions et la cible sont copiées dans le Presse-papiers. L'utilisation du contenu du Presse-papiers n'est pas uniquement réservée à l'Decision List Viewer car ce contenu peut également être collé dans d'autres applications. Par exemple, lorsqu'il est collé dans un éditeur de texte, les conditions et la cible sont collées au format XML.

Pour copier ou couper des segments de modèle :

- ▶ Sélectionnez le segment de modèle que vous souhaitez utiliser dans un autre modèle.
- ▶ Dans le menu Edition, sélectionnez **Copier** (ou **Couper**), ou cliquez avec le bouton droit de la souris sur le segment de modèle, et sélectionnez **Copier** ou **Couper**.
- ▶ Ouvrez le modèle approprié (dans lequel le segment de modèle sera collé).

- Sélectionnez l'un des segments de modèle et cliquez sur **Coller**.

Remarque : à la place des commandes **Couper**, **Copier** et **Coller**, vous pouvez utiliser les combinaisons de touches : **Ctrl+X**, **Ctrl+C**, et **Ctrl+V**.

Le segment copié (ou coupé) est inséré au-dessus du segment de modèle précédemment sélectionné. Les mesures du segment collé et des segments situés au-dessous sont recalculées.

Remarque : les deux modèles de cette procédure doivent être basés sur le même modèle sous-jacent et contenir la même cible, sans quoi un message d'erreur apparaît.

Modèles alternatifs

Lorsqu'il y a plus d'un résultat, l'onglet Alternatives affiche les résultats de chaque tâche d'exploration. Chaque résultat comprend les conditions des données sélectionnées correspondant le plus à la cible, ainsi que les alternatives « acceptables ». Le nombre total d'alternatives affichées dépend des critères de recherche utilisés dans le processus d'analyse.

Pour afficher les modèles alternatifs :

- Cliquez sur le modèle alternatif dans l'onglet Alternatives. Les segments du modèle alternatif apparaissent ou remplacent les segments du modèle actuel dans le panneau Aperçu de l'alternative.
- Pour travailler avec un modèle alternatif dans le panneau de modèle de travail, sélectionnez le modèle puis cliquez sur Charger dans le panneau Aperçu de l'alternative ou cliquez avec le bouton droit de la souris sur le nom de l'alternative dans l'onglet Alternatives et sélectionnez Charger.

Remarque : les modèles alternatifs ne sont pas enregistrés lorsque vous générez un nouveau modèle.

Personnalisation d'un modèle

Les données ne sont pas statiques. Les clients déménagent, se marient et changent d'emploi. Les produits perdent leur positionnement sur le marché et deviennent obsolètes.

L'Decision List Viewer offre aux utilisateurs professionnels la possibilité d'adapter les modèles à de nouvelles situations rapidement et aisément. Vous pouvez modifier un modèle en procédant à une édition, à un classement par ordre de priorité, à une suppression ou à la désactivation de segments de modèle spécifiques.

Classement de segments par ordre de priorité

Vous pouvez classer les règles de modèle dans n'importe quel ordre. Par défaut, les segments de modèle apparaissent dans l'ordre de priorité, le premier segment présentant la priorité la plus élevée. Lorsque vous attribuez une priorité différente à un ou plusieurs segments, le modèle est modifié en conséquence. Vous pouvez modifier le modèle selon les besoins en déplaçant des segments vers un niveau de priorité supérieur ou inférieur.

Pour classer des segments de modèle par ordre de priorité :

- ▶ Sélectionnez le segment de modèle auquel vous souhaitez attribuer une priorité différente.
- ▶ Cliquez sur l'un des deux boutons fléchés de la barre d'outils du panneau du modèle de travail pour déplacer le segment de modèle sélectionné vers le haut ou le bas de la liste.

Une fois le classement par ordre de priorité effectué, tous les résultats d'évaluation précédents sont recalculés et les nouvelles valeurs sont affichées.

Suppression de segments**Pour supprimer un ou plusieurs segments :**

- ▶ Sélectionnez un segment de modèle.
- ▶ Dans le menu Edition, sélectionnez Supprimer le segment, ou cliquez sur le bouton de suppression de la barre d'outils du panneau du modèle de travail.

Les mesures du modèle modifié sont recalculées et le modèle est actualisé en conséquence.

Exclusion de segments

En recherchant des groupes spécifiques, vous baserez probablement vos actions sur une sélection de segments du modèle. Lorsque vous déployez un modèle, vous pouvez en exclure des segments. Les segments exclus sont évalués en tant que valeurs nulles. Exclure un segment ne signifie pas que ce segment n'est pas utilisé ; cela signifie que tous les enregistrements correspondant à cette règle sont exclus de la liste de diffusion. La règle est toujours appliquée, mais d'une façon différente.

Pour exclure des segments de modèle spécifiques :

- ▶ Sélectionnez un segment dans le panneau du modèle de travail.
- ▶ Cliquez sur le bouton Basculer l'exclusion de segment situé sur la barre d'outils du panneau du modèle de travail. La mention Exclusion apparaît à présent dans la colonne Cible sélectionnée du segment sélectionné.

Remarque : contrairement aux segments supprimés, les segments exclus peuvent être réutilisés dans le modèle final. Les segments exclus affectent les résultats des graphiques.

Utiliser une autre valeur cible

La boîte de dialogue Changer la valeur cible permet de modifier la valeur cible du champ cible actuel.

Les instantanés et les résultats de session comportant une valeur cible différente de celle du modèle de travail sont identifiés via l'attribution de la couleur jaune à l'arrière-plan de cette ligne. Ceci indique que le résultat d'instantané/de session est obsolète.

La boîte de dialogue **Créer/Editer une tâche d'exploration** affiche la valeur cible du modèle de travail actuel. La valeur cible n'est pas enregistrée avec la tâche d'exploration. Elle est, au contraire, extraite de la valeur du modèle de travail.

Lorsque vous convertissez un modèle enregistré en modèle de travail comportant une valeur cible différente de celle du modèle de travail actuel (par exemple, en éditant un résultat alternatif ou une copie d'un instantané), la valeur cible du modèle enregistré est modifiée de manière à être identique à celle du modèle de travail (la valeur cible affichée dans le panneau du modèle de travail n'est pas modifiée). Les mesures du modèle sont réévaluées avec la nouvelle cible.

Générer un nouveau modèle

La boîte de dialogue Générer un nouveau modèle fournit des options permettant de nommer le modèle et de sélectionner l'emplacement où doit être créé le noeud.

Nom du modèle. Sélectionnez Personnalisé pour modifier le nom généré automatiquement ou créer un nom unique pour le noeud selon son affichage dans l'espace de travail.

Créer le noeud sur. Sélectionner Espace de travail place le nouveau modèle dans l'espace de travail, sélectionner Palette GM le place dans la palette Modèles et sélectionner Les deux le place à la fois dans l'espace de travail et dans la palette Modèles.

Inclure l'état de session interactive. Lorsque cette option est activée, l'état de session interactive est préservé dans le modèle généré. Lorsque vous générez par la suite un noeud de modélisation à partir du modèle, l'état est repris et utilisé pour initialiser la session interactive. Que l'option soit sélectionnée ou non, le modèle évalue lui-même les nouvelles données de manière identique. Lorsque l'option n'est pas sélectionnée, le modèle peut toujours créer un noeud de création, mais un noeud de création plus générique qui lance une nouvelle session interactive au lieu de reprendre là où l'ancienne session s'est arrêtée. Si vous modifiez les paramètres de noeud, mais procédez à l'exécution avec un état enregistré, les paramètres modifiés sont ignorés au bénéfice des paramètres issus de l'état enregistré.

Remarque : les mesures standard sont les seules mesures conservées avec le modèle. Les autres mesures sont préservées avec l'état interactif. Le modèle généré ne représente pas l'état de tâche d'exploration interactif. Une fois l'Decision List Viewer lancé, il affiche les paramètres initialement définis via l'afficheur.

[Pour plus d'informations, reportez-vous à la section Régénération d'un noeud de modélisation dans le chapitre 3 sur p. 72.](#)

Evaluation du modèle

Une modélisation réussie nécessite d'évaluer avec soin le modèle avant que la mise en oeuvre dans l'environnement de production n'ait lieu. L'Decision List Viewer fournit des mesures statistiques et commerciales permettant d'évaluer l'impact d'un modèle dans le monde réel. Ces mesures incluent des graphiques de gains et une interopérabilité totale avec Excel, ce qui permet de simuler des scénarios coût-bénéfice pour l'évaluation de l'impact du déploiement.

Vous pouvez évaluer votre modèle de plusieurs manières :

- En utilisant des mesures de modèles statistiques et commerciales prédéfinies disponibles dans l'Decision List Viewer (probabilité, effectif).
- En évaluant des mesures importées à partir de Microsoft Excel.
- En visualisant le modèle à l'aide d'un graphique de gains.

Organisation des mesures de modèle

L'Decision List Viewer fournit des options permettant de définir les mesures calculées et affichées sous forme de colonnes. Chaque segment peut inclure les mesures de couverture, d'effectif, de probabilité et d'erreur par défaut représentées sous forme de colonnes. Vous pouvez également créer des mesures qui seront affichées sous forme de colonnes.

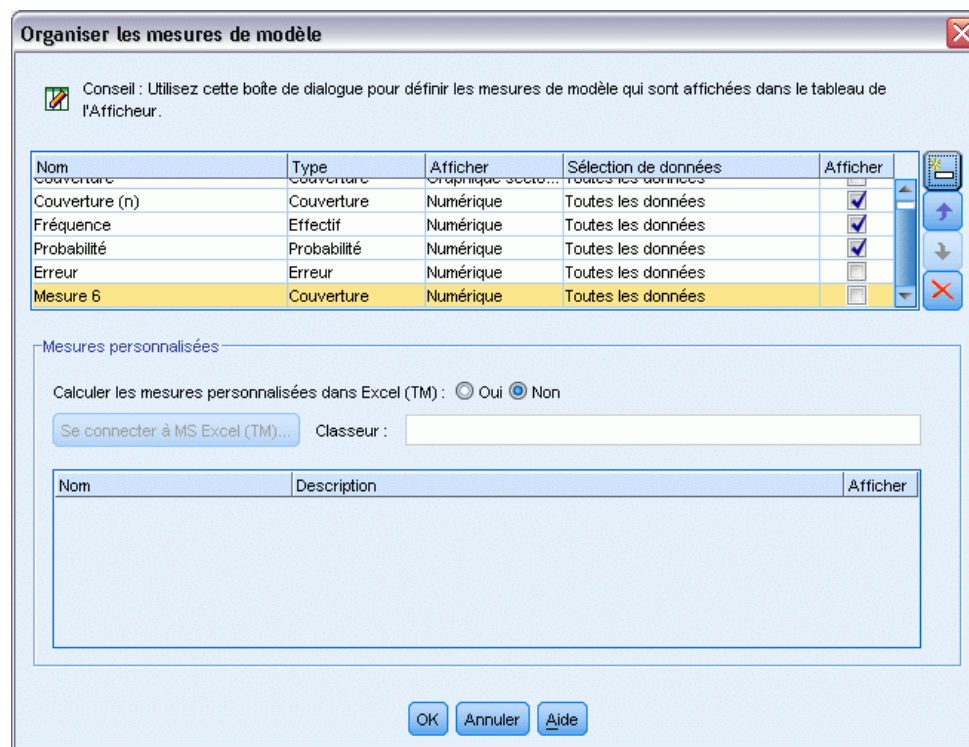
Définition de mesures de modèle

Pour ajouter une mesure à votre modèle, ou définir une mesure existante :

- ▶ Dans le menu Outils, choisissez Organiser les mesures du modèle ou cliquez avec le bouton droit de la souris sur le modèle pour effectuer cette sélection. La boîte de dialogue Organiser les mesures du modèle apparaît.

Figure 9-18

Boîte de dialogue Organiser les mesures de modèle



- ▶ Cliquez sur le bouton Ajouter une nouvelle mesure de modèle (à droite de la colonne Afficher). Une nouvelle mesure apparaît dans la table.

- ▶ Spécifiez un nom de mesure et sélectionnez une option d'affichage, une sélection et un type appropriés. La colonne Afficher indique si la mesure sera ou non affichée pour le modèle de travail. Lorsque vous définissez une mesure existante, sélectionnez une mesure et une sélection appropriées, et indiquez si la mesure sera affichée pour le modèle de travail.
- ▶ Cliquez sur OK pour revenir à l'espace de travail de l'Decision List Viewer. Si la colonne Afficher de la nouvelle mesure a été sélectionnée, la nouvelle mesure est affichée pour le modèle de travail.

Mesures personnalisées dans Excel

Pour plus d'informations, reportez-vous à la section [Evaluation dans Excel sur p. 260](#).

Rafraîchir les mesures

Dans certains cas, il peut être nécessaire de recalculer les mesures de modèle (par exemple, lorsque vous appliquez un modèle existant à un nouvel ensemble de clients).

Pour recalculer (Rafraîchir) les mesures de modèle :

- Dans le menu Edition, sélectionnez Rafraîchir toutes les mesures.

ou

- Appuyez sur la touche F5.

Toutes les mesures sont recalculées et les nouvelles valeurs sont affichées pour le modèle de travail.

Evaluation dans Excel

L'Decision List Viewer peut être intégré à Microsoft Excel, ce qui vous permet d'utiliser vos propres calculs de valeur et formules de profit directement dans le processus de création de modèles pour simuler des scénarios coût-bénéfice. Le lien à Excel vous permet d'exporter des données vers Excel, où vous pouvez les utiliser pour créer des graphiques de présentation et calculer des mesures personnalisées (par exemple, des mesures de profit et de retour sur investissement complexes) et les afficher dans l'Decision List Viewer pendant la création du modèle.

Pour plus d'informations, reportez-vous à la section [Calcul des mesures personnalisées avec Excel dans le chapitre 11 dans *Guide d'applications de IBM SPSS Modeler 15*](#).

Remarque : pour que vous puissiez utiliser une feuille de calcul Excel, l'expert CRM analytique doit définir des informations de configuration pour la synchronisation de l'Decision List Viewer avec Microsoft Excel. La configuration est stockée dans un fichier de feuilles de calcul Excel et indique les informations transférées de l'Decision List Viewer vers Excel et vice versa.

Les étapes suivantes ne sont valides que si MS Excel est installé. Si Excel n'est pas installé, les options de synchronisation de modèles avec Excel ne sont pas affichées.

Pour synchroniser des modèles avec MS Excel :

- ▶ Ouvrez le modèle, lancez une session interactive, puis choisissez Organiser les mesures de modèle dans le menu Outils.
- ▶ Sélectionnez Oui pour l'option Calculer des mesures personnalisées dans Excel. Le champ Classeur s'active, ce qui vous permet de sélectionner un modèle de classeur Excel préconfiguré.
- ▶ Cliquez sur le bouton Connecter à Excel. La boîte de dialogue Ouvrir qui apparaît vous permet de naviguer jusqu'à l'emplacement du modèle préconfiguré sur votre système de fichiers local ou réseau.
- ▶ Sélectionnez le modèle Excel approprié et cliquez sur Ouvrir. Le modèle Excel sélectionné est lancé; utilisez la barre des tâches Windows (ou appuyez sur Alt+Tab) pour revenir à la boîte de dialogue Choisir les entrées de mesures personnalisées.
- ▶ Sélectionnez les mises en correspondance appropriées entre les noms de mesure définis dans le modèle Excel et les noms de mesure du modèle, puis cliquez sur OK.

Une fois le lien établi, Excel démarre avec le modèle Excel préconfiguré affichant les règles du modèle dans la feuille de calcul. Les résultats calculés dans Excel sont affichés en tant que nouvelles colonnes dans l'Decision List Viewer.

Remarque : Les mesures Excel ne sont pas conservées lorsque le modèle est enregistré ; elles ne sont valides que pendant la session active. Cependant, vous pouvez créer des instantanés incluant des mesures Excel. Les mesures Excel enregistrées dans les instantanés ne sont valides qu'à des fins de comparaison historique et ne sont pas réactualisées lorsqu'elles sont rouvertes. [Pour plus d'informations, reportez-vous à la section Onglet Instantanés sur p. 245.](#) Les mesures Excel n'apparaissent dans les instantanés que lorsque vous rétablissez la connexion au modèle Excel.

Configuration de l'intégration à MS Excel

L'intégration entre l'Decision List Viewer et Microsoft Excel s'effectue via l'utilisation d'un modèle de feuille de calcul Excel préconfiguré. Ce modèle est constitué de trois feuilles de calcul :

Mesures de modèle. Affiche les mesures importées de l'Decision List Viewer, les mesures Excel personnalisées et les totaux des calculs (définis dans la feuille de calcul Paramètres).

Paramètres. Fournit les variables pour générer des calculs en fonction des mesures de l'Decision List Viewer importées et des mesures Excel personnalisées.

Configuration : Fournit des options permettant de spécifier les mesures importées à partir de l'Decision List Viewer et de définir les mesures Excel personnalisées.

Avertissement : La structure de la feuille de calcul Configuration est définie de façon rigoureuse. Ne **PAS** éditer les cellules de la zone ombrée verte.

- **Mesures depuis le modèle.** Indique les mesures Decision List Viewer utilisées dans les calculs.
- **Mesures vers le modèle.** Indique la ou les mesures générées par Excel qui vont être renvoyées à l'Decision List Viewer. Les mesures générées par Excel sont affichées en tant que nouvelles colonnes de mesures dans l'Decision List Viewer.

Remarque : Les mesures Excel ne sont pas conservées avec le modèle lorsque vous générez un nouveau modèle ; elles ne sont valides que pendant la session active.

Modification des mesures de modèle

Les exemples suivants expliquent les différentes façons de modifier les mesures de modèle :

- Changer une mesure existante.
- Importer une mesure standard supplémentaire depuis le modèle.
- Exporter une mesure standard supplémentaire vers le modèle.

Changer une mesure existante

- ▶ Ouvrez le modèle, et sélectionnez la feuille de calcul Configuration.
- ▶ Editez le Nom ou la Description en les surlignant et en tapant un nouveau contenu.

Remarque : si l'on souhaite modifier une mesure, par exemple pour inviter l'utilisateur à choisir Probabilité au lieu d'Effectif, il suffit de modifier le nom et la description dans Mesures depuis le modèle : cela s'affiche alors dans le modèle, ce qui permet à l'utilisateur de choisir la mesure appropriée à mapper.

Importez une mesure standard supplémentaire depuis le modèle.

- ▶ Ouvrez le modèle, et sélectionnez la feuille de calcul Configuration.
- ▶ A partir des menus, sélectionnez :
Outils > Verrouillage > Déverrouillez la feuille
- ▶ Sélectionnez la cellule A5, qui est ombrée jaune et contient le mot Fin.
- ▶ A partir des menus, sélectionnez :
Insertion > Rows
- ▶ Entrez le Nom et la Description de la nouvelle mesure. Par exemple, Erreur et Erreur associée à un segment.
- ▶ Dans la cellule C5, entrez la formule =COLUMN('Model Measures'!N3).
- ▶ Dans la cellule C5, entrez la formule =ROW('Model Measures'!N3)+1.

Ces formules permettront d'afficher la nouvelle mesure dans la colonne N de la feuille de calcul Mesures de modèle qui est actuellement vide.

- ▶ A partir des menus, sélectionnez :
Outils > Verrouillage > Verrouillez la feuille
- ▶ Cliquez sur OK.
- ▶ Sur la feuille de calcul Mesures de modèle, assurez-vous que la cellule N3 a Erreur comme titre pour la nouvelle colonne.
- ▶ Sélectionnez toute la colonne N.

- ▶ A partir des menus, sélectionnez :
Format > Cellules
- ▶ Par défaut, toutes les cellules ont un format de nombre Standard. Cliquez sur Pourcentage pour modifier l’affichage des chiffres. En plus de vous aider à vérifier vos chiffres sous Excel, cela vous permet d’utiliser les données d’autres façons, par exemple pour une sortie sous forme de graphique.
- ▶ Cliquez sur OK.
- ▶ Enregistrez la feuille de calcul en tant que modèle Excel 2003, avec un nom unique et l’extension de fichier *.xlt*. Pour retrouver facilement le nouveau modèle, nous vous recommandons de l’enregistrer dans l’emplacement préconfiguré pour les modèles sur votre système de fichiers local ou réseau.

Exportez une mesure personnalisée supplémentaire vers le modèle.

- ▶ Ouvrez le modèle auquel vous avez ajouté la colonne Erreur dans l’exemple précédent ; sélectionnez la feuille de calcul Configuration.
- ▶ A partir des menus, sélectionnez :
Outils > Verrouillage > Déverrouillez la feuille
- ▶ Sélectionnez la cellule A14, qui est ombrée jaune et contient le mot Fin.
- ▶ A partir des menus, sélectionnez :
Insertion > Rows
- ▶ Entrez le Nom et la Description de la nouvelle mesure. Par exemple, Erreur mise à l’échelle et Mise à l’échelle appliquée à une erreur provenant d’Excel.
- ▶ Dans la cellule C14, entrez la formule =COLUMN('Model Measures'!O3).
- ▶ Dans la cellule D14, entrez la formule =ROW('Model Measures'!O3)+1.
Ces formules indiquent que la colonne O contiendra la nouvelle mesure pour le modèle.
- ▶ Sélectionnez la feuille de calcul Paramètres.
- ▶ Dans la cellule A17, entrez la description '- Erreur mise à l’échelle.
- ▶ Dans la cellule B17, entrez le facteur d’échelle 10.
- ▶ Sur la feuille de calcul Mesures de modèle, entrez la description Erreur mise à l’échelle dans la cellule O3 comme titre de la nouvelle colonne.
- ▶ Dans la cellule O4, entrez la formule =N4*Settings!\$B\$17.
- ▶ Cliquez sur le coin de la cellule O4, et faites-le glisser vers la cellule O22 pour recopier la formule dans chaque cellule.
- ▶ A partir des menus, sélectionnez :
Outils > Verrouillage > Verrouillez la feuille
- ▶ Cliquez sur OK.

- Enregistrez la feuille de calcul en tant que modèle Excel 2003, avec un nom unique et l'extension de fichier *.xlt*. Pour retrouver facilement le nouveau modèle, nous vous recommandons de l'enregistrer dans l'emplacement préconfiguré pour les modèles sur votre système de fichiers local ou réseau.

Lors de la connexion à Excel en utilisant ce modèle, la valeur Erreur est disponible comme nouvelle mesure personnalisée.

Visualisation de modèles

La meilleure manière de comprendre l'impact d'un modèle est de le visualiser. Un graphique de gains permet une compréhension précieuse et au jour le jour des activités et du bénéfice technique de votre modèle via l'examen de l'effet de plusieurs alternatives en temps réel. La section [Diagramme des gains](#) présente l'avantage d'un modèle par rapport à une prise de décision aléatoire et permet la comparaison directe de plusieurs graphiques lorsqu'il existe des modèles alternatifs.

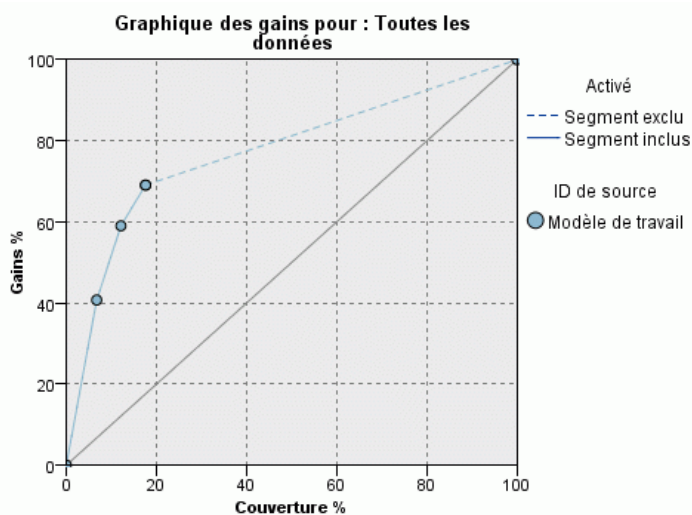
Diagramme des gains

Le graphique de gains représente les valeurs affichées dans la colonne *Gain (%)* du tableau. Les gains sont définis en tant que proportion des correspondances de chaque incrément par rapport au nombre total de correspondances de l'arbre via l'équation suivante :

(correspondances de l'incrément / nombre total de correspondances) x 100 %

Les graphiques de gains permettent d'illustrer efficacement dans quelle mesure vous devez élargir vos perspectives pour obtenir un pourcentage précis de correspondances de l'arbre. La diagonale représente la réponse attendue pour l'échantillon complet si vous n'utilisiez pas le modèle. Dans ce cas, le taux de réponse serait constant, car une personne est susceptible de répondre comme une autre. Pour doubler ce nombre, vous devriez interroger deux fois plus de personnes. La courbe fournit la valeur d'amélioration possible du taux de réponse en n'incluant que les valeurs comprises dans les centiles plus élevés basés sur les gains. Par exemple, si les premiers 50 % sont ajoutés, vous pouvez obtenir plus de 70 % des réponses positives. Plus la courbe s'accroît, plus les gains sont élevés.

Figure 9-19
Onglet Gains



Pour afficher un graphique de gains :

- Ouvrez un flux contenant un noeud Liste de décision et lancez une session interactive à partir du noeud.
- Cliquez sur l'onglet Gains. En fonction des partitions spécifiées, un ou deux graphiques peuvent apparaître (par exemple, deux graphiques si les partitions d'apprentissage et de test sont définies pour les mesures de modèle).

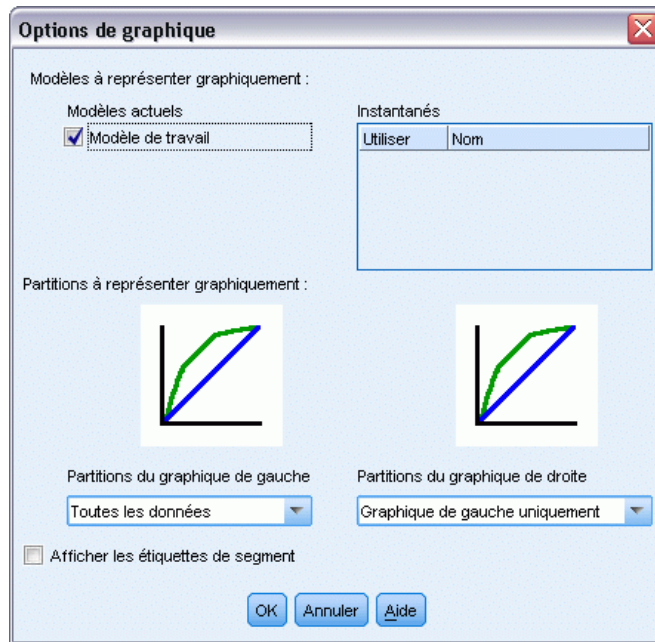
Par défaut, les graphiques s'affichent en tant que segments. Vous pouvez passer à un affichage de quantiles en sélectionnant Quantiles, puis en sélectionnant la méthode de quantile appropriée dans le menu déroulant.

Remarque : Reportez-vous à [Modification des visualisations](#) pour obtenir des informations sur l'utilisation des graphiques.

Options de diagramme

La fonction Options de graphique fournit des options permettant de sélectionner les modèles et instantanés représentés graphiquement, ainsi que les partitions tracées. Elle permet également d'indiquer si les étiquettes de segment sont affichées.

Figure 9-20
Boîte de dialogue Options de graphique



Modèles à représenter graphiquement

Modèles actuels. Permet de sélectionner les modèles à représenter par graphique. Vous pouvez sélectionner le modèle de travail ou n'importe quel modèle d'instantané créé.

Partitions à représenter graphiquement

Partitions du graphique de gauche. La liste déroulante fournit des options permettant d'afficher toutes les partitions définies ou toutes les données.

Partitions du graphique de droite. La liste déroulante fournit des options permettant d'afficher toutes les partitions définies, toutes les données ou uniquement le graphique de gauche. Lorsque l'option Représenter uniquement les éléments de gauche est sélectionnée, seul le graphique de gauche est affiché.

Afficher les étiquettes de segment. Lorsque cette option est activée, chaque étiquette de segment est affichée dans les graphiques.

Modèles statistiques

Les modèles statistiques utilisent des équations mathématiques pour coder les informations extraites à partir des données. Dans certains cas, les techniques de modélisation statistique peuvent fournir des modèles adéquats très rapidement. Même dans des situations où des techniques d'apprentissage automatique (telles que les réseaux de neurones) peuvent donner de meilleurs résultats, vous pouvez utiliser des modèles statistiques en tant que modèles prédictifs de référence qui permettent d'évaluer les performances de techniques plus avancées.

Les noeuds de modélisation statistique suivants sont disponibles.



Les modèles de régression linéaire prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites. [Pour plus d'informations, reportez-vous à la section Modèles linéaires sur p. 269.](#)



La régression logistique est une technique statistique de classification des enregistrements sur la base des valeurs des champs d'entrée. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est similaire à la régression linéaire. [Pour plus d'informations, reportez-vous à la section Noeud Logistique sur p. 288.](#)



Le noeud ACP/Analyse factorielle propose des techniques de factorisation puissantes qui vous permettent de réduire la complexité de vos données. L'analyse en composantes principales (ACP) recherche les combinaisons linéaires des champs d'entrée qui permettent de capturer au mieux la variance dans l'ensemble de champs, où les composantes sont orthogonales (perpendiculaires) les unes par rapport aux autres. L'analyse factorielle a pour but d'identifier les facteurs sous-jacents qui expliquent la tendance des corrélations dans un ensemble de champs observés. Quelle que soit l'approche choisie, le but consiste à trouver un nombre limité de champs dérivés récapitulant les informations contenues dans l'ensemble de champs d'origine. [Pour plus d'informations, reportez-vous à la section Noeud ACP/Analyse factorielle sur p. 310.](#)



L'analyse discriminante crée des hypothèses plus strictes que la régression logistique mais peut constituer une alternative ou un complément précieux à une analyse de régression logistique lorsque ces hypothèses sont réunies. [Pour plus d'informations, reportez-vous à la section Noeud discriminant sur p. 318.](#)



La procédure Modèles linéaires généralisés développe le modèle linéaire général de sorte que la variable dépendante soit linéairement reliée aux facteurs et covariables via une fonction de lien précise. En outre, le modèle permet à la variable dépendante de suivre une distribution non normale. Il couvre les fonctionnalités d'un grand nombre de modèles statistiques, notamment le modèle de régression linéaire, le modèle de régression logistique, le modèle log-linéaire pour les données d'effectif et le modèle de survie avec censure par intervalle. [Pour plus d'informations, reportez-vous à la section Noeud Modèles linéaires généralisés sur p. 327.](#)



Un modèle mixte linéaire généralisé (MMLG) élargit le modèle linéaire de sorte que la cible puisse avoir une distribution non normale, qu'elle soit liée linéairement aux facteurs et covariables via une fonction de lien spécifiée, et que les observations puissent être corrélées. Les modèles mixtes linéaires généralisés couvrent une large variété de modèles, depuis les modèles de régression linéaire simple aux modèles multi-niveaux complexes destinés aux données longitudinales non normales. [Pour plus d'informations, reportez-vous à la section Noeud MMLG sur p. 342.](#)



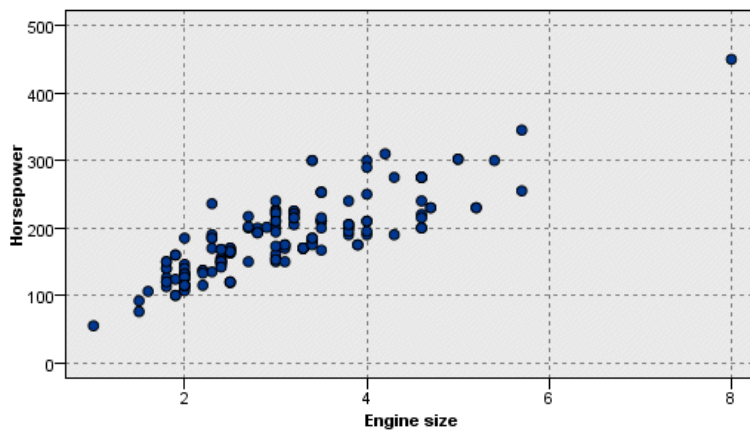
Le noeud de régression de Cox vous permet de créer un modèle de survie pour les données de durée jusqu'à l'événement en présence d'enregistrements censurés. Ce modèle produit une fonction de survie qui prédit la probabilité que l'événement en question se soit produit à un moment (t) pour des valeurs données des variables d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud de Cox sur p. 371.](#)

Noeud linéaire

La régression linéaire est une technique statistique ordinaire 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.

Figure 10-1

Diagramme de régression linéaire simple



Conditions requises. Seuls les champs numériques sont autorisés dans un modèle de régression linéaire. Vous devez avoir exactement un champ cible (avec le rôle défini sur *Cible*) et une ou plusieurs valeurs prédites (avec le rôle défini sur *Entrée*). Les champs avec un rôle de type *Les deux* ou *Aucun* sont ignorés, tout comme les champs non numériques. (Si nécessaire, les champs non numériques peuvent être recodés grâce à un noeud Calculer. [Pour plus d'informations, reportez-vous à la section Recodage des valeurs à l'aide du noeud Calculer dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#))

Puissance. Les modèles de régression linéaire sont relativement simples et produisent une formule mathématique de génération de prévisions pouvant facilement être interprétée. Dans la mesure où la régression linéaire est une technique statistique ancienne, les propriétés des modèles qu'elle génère sont bien connues. D'autre part, leur apprentissage est très rapide. Le noeud

linéaire contient des méthodes de sélection automatique des champs, ce qui permet d'éliminer de l'équation les champs d'entrée non pertinents.

Remarque : lorsque le champ cible est un champ catégoriel plutôt qu'un intervalle continu, par exemple *Oui/Non* ou *Attrition/Absence d'attrition*, la régression logistique est une alternative. La régression logistique offre également une prise en charge des entrées non numériques, supprimant ainsi la nécessité de recoder ces champs. [Pour plus d'informations, reportez-vous à la section Noeud Logistique sur p. 288.](#)

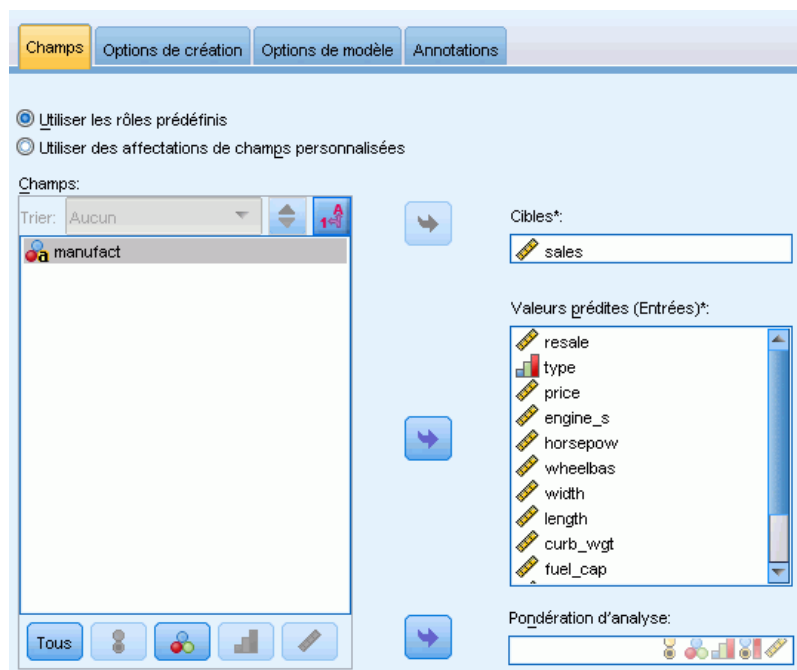
Modèles linéaires

Les modèles linéaires prédisent une cible continue en fonction de relations linéaires entre la cible et une ou plusieurs variables prédites.

Les modèles linéaires sont relativement simples et produisent une formule mathématique pouvant facilement être évaluée. Les propriétés de ces modèles sont bien comprises et peuvent généralement être créées très rapidement par rapport à d'autres types de modèles (tels que les réseaux neuronaux ou les arbres de décision) sur le même ensemble de données.

Exemple : Une compagnie d'assurances disposant de ressources restreintes pour enquêter sur les demandes de remboursement des propriétaires de biens immobiliers, souhaite construire un modèle pour estimer les coûts des réclamations. En déployant ce modèle vers les centres de service, les représentants ont la possibilité d'entrer les informations relatives à la réclamation pendant leur conversation avec les clients et d'obtenir immédiatement une estimation du coût sur la base de réclamations antérieures.

Figure 10-2
Onglet Champs



Exigences concernant les champs. Il doit y avoir une cible et au moins une entrée. Par défaut, les champs avec des rôles prédéfinis sur Les deux ou Aucun ne sont pas utilisés. La cible doit être continue (échelle). Il n'existe pas de restrictions des niveaux de mesure sur les variables indépendantes (d'entrée) ; les champs catégoriels (booléens, nominaux, et ordinaux) sont utilisés comme facteurs dans le modèle et les champs continus sont utilisés comme covariables. [Pour plus d'informations, reportez-vous à la section Options de champs des noeuds de modélisation dans le chapitre 3 sur p. 39.](#)

Objectifs

Figure 10-4
Paramètres des Objectifs

Sélectionnez un élément :

Objectifs

Bases

Choix du modèle

Ensembles

Options avancées

Que souhaitez-vous faire ?

Créer un nouveau modèle Poursuivre le modèle d'apprentissage existant

Quel est votre objectif principal ?

Créer un modèle standard

Améliorer la précision du modèle (amélioration)

Améliorer la stabilité du modèle (bagging)

Créer un modèle pour des ensembles de données volumineux (nécessite un serveur)

Description

Crée un modèle unique standard expliquant les relations entre les champs. Les modèles standard sont plus faciles à interpréter et peuvent être plus rapidement évalués que les ensembles de données volumineux, améliorés ou agrégés.

Que souhaitez-vous faire ?

- **Créer un nouveau modèle.** Créer un modèle entièrement nouveau. Il s'agit de l'opération ordinaire du noeud.
- **Poursuivre l'apprentissage d'un modèle existant.** L'apprentissage se poursuit sur le dernier modèle généré par le noeud. Cela permet de mettre à jour ou d'actualiser un modèle existant sans avoir à accéder aux données d'origine et peut permettre d'obtenir des performances nettement plus rapides car seuls les enregistrements nouveaux ou mis à jour sont acheminés dans le flux. Les détails relatifs au modèle précédent sont stockés avec le noeud de modélisation, ce qui permet d'utiliser cette option même si le nugget de modèle précédent n'est plus disponible dans le flux ou dans la palette de modèles.

Remarque : Lorsque cette option est activée, toutes les autres commandes des onglets Champ et Options de création sont désactivés.

Quel est votre objectif principal ?

- **Créer un modèle standard.** La méthode crée un modèle unique afin de prédire la cible à l'aide des variables indépendantes. En général, les modèles standard sont plus faciles à interpréter et peuvent être plus rapides à évaluer que des ensembles de données améliorés, agrégés ou volumineux.
- **Améliorer la précision du modèle (amélioration).** La méthode crée un modèle d'ensemble à l'aide d'une amélioration, qui génère une séquence de modèles pour obtenir des prédictions plus précises. La création et l'évaluation des ensembles peut prendre davantage de temps qu'un modèle standard.

Le boosting produit une succession de « modèles de composant », chacun généré à partir de l'ensemble de données entier. Avant la création de chaque modèle de composant successif, les enregistrements sont pondérés en fonction des résidus du modèle de composant précédent. Les observations possédant des résidus élevés reçoivent des pondérations d'analyse relativement supérieure de manière à ce que le modèle de composant suivant donne une meilleure prédiction de ces enregistrements. Ensemble ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.

- **Améliorer la stabilité du modèle (agrégation).** La méthode crée un modèle d'ensemble à l'aide du bagging (agrégation de bootstrap), qui génère plusieurs modèles pour obtenir des prédictions plus fiables. La création et l'évaluation des ensembles peut prendre davantage de temps qu'un modèle standard.

L'agrégation de bootstrap (bagging) produit des doubles de l'ensemble de données d'apprentissage en réalisant un échantillonnage avec remplacement à partir de l'ensemble de données d'origine. Ceci permet de créer des échantillons de bootstrap de taille égale à l'ensemble de données d'origine. Ensuite, un « modèle de composant » est créé à partir de chaque double. Ensemble ces modèles de composant forment un modèle d'ensemble. Le modèle d'ensemble évalue les nouveaux enregistrements à l'aide d'une règle de combinaison ; les règles disponibles dépendent du niveau de mesure de la cible.

- **Créer un modèle pour de très grands ensembles de données (requiert IBM® SPSS® Modeler Server).** La méthode crée un modèle d'ensemble en divisant l'ensemble de données en des blocs de données distincts. Sélectionnez cette option si votre ensemble de données est trop volumineux pour créer l'un des modèles ci-dessus, ou pour créer un modèle incrémentiel. Cette option peut prendre moins de temps à créer, mais davantage à évaluer qu'un modèle standard. Cette option nécessite une connexion SPSS Modeler Server.

Voir [Ensembles](#) sur p. 275 pour les paramètres liés au boosting, au bagging et aux très grands ensembles de données.

Bases

Figure 10-5
Paramètres de base

The screenshot shows a software interface with four tabs: 'Champs', 'Options de création', 'Options de modèle', and 'Annotations'. The 'Options de création' tab is active. On the left, a sidebar lists 'Objectifs', 'Bases', 'Choix du modèle', 'Ensembles', and 'Options avancées', with 'Bases' highlighted. The main content area has the heading 'Sélectionnez un élément :'. It contains a checked checkbox labeled 'Préparer automatiquement les données' with the description: 'Prépare les données dans le but d'améliorer la puissance de prévision de cette procédure. Non disponible pour des ensembles de données volumineux.' Below this is a spinner control for 'Niveau de confiance (%)' set to '95,0'.

Préparer automatiquement les données. Cette option permet à la procédure de transformer en interne la cible et les variables indépendantes afin de maximiser la puissance de prédiction du modèle. Toute modification est enregistrée avec le modèle et appliquée aux nouvelles données pour le scoring. Les versions originales de champs transformés sont exclues du modèle. Par défaut, les préparations automatiques de données suivantes sont réalisées.

- **Gestion de la date et de l'heure.** Chaque variable indépendante de date est transformée en une nouvelle variable indépendante continue qui contient la durée écoulée depuis une date de référence (01/01/1970). Chaque variable indépendante d'heure est transformée en une nouvelle variable indépendante continue qui contient la durée écoulée depuis une heure de référence (00:00:00).
- **Régler le niveau de mesure.** Les variables indépendantes continues ayant moins de 5 valeurs distinctes sont reconverties en variables indépendantes ordinales. Les variables indépendantes ordinales ayant plus de 10 valeurs distinctes sont reconverties en variables indépendantes continues.
- **Gestion des valeurs éloignées.** Les valeurs de variables indépendantes continues qui se trouvent au-delà d'une valeur de césure (écart-type de 3 par rapport à la moyenne) sont définies sur la valeur de césure.
- **Gestion des valeurs manquantes.** Les valeurs manquantes de variables indépendantes nominales sont remplacées par le mode de la partition d'apprentissage. Les valeurs manquantes de variables indépendantes ordinales sont remplacées par la médiane de la

partition d'apprentissage. Les valeurs manquantes de variables indépendantes continues sont remplacées par la moyenne de la partition d'apprentissage.

- **Fusion supervisée.** Ceci crée un modèle plus petit en réduisant le nombre de champs à traiter en association avec la cible. Les modalités similaires sont identifiées en fonction de la relation entre l'entrée et la cible. Les modalités ne différant pas de manière significative (c'est-à-dire ayant une valeur p supérieure à 0,1), sont fusionnées. Si toutes les catégories sont fusionnées en une seule, les versions d'origine et dérivées du champ sont exclues du modèle car elles n'ont pas de valeur de variable indépendante.

Niveau de confiance. Il s'agit du niveau de confiance utilisé pour calculer les estimations d'intervalle des coefficients de modèle dans la vue [Coefficients](#). Définissez une valeur supérieure à 0 et inférieure à 100. La valeur par défaut est 95.

Choix du modèle

Figure 10-6
Paramètres du choix du modèle

Sélectionnez un élément :

Objectifs

Bases

Choix du modèle

Ensembles

Options avancées

Méthode de choix du modèle: Pas à pas ascendante

Choix de la méthode Pas à pas ascendante

Critères d'entrée/suppression: Critère d'informations (AICC)

Inclure les effets avec des valeurs p inférieures à: 0,05

Supprimer les effets avec des valeurs p supérieures à: 0,1

Personnaliser un maximum d'effets dans le modèle final

Nombre maximal d'effets:

Personnaliser un maximum d'étapes

Nombre d'étapes maximal:

Sélection des meilleurs sous-ensembles

Critères d'entrée/suppression: Critère d'informations (AICC)

Méthodes de choix du modèle. Choisissez l'une des méthodes de sélection du modèle (détails ci-dessous) ou l'option Inclure toutes les variables indépendantes, qui entre simplement toutes les variables indépendantes disponibles en tant que termes des effets principaux du modèle. Le modèle Pas à pas ascendant est utilisé par défaut.

Choix de la méthode Pas à pas ascendante. Elle commence sans effet dans le modèle et ajoute et supprime des effets une étape à la fois jusqu'à ce qu'aucune autre ne puisse être ajoutée ou supprimée en fonction des critères pas à pas.

- **Critères d'entrée/suppression.** Il s'agit des statistiques utilisées pour savoir si un effet doit être ajouté ou supprimé du modèle. Critère d'information (AICC) est basé sur la vraisemblance du modèle fourni à l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Statistiques F est basé sur un test statistique de l'amélioration dans l'erreur d'un modèle. R-deux ajusté est basé sur l'adéquation de l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Le Critère de prévention du surajustement (ASE) est basé sur l'adéquation (erreur de la moyenne des carrés) de l'ensemble de prévention de surajustement. L'ensemble de prévention de surajustement est un sous-échantillon aléatoire contenant approximativement 30 % des données de l'ensemble de données original. Il n'est pas utilisé pour former le modèle.

Si un autre critère que Statistiques F est sélectionné, à chaque étape l'effet qui correspond à l'accroissement positif le plus important dans le critère est ajouté au modèle. Tous les effets du modèle qui correspondent à une diminution du critère sont supprimés.

Si Statistiques F est sélectionné en tant que critère, à chaque étape l'effet ayant la plus petite valeur *p* inférieure au seuil spécifié, Inclure les effets avec des valeurs *p* inférieures à, est ajouté au modèle. La valeur par défaut est 0.05. Tous les effets du modèle ayant une valeur *p* supérieure au seuil spécifié, Supprimer les effets ayant des valeurs *p* supérieures à, sont supprimés. La valeur par défaut est 0.10.

- **Personnaliser le nombre maximum d'effets dans le modèle final.** Par défaut, tous les effets disponibles peuvent être entrés dans le modèle. Si l'algorithme pas à pas se termine à une étape avec le nombre spécifié d'effets, l'algorithme s'arrête à l'ensemble d'effets en cours.
- **Personnaliser le nombre maximal d'étapes.** L'algorithme pas à pas s'arrête après un certain nombre d'étapes. Par défaut, il s'agit de 3 fois le nombre d'effets disponibles. Vous pouvez également spécifier un nombre entier positif maximum d'étapes.

Sélection des meilleurs sous-ensembles. Ceci permet de vérifier "tous les modèles possibles" ou au moins un sous-ensemble plus important des modèles possibles qu'en pas à pas ascendant, pour choisir le meilleur en fonction du critère des meilleurs sous-ensembles. Critère d'information (AICC) est basé sur la vraisemblance du modèle fourni à l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. R-deux ajusté est basé sur l'adéquation de l'ensemble d'apprentissage, et est ajusté afin de pénaliser des modèles trop complexes. Le Critère de prévention du surajustement (ASE) est basé sur l'adéquation (erreur de la moyenne des carrés) de l'ensemble de prévention de surajustement. L'ensemble de prévention de surajustement est un sous-échantillon aléatoire contenant approximativement 30 % des données de l'ensemble de données original. Il n'est pas utilisé pour former le modèle.

Le modèle ayant la plus grande valeur de critère est sélectionné comme meilleur modèle.

Remarque : La sélection des meilleurs sous-ensembles demande plus de ressources de calcul que la sélection pas à pas ascendante. Lorsque la sélection des meilleurs sous-ensemble est effectuée en conjonction avec le boosting, le bagging ou le traitement d'ensembles très volumineux, elle peut être plus longue que la création d'un modèle standard à l'aide de la sélection pas à pas ascendante.

Ensembles

Figure 10-7
Paramètres des ensembles

Sélectionnez un élément :

- Objectifs
- Bases
- Choix du modèle
- Ensembles**
- Options avancées

Ces paramètres déterminent le comportement de création des ensembles lorsque l'amélioration, l'agrégation ou de très grands ensembles de données sont requis dans les objectifs.
Les options ne s'appliquant pas sont ignorées.

La méthode du Bagging et les très grands ensembles de données

Règle de combinaison par défaut pour les cibles continues: Moyenne

Amélioration et agrégation

Nombre des modèles de composant pour l'amélioration et/ou l'agrégation: 10

Ces paramètres déterminent le comportement d'assemblage qui se produit lors du boosting, du bagging ou lorsque que des ensembles volumineux de données sont requis dans les objectifs. Les options qui ne s'appliquent pas à l'objectif sélectionné sont ignorées.

Bagging et très grands ensembles de données. Lors de l'évaluation d'un ensemble, il s'agit de la règle utilisée pour combiner les valeurs prédites à partir des modèles de base pour calculer la valeur de score d'un ensemble.

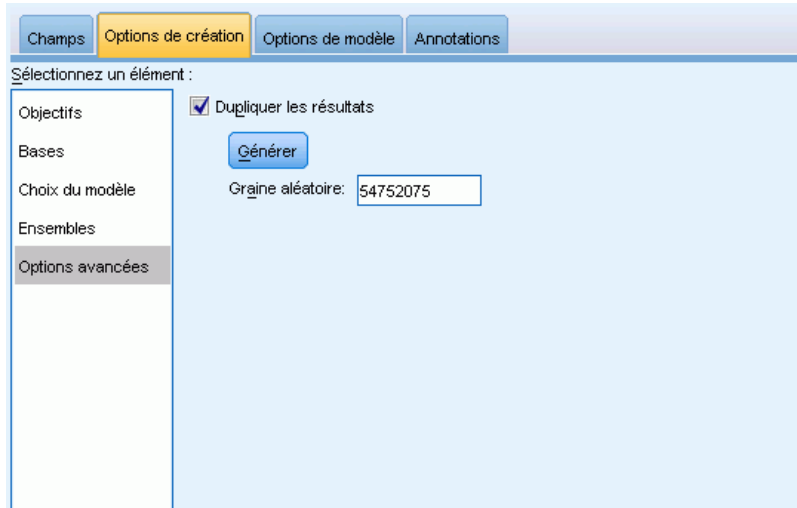
- **Règles de combinaison par défaut pour les cibles continues.** Des valeurs prédites d'ensemble pour des cibles continues peuvent être combinées à l'aide de la moyenne ou de la médiane des valeurs prédites à partir des modèles de base.

Veillez noter que lorsque l'objectif consiste à améliorer la précision du modèle, les sélections de règles de combinaisons sont ignorées. Le boosting utilise toujours un vote majoritaire pondéré pour évaluer des cibles catégorielles et une médiane pondérée pour évaluer des cibles continues.

Boosting et bagging. Spécifiez le nombre de modèles de base à créer lorsque l'objectif est d'améliorer la précision ou la stabilité du modèle ; pour le bagging, il s'agit du nombre d'échantillons de bootstrap. Il doit s'agir d'un entier positif.

Avancé

Figure 10-8
Paramètres avancés

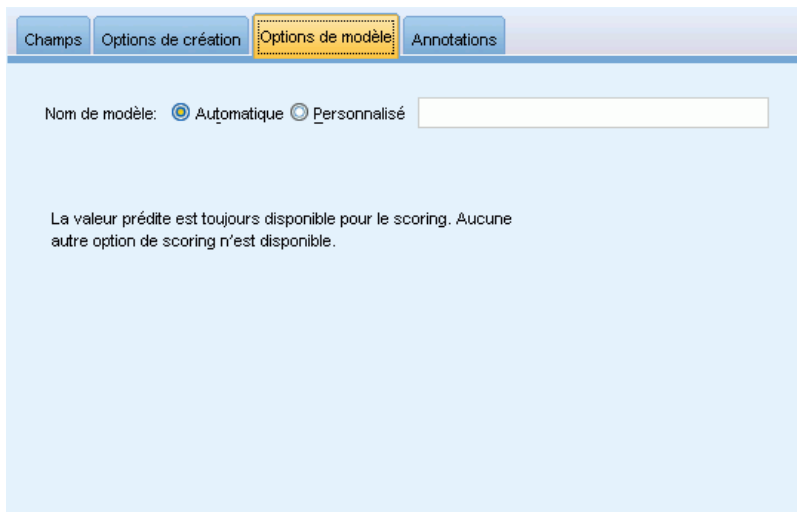


The screenshot shows a software interface with a top navigation bar containing four tabs: 'Champs', 'Options de création', 'Options de modèle', and 'Annotations'. The 'Options de création' tab is currently selected and highlighted in yellow. Below the tabs, there is a section titled 'Sélectionnez un élément :'. On the left side of this section is a vertical list of menu items: 'Objectifs', 'Bases', 'Choix du modèle', 'Ensembles', and 'Options avancées'. The 'Options avancées' item is highlighted with a grey background. To the right of this list, there is a checked checkbox labeled 'Dupliquer les résultats'. Below the checkbox is a blue button labeled 'Générer'. Underneath the button is a text label 'Graine aléatoire:' followed by a text input field containing the value '54752075'.

Dupliquer les résultats. Définir une valeur de départ aléatoire vous permet de dupliquer des analyses. Le générateur de nombres aléatoires est utilisé pour choisir les enregistrements de l'ensemble de prévention de surajustement. Spécifiez un entier ou cliquez sur Générer, ce qui crée un entier pseudo-aléatoire compris entre 1 et 2147483647, inclus. La valeur par défaut est 54752075.

Options de modèle

Figure 10-9
Onglet Options de modèle



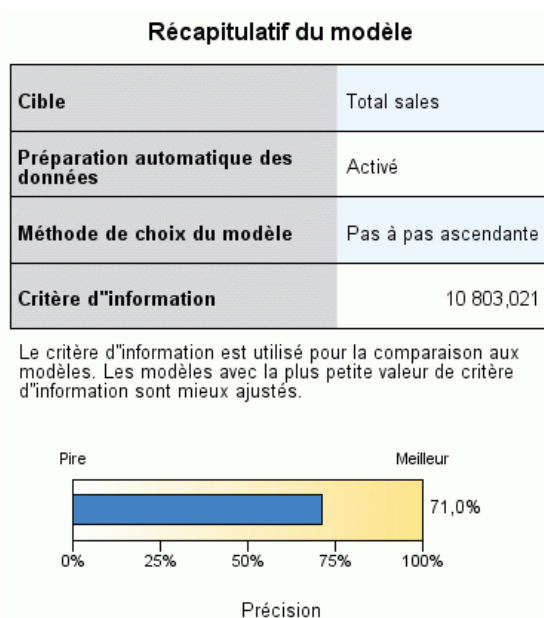
The screenshot shows the same software interface as Figure 10-8, but with the 'Options de modèle' tab selected and highlighted in yellow. Below the tabs, there is a section titled 'Nom de modèle:'. This section contains two radio buttons: 'Automatique' (which is selected) and 'Personnalisé'. To the right of these radio buttons is a text input field. Below this section, there is a paragraph of text: 'La valeur prédite est toujours disponible pour le scoring. Aucune autre option de scoring n'est disponible.'

Nom de modèle. Vous pouvez générer automatiquement le nom du modèle en fonction des champs cible ou spécifier un nom personnalisé. Le nom généré automatiquement est le nom du champ cible.

Veillez noter que la valeur prédite est toujours calculée lorsque le modèle est évalué. Le nom du nouveau champ est le nom du champ cible auquel le préfixe *\$L-* a été ajouté. Par exemple, si un champ cible est nommé *ventes*, le nouveau champ sera alors intitulé *\$L-ventes*.

Récapitulatif de modèle

Figure 10-10
Vue récapitulative du modèle



La vue récapitulative du modèle est un instantané, un récapitulatif accessible d'un coup d'œil du modèle et de son ajustement.

Tableau. Le tableau identifie certains paramètres de modèle de niveau supérieur, notamment :

- Le nom de la cible spécifié sur l'onglet [Champs](#),
- si la préparation automatique des données a été exécutée tel qu'indiqué sur l'onglet [Options de base](#),
- la méthode de sélection de modèle et le critère de sélection spécifiés dans les paramètres de [sélection de modèle](#). La valeur du critère de sélection du modèle final est également affichée et est présentée en plus petit, pour une meilleure mise en forme.

Diagramme. Le diagramme affiche la précision du modèle final, qui est présenté en plus grand, disposant d'un meilleur format. La valeur est de $100 \times R^2$ ajusté pour le modèle final.

Préparation automatique des données

Figure 10-11
Vue Préparation automatique des données

Préparation automatique des données		
Cible : Total sales		
Champ	Rôle	Actions entreprises
Age category	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible
Primary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible
Promotion	Valeur prédite	Modifier le niveau de mesure de continu en ordinal
Secondary keyword set	Valeur prédite	Fusionner les catégories pour maximiser l'association avec la cible

Si le nom du champ d'origine est X, le nom du champ transformé est X transformé. Le champ d'origine est exclu de l'analyse et le champ transformé est inclus à sa place.

Cette vue affiche des informations concernant les champs qui ont été exclus et la façon dont les champs transformés ont été dérivés dans l'étape de préparation automatique des données (ADP). Pour chaque champ transformé ou exclu, le tableau répertorie le nom du champ, son rôle au sein de l'analyse et l'action entreprise par l'étape ADP. Les champs sont triés selon l'ordre alphabétique croissant des noms de champ. Les actions possibles pour chaque champ comprennent :

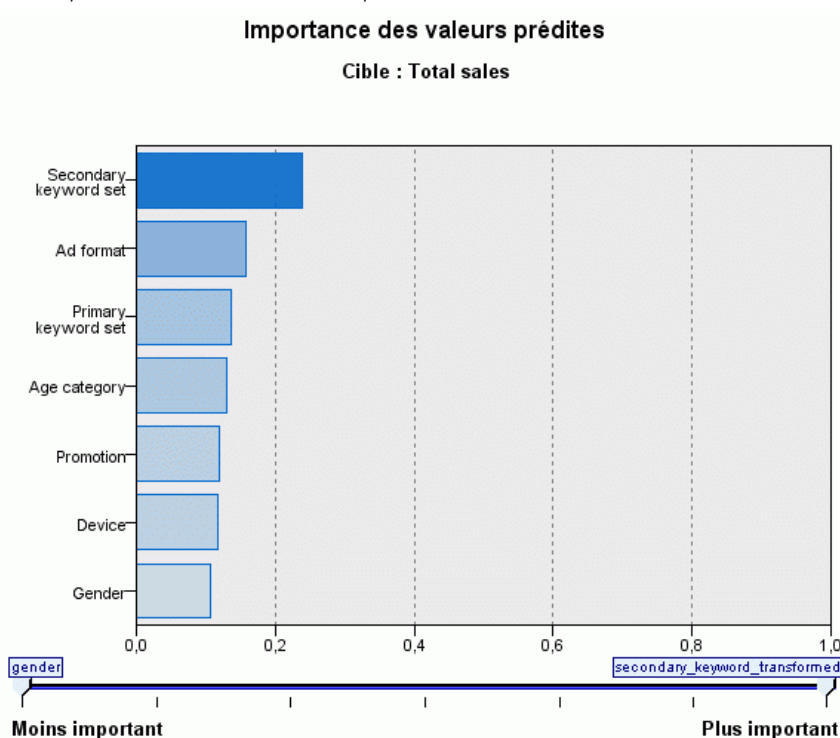
- Calculer la durée : mois calcule le temps écoulé en mois à partir des valeurs d'un champ de dates et de la date système en cours.
- Calculer la durée : heures calcule le temps écoulé en heures à partir des valeurs d'un champ contenant des heures et de l'heure système en cours.
- Modifier le niveau de mesure de continu en ordinal convertit les champs continus de moins de cinq valeurs uniques en champs ordinaux.
- Modifier le niveau de mesure d'ordinal en continu convertit les champs ordinaux de plus de dix valeurs uniques en champs continus.
- Supprimer les valeurs éloignées définit les valeurs de variables indépendantes continues qui se trouvent au-delà d'une valeur de césure (écart-type de 3 par rapport à la moyenne) sur la valeur de césure.
- Remplacer les valeurs manquantes remplace les valeurs manquantes des champs nominaux par le mode, des champs ordinaux par la médiane et des champs continus par la moyenne.

- Fusionner les catégories pour optimiser l'association à la cible identifie les modalités similaires des variables indépendantes en fonction de la relation entre l'entrée et la cible. Les modalités ne différant pas de manière significative (c'est-à-dire ayant une valeur p supérieure à 0,05), sont fusionnées.
- Exclure les variables indépendantes constantes / après le traitement des valeurs éloignées / après la fusion des catégories supprime les variables indépendantes qui comportent une seule valeur, après que les autres actions ADP ont été réalisées.

Importance des variables indépendantes

Figure 10-12

Vue Importance des variables indépendantes



En général, vous préférerez concentrer vos efforts de modélisation sur les champs prédictifs les plus importants et abandonner ou ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des valeurs prédites affichée est 1,0. L'importance des des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

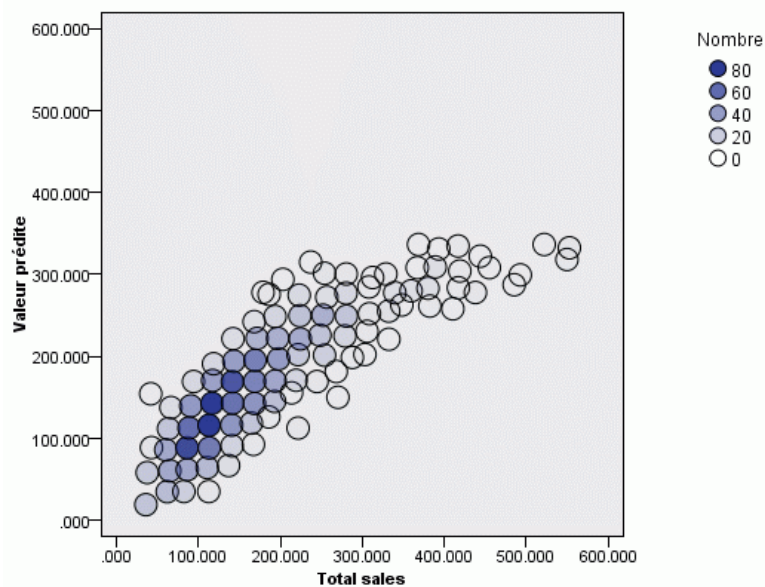
Valeurs prévues en fonction des valeurs observées

Figure 10-13

Vue Valeurs prédites en fonction des valeurs observées

Valeurs prévues en fonction des valeurs observées

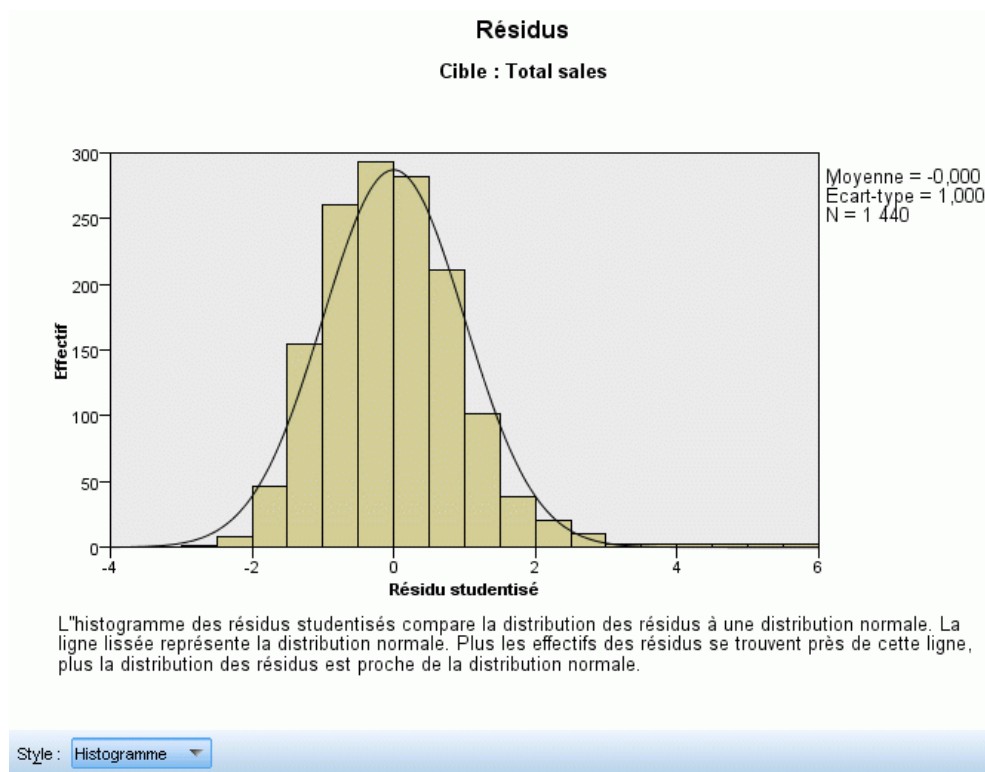
Cible : Total sales



Ceci affiche un diagramme de dispersion mis en intervalles des valeurs prédites sur l'axe vertical par les valeurs observées sur l'axe horizontal. Idéalement, les points devraient se trouver sur une ligne de 45 degrés ; cette vue peut vous indiquer si des enregistrements sont particulièrement mal prédits par le modèle.

Résidus

Figure 10-14
Vue Résidus, style d'histogramme



Ceci affiche un diagramme de diagnostic des résidus du modèle.

Styles de diagramme. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style.

- **Histogramme.** Il s'agit d'un histogramme à intervalles des résidus studentisés avec une superposition de la distribution normale. Les modèles linéaires supposent que les résidus ont une distribution normale, de sorte que l'histogramme doit, dans l'idéal, approcher étroitement la ligne de lissage.
- **Diagramme P-P.** Il s'agit d'un diagramme probabilité-probabilité mis en intervalles qui compare des résidus studentisés à une distribution normale. Si la pente des points représentés est moins forte que la ligne normale, les résidus affichent une plus grande variabilité qu'une distribution normale ; si la pente est plus forte, les résidus affichent une moins grande variabilité qu'une distribution normale. Si les points représentés ont une courbe en S, la distribution des résidus est asymétrique.

Valeurs éloignées

Figure 10-15
Vue Valeurs éloignées

Valeurs éloignées

Cible : Total sales

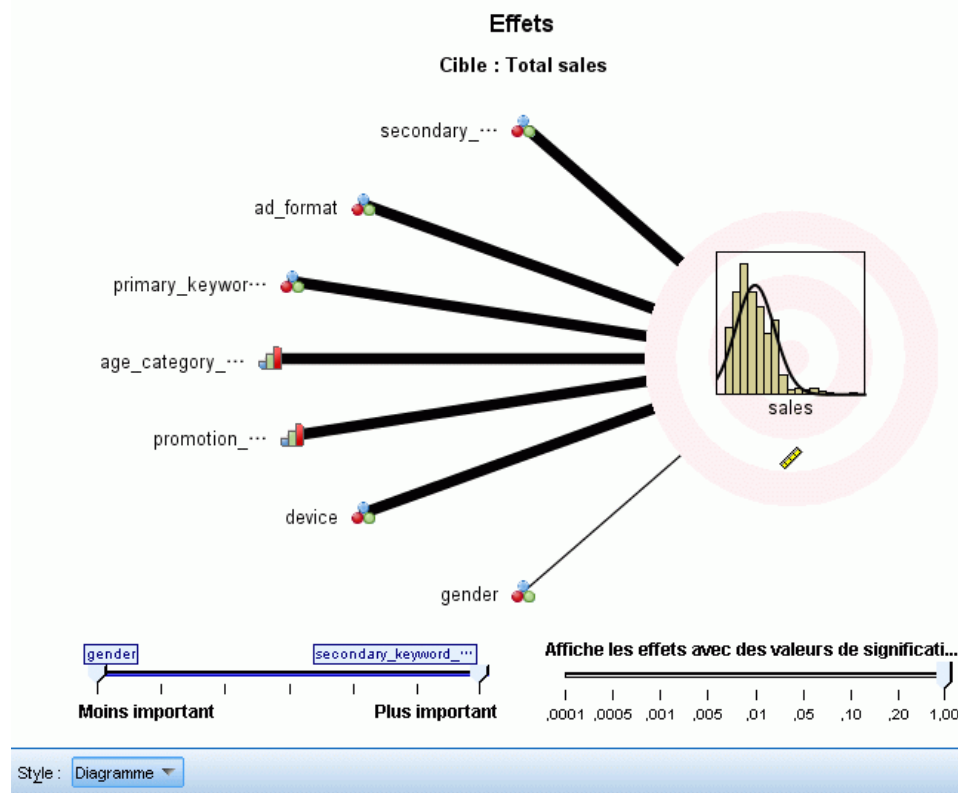
Total sales	Distance de Cook
560.040	0,026
566.440	0,025
548.990	0,018
539.630	0,018
485.430	0,014
543.240	0,014

Ce tableau répertorie les enregistrements qui exercent une influence excessive sur le modèle et affiche l'ID d'enregistrement (s'il est spécifié dans l'onglet Champs), la valeur cible et la distance de Cook. La distance de Cook est une mesure du degré de modification des résidus de tous les enregistrements si un enregistrement donné est exclu des calculs des coefficients de modèle. Une distance de Cook importante signifie que l'exclusion d'un enregistrement modifie de manière importante les coefficients et doit donc être considérée comme ayant une influence.

Les enregistrements ayant une influence doivent être examinés soigneusement afin de déterminer si vous pouvez leur octroyer une pondération inférieure dans l'estimation du modèle, tronquer les valeurs éloignées à un seuil acceptable ou supprimer complètement les enregistrements ayant une influence.

Effets

Figure 10-16
vue Effets, style de diagramme



Cette vue affiche la taille de chaque effet dans le modèle.

Styles. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style.

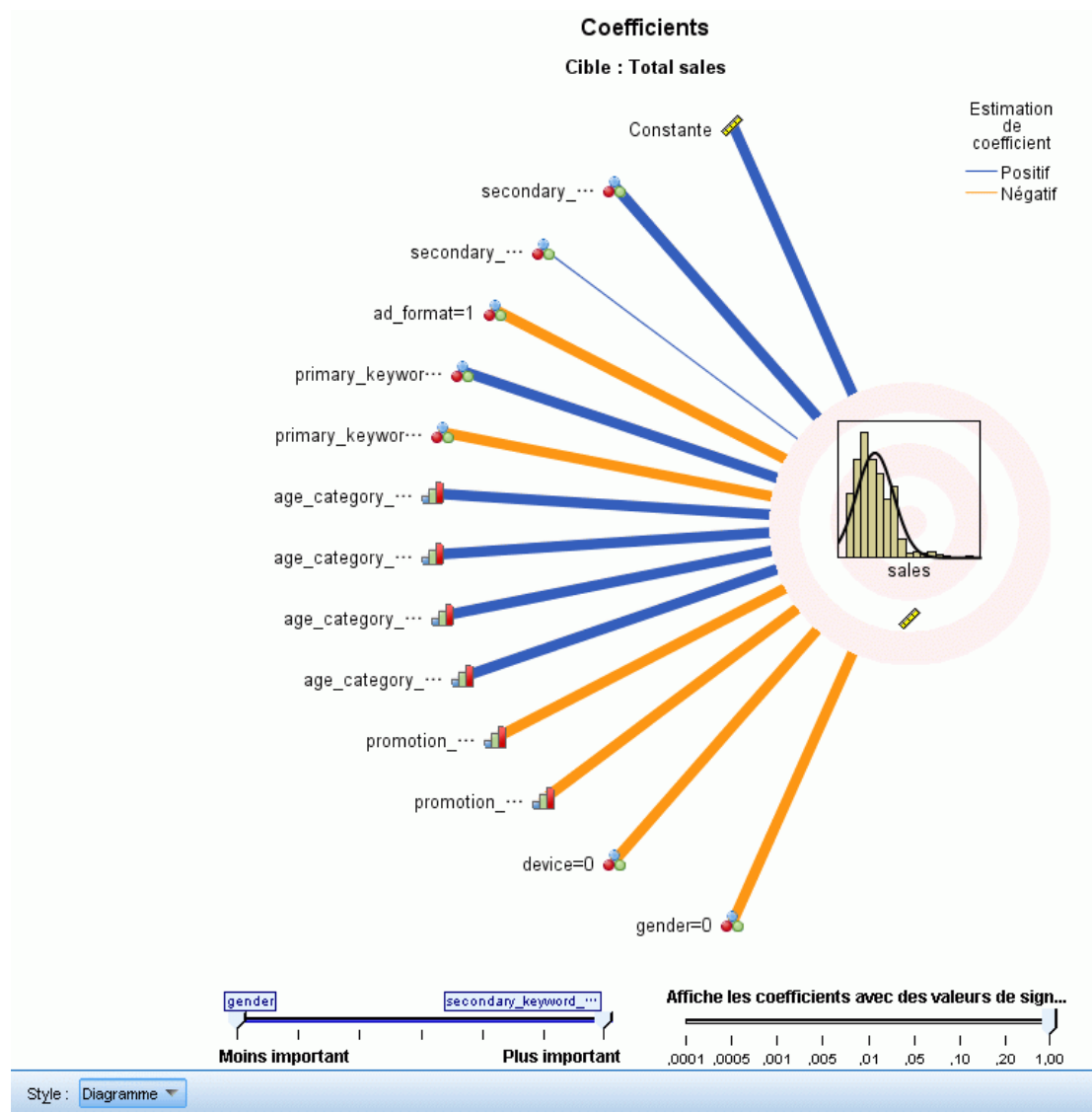
- **Diagramme.** Il s'agit d'un diagramme dans lequel les effets sont triés de haut en bas en diminuant l'importance de la variable indépendante. Les lignes de connexion du diagramme sont pondérées en fonction de la signification de l'effet, une largeur de ligne plus importante correspondant à des effets plus importants (valeurs p plus petites). Lorsque vous placez la souris sur une ligne de connexion, une info-bulle apparaît et affiche la valeur p et l'importance de l'effet. Il s'agit de la valeur par défaut.
- **Tableau.** Il s'agit d'un tableau ANOVA pour le modèle général et les effets de modèle individuels. Il s'agit d'effets individuels triés de haut en bas en diminuant l'importance de la variable indépendante. Remarque : par défaut, le tableau est réduit et n'affiche que les résultats du modèle global. Pour afficher les résultats des effets du modèle individuel, cliquez sur la cellule *Modèle corrigé* dans le tableau.

Importance des variables indépendantes. Il existe un curseur de l'importance des variables indépendantes qui contrôle celles qui sont affichées dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les variables indépendantes les plus importantes. Par défaut, les 10 premiers effets sont affichés.

Signification. Il existe un curseur de signification qui offre des commandes plus avancées sur les effets affichés dans la vue, en plus de celles affichées en fonction de l'importance des variables indépendantes. Les effets ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les effets les plus importants. Par défaut, la valeur est de 1,00, de sorte qu'aucun effet n'est filtré en fonction de la signification.

Coefficients

Figure 10-17
Vue Coefficients, style de diagramme



Cette vue affiche la valeur de chaque coefficient du modèle. Veuillez noter que les facteurs (variables indépendantes catégorielles) sont codés par un indicateur dans le modèle, de sorte que les **effets** comportant des facteurs ont généralement plusieurs **coefficients** associés, un pour chaque catégorie exceptée la catégorie correspondant au paramètre redondant (de référence).

Styles. Il existe différents styles d’affichage accessibles depuis la liste déroulante Style.

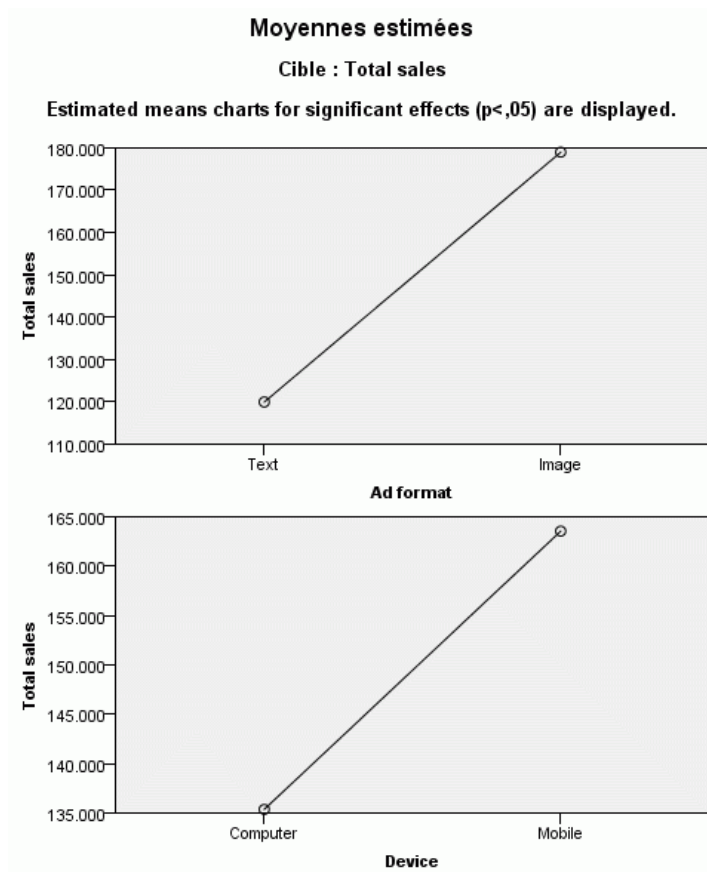
- **Diagramme.** Il s’agit d’un diagramme qui affiche d’abord la constante, puis trie les effets de haut en bas en diminuant l’importance de la variable indépendante. Au sein des effets contenant des facteurs, les coefficients sont triés dans l’ordre croissant de la valeur des données. Les lignes de connexion du diagramme sont coloriées en fonction du signe du coefficient (voir la légende du diagramme) et pondérées en fonction de la signification du coefficient, avec une largeur de ligne plus importante correspondant à des coefficient plus importants (valeurs p plus petites). Lorsque vous placez la souris sur une ligne de connexion, une info-bulle apparaît et affiche la valeur du coefficient, sa valeur p et l’importance de l’effet auquel est associé le paramètre. Il s’agit du style par défaut.
- **Tableau.** Affiche les valeurs, les tests de signification et les intervalles de confiance des coefficients de modèles individuels. Après la constante, les effets sont triés de haut en bas en diminuant l’importance de la variable indépendante. Au sein des effets contenant des facteurs, les coefficients sont triés dans l’ordre croissant de la valeur des données. Remarque : par défaut, le tableau est réduit et n’affiche que le coefficient, la signification et l’importance de chaque paramètre du modèle. Pour afficher l’erreur standard, la statistique t et l’intervalle de confiance, cliquez sur la cellule Coefficient dans le tableau. Lorsque vous placez la souris sur le nom d’un paramètre de modèle dans le tableau, une info-bulle apparaît et affiche le nom du paramètre, l’effet auquel est associé le paramètre et (pour les variables indépendantes catégorielles) les étiquettes de valeur associées au paramètre du modèle. Ceci est particulièrement utile pour afficher les nouvelles catégories créées lorsque la préparation automatique des données fusionne les catégories similaires d’une variable indépendante catégorielle.

Importance des variables indépendantes. Il existe un curseur de l’importance des variables indépendantes qui contrôle celles qui sont affichées dans la vue. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les variables indépendantes les plus importantes. Par défaut, les 10 premiers effets sont affichés.

Signification. Il existe un curseur de signification qui offre des commandes plus avancées sur les coefficients affichés dans la vue, en plus de celle affichée en fonction de l’importance des variables indépendantes. Les coefficients ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les coefficients les plus importants. Par défaut, la valeur est de 1,00, de sorte qu’aucun coefficient n’est filtré en fonction de la signification.

Moyennes estimées

Figure 10-18
Vue Moyennes estimées



Il s'agit de diagrammes affichés pour des variables indépendantes significatives. Le diagramme affiche la valeur estimée par le modèle de la cible sur l'axe vertical pour chaque valeur de la variable indépendante de l'axe horizontal en conservant toutes les autres variables indépendantes. Il offre une visualisation pratique des effets des coefficients de chaque variable prédite sur la cible.

Remarque : si aucune variable indépendante n'est significative, aucune moyenne estimée n'est générée.

Récapitulatif de création de modèle

Figure 10-19

Vue Récapitulatif de création de modèle, algorithme pas à pas ascendant

Récapitulatif de création de modèle

Cible : Total sales

	Etape						
	1	2	3	4	5	6	7
Critère d'information	11 949,413	11 597,758	11 347,000	11 118,878	10 965,287	10 816,338	10 803,021
secondary_keyword_transformed	✓	✓	✓	✓	✓	✓	✓
ad_format		✓	✓	✓	✓	✓	✓
primary_keyword_transformed			✓	✓	✓	✓	✓
Effet age_category_transformed				✓	✓	✓	✓
promotion_transformed					✓	✓	✓
device						✓	✓
gender							✓

La méthode de création de modèle est la méthode ascendante pas à pas avec le critère d'information.
Une coche signifie que l'effet est présent dans le modèle à cette étape.

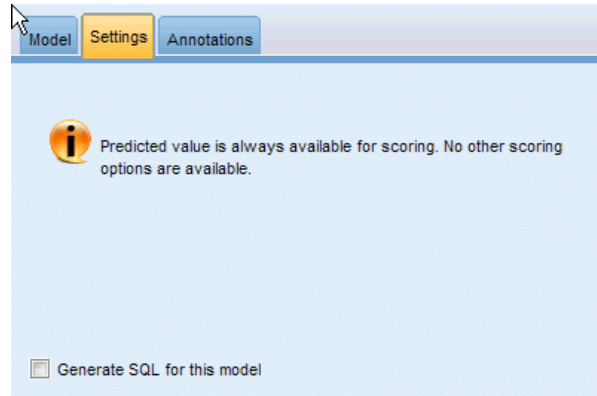
Lorsqu'un algorithme de sélection de modèle différent de Aucun est sélectionné dans les paramètres de sélection de modèles, il propose certains détails concernant le processus de création du modèle.

Pas à pas ascendant. Lorsque l'algorithme de sélection est pas à pas ascendant, le tableau affiche les 10 dernières étapes de l'algorithme pas à pas. Pour chaque étape, la valeur du critère de sélection et les effets du modèle à cette étape sont affichés. Ceci vous offre un aperçu de l'ampleur de la contribution de chaque étape au modèle. Chaque colonne vous permet de trier les lignes afin que vous puissiez voir plus facilement les effets qui se trouvent dans le modèle à chaque étape donnée.

Meilleurs sous-ensembles. Lorsque l'algorithme de sélection est Meilleurs sous-ensembles, le tableau affiche les 10 meilleurs modèles. Pour chaque modèle, la valeur du critère de sélection et les effets du modèle sont affichés. Ceci vous donne un aperçu de la stabilité des meilleurs modèles ; s'ils ont tendance à avoir des effets similaires avec quelques différences, vous pouvez alors avoir une confiance raisonnable dans le "meilleur" modèle ; s'ils ont tendance à avoir des effets très différents, certains des effets peuvent être trop similaires et doivent être associés (ou l'un d'entre eux doit être supprimé). Chaque colonne vous permet de trier les lignes afin que vous puissiez voir plus facilement les effets qui se trouvent dans le modèle à chaque étape donnée.

Paramètres

Figure 10-20
onglet Paramètres



Veillez noter que la valeur prédite est toujours calculée lorsque le modèle est évalué. Le nom du nouveau champ est le nom du champ cible auquel le préfixe *\$L-* a été ajouté. Par exemple, si un champ cible est nommé *ventes*, le nouveau champ sera alors intitulé *\$L-ventes*.

Générer SQL pour ce modèle. Lorsque vous utilisez des données provenant d'une base de données, le code SQL peut être renvoyé à la base de données pour exécution, ce qui assure des performances supérieures pour de nombreuses opérations. [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*.](#)

Evaluer en convertissant en SQL natif. Si cette option est sélectionnée, le SQL est généré pour évaluer le modèle de manière native dans l'application.

Noeud Logistique

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. Excepté le fait qu'elle utilise un champ cible catégoriel et non pas numérique, cette régression est semblable à la régression linéaire. A la fois les modèles binomiaux (pour les cibles avec deux catégories discrètes) et les modèles multinomiaux (pour les cibles avec plus de deux catégories) sont pris en charge.

La régression logistique crée un ensemble d'équations qui associent les valeurs de champ d'entrée aux probabilités rattachées à chacune des catégories de champ de sortie. Une fois le modèle créé, il peut être utilisé pour générer des probabilités sur de nouvelles données. Pour chaque enregistrement, une probabilité d'appartenance est calculée pour chaque catégorie de sortie. La catégorie cible présentant la plus forte probabilité devient la valeur de sortie prédite de l'enregistrement.

Exemple binomial. Un fournisseur de télécommunications souhaite connaître le nombre de clients qui partent à la concurrence. Grâce aux données d'utilisation du service, vous pouvez créer un modèle binomial pour prévoir quels sont les clients susceptibles de s'adresser à un autre fournisseur et pour personnaliser les offres de façon à fidéliser autant de clients que possible. Le

modèle binomial est justifié par le fait que la cible présente deux catégories distinctes (passer à un autre fournisseur/ne pas passer à un autre fournisseur). [Pour plus d'informations, reportez-vous à la section Attrition dans le domaine des télécommunications \(régression logistique binomiale\) dans le chapitre 13 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Remarque : Pour les modèles binomiaux uniquement, les champs de type chaîne doivent être limités à huit caractères. Si nécessaire, il est possible de recoder les chaînes plus longues à l'aide d'un 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*.](#) [Pour plus d'informations, reportez-vous à la section Réduction de la longueur des chaînes de données d'entrée \(Reclassifier\) dans le chapitre 10 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Exemple multinomial. Un fournisseur de services de télécommunication a segmenté sa base de clients par type d'utilisation des services en catégorisant les clients en quatre groupes. En utilisant des données démographiques pour prévoir l'affectation des groupes, vous pouvez créer un modèle multinomial pour classer les clients potentiels dans des groupes, puis personnaliser les offres pour chacun d'entre eux. [Pour plus d'informations, reportez-vous à la section Classification des clients de télécommunications \(régression logistique multinomiale\) dans le chapitre 12 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Un ou plusieurs champs d'entrée et un seul champ de cible catégorielle avec deux catégories au minimum. Un modèle binomial requiert une cible avec un niveau de mesure *Booléen*. Pour un modèle multinomial, la cible peut avoir un niveau de mesure *Booléen*, ou *Nominal* avec deux catégories minimum. Les champs paramétrés sur *Les deux* ou *Aucun* sont ignorés. Les types des champs utilisés dans ce modèle doivent être complètement instanciés.

Puissance. Les modèles de régression logistique sont souvent très précis. Ils peuvent traiter des champs d'entrée symboliques et numériques. Ils peuvent fournir des probabilités prédites pour toutes les catégories cible, ce qui permet d'obtenir facilement une deuxième meilleure prévision. Les modèles logistiques montrent une efficacité optimale lorsque l'affectation des groupes représente un champ réellement catégoriel ; si l'affectation des groupes est basée sur les valeurs d'un champ d'intervalle continu (par exemple, QI élevé contre QI faible), envisagez d'utiliser une régression linéaire pour bénéficier des informations plus riches offertes par l'intervalle complet de valeurs. Les modèles logistiques peuvent également exécuter une sélection automatique des champs, même si d'autres approches comme le composant Sélection de fonction ou les modèles d'arbre exécutent cette sélection plus rapidement sur les ensembles de données volumineux. Enfin, étant donné que les modèles logistiques sont bien maîtrisés par de nombreux analystes et Data miners, ils peuvent constituer une référence par rapport à laquelle il est possible de comparer d'autres techniques de modélisation.

Lors du traitement d'ensembles de données volumineux, vous pouvez considérablement améliorer les performances en désactivant l'option Tests de rapport de vraisemblance (option de résultat avancée). [Pour plus d'informations, reportez-vous à la section Sorties avancées du noeud Régression logistique sur p. 299.](#)

Noeud Logistique - 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

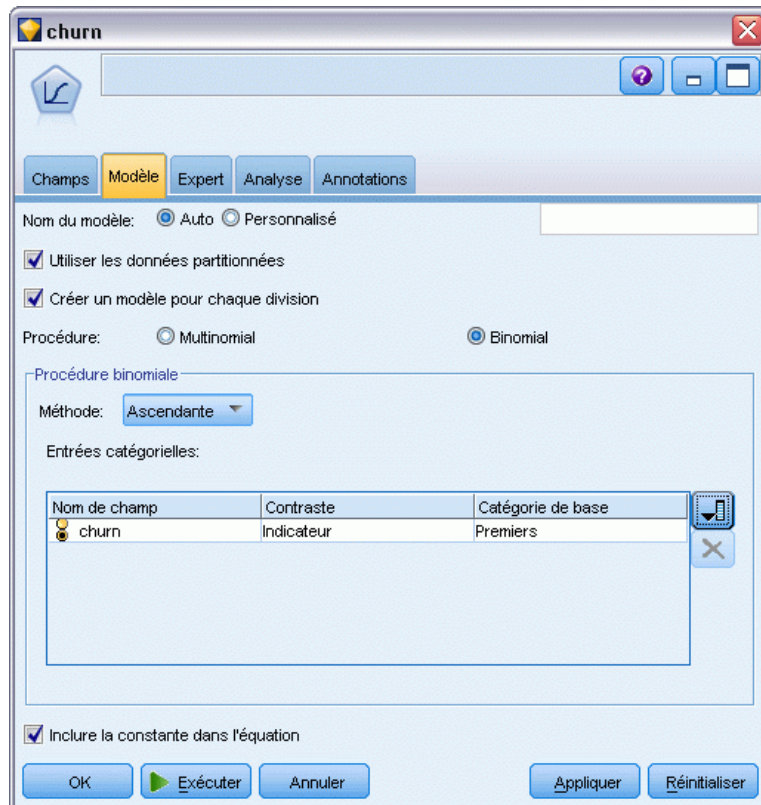
Procédure. Spécifie la création d'un modèle binomial ou d'un modèle multinomial. Les options disponibles dans la boîte de dialogue varient selon le type de procédure de modélisation sélectionné.

- **Binomiale.** Utilisé lorsque le champ cible est un champ booléen ou nominal avec deux valeurs discrètes (dichotomiques), telles que *oui/non*, *activé/désactivé*, *mâle/femelle*.
- **Multinomial.** Option utilisée lorsque le champ cible est un champ nominal avec plus de deux valeurs. Vous pouvez spécifier Effets principaux, Factoriel complet ou Personnalisé.

Inclure terme constant dans l'équation : Cette option détermine si les équations finales incluront ou non une constante. Il est conseillé de sélectionner cette option dans la plupart des situations.

Modèles binomiaux

Figure 10-21
Noeud Logistique, options du modèle binomial



Pour les modèles binomiaux, les méthodes et options suivantes sont disponibles :

Méthode : Indiquez la méthode devant être utilisée pour créer le modèle de régression logistique.

- **Entrée.** Cette méthode par défaut intègre directement toutes les caractéristiques dans l'équation. Aucune sélection de champ n'intervient dans la création du modèle.
- **Ascendante.** La méthode de sélection de champ Ascendante crée le modèle selon une progression étape par étape. Avec cette méthode, le modèle initial est le modèle le plus simple, et seules la constante et les caractéristiques peuvent y être ajoutées. A chaque étape, les caractéristiques non encore intégrées au modèle sont testées : le système évalue leur contribution à l'amélioration du modèle et les meilleures d'entre elles sont ajoutées au modèle final. Lorsqu'il est impossible d'ajouter des caractéristiques supplémentaires, ou que le meilleur candidat lui-même n'améliore pas le modèle de façon significative, le modèle final est généré.
- **Descendante.** Cette méthode est le contraire de la méthode Ascendante. Avec cette méthode, le modèle initial utilise toutes les caractéristiques en tant que valeurs prédites. Il est donc seulement possible de supprimer des caractéristiques du modèle. Les caractéristiques de modèle contribuant peu à l'amélioration du modèle sont supprimées une à une : lorsque la suppression d'une autre caractéristique ne peut que dégrader le modèle, le modèle final est généré.

Entrées catégorielles. Répertorie les champs identifiés comme catégoriels, c'est-à-dire ceux avec un niveau de mesure booléen, nominal ou ordinal. Vous pouvez définir le contraste et la catégorie de base de chaque champ catégoriel.

- **Nom du champ.** Cette colonne indique les noms de champ des entrées catégorielles ; elle contient toutes les valeurs booléennes et nominales des données. Pour ajouter des entrées continues ou numériques à cette colonne, cliquez sur l'icône Ajouter un champ à droite de la liste et sélectionnez les entrées requises.
- **Contraste.** L'interprétation des coefficients de régression d'un champ catégoriel dépend des contrastes utilisés. Le contraste détermine le mode de définition des tests d'hypothèse pour la comparaison des moyennes estimées. Par exemple, si vous savez qu'un champ catégoriel présente un ordre implicite, tel qu'un patron ou un regroupement, vous pouvez utiliser le contraste pour modéliser cet ordre. Les contrastes disponibles sont les suivants :

Indicateur : Les contrastes indiquent la présence ou l'absence d'appartenance à la modalité. Il s'agit de la méthode par défaut.

Simple : Chaque catégorie du champ prédicteur, à l'exception de la catégorie de référence, est comparée à la catégorie de référence.

Différence : Chaque catégorie du champ prédicteur, à l'exception de la première catégorie, est comparée à l'effet moyen des catégories précédentes. (Aussi connu sous le nom de contrastes inversés d'Helmert.)

Helmert : Chaque catégorie du champ prédicteur, à l'exception de la dernière catégorie, est comparée à l'effet moyen des catégories suivantes.

Répété : Chaque catégorie du champ prédicteur, à l'exception de la première catégorie, est comparée à la catégorie qui la précède.

Modèle polynomial : Contraste polynomial orthogonal. On part de l'hypothèse que les modalités sont espacées de manière équivalente. Les contrastes polynomiaux ne sont disponibles que pour les champs numériques.

type. Chaque catégorie du champ prédicteur, à l'exception de la catégorie de référence, est comparée à l'effet global.

- **Catégorie de base.** Spécifie la façon dont la catégorie de référence est déterminée pour le type de contraste sélectionné. Sélectionnez Premiers afin d'utiliser la première catégorie pour le champ d'entrée (trié dans l'ordre alphabétique) ou choisissez Derniers pour utiliser la dernière catégorie. La valeur par défaut est Premiers.

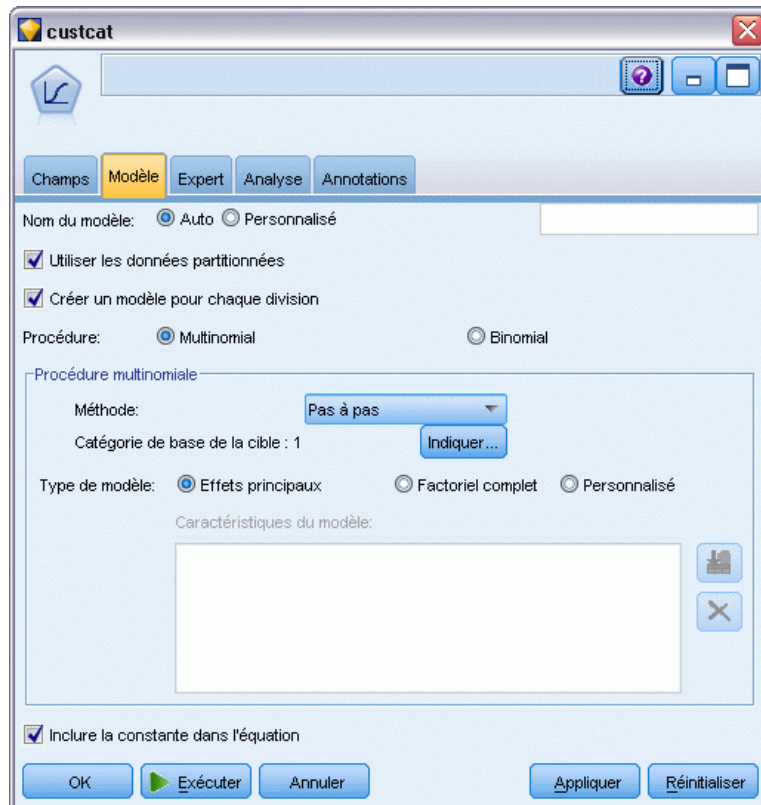
Remarque : ce champ n'est pas disponible si le paramètre de contraste est Différence, Helmert, Répété ou Polynomial.

L'estimation de l'effet de chaque champ sur la réponse générale est calculée comme une augmentation ou une diminution de la vraisemblance de chacune des autres catégories en rapport avec la catégorie de référence. Vous pouvez alors éventuellement identifier les champs et valeurs dont la probabilité de fournir une réponse spécifique est plus élevée.

La catégorie de base apparaît dans la sortie sous la forme 0,0. En effet, la comparaison de la catégorie de base avec elle-même engendre un résultat nul. Toutes les autres catégories sont présentées sous forme d'équations en rapport avec la catégorie de base. [Pour plus d'informations, reportez-vous à la section Modèle de nugget logistique - Détails sur p. 303.](#)

Modèles multinomiaux

Figure 10-22
Noeud Logistique, options du modèle multinomial



Pour les modèles multinomiaux, les méthodes et options suivantes sont disponibles :

Méthode : Indiquez la méthode devant être utilisée pour créer le modèle de régression logistique.

- **Entrée.** Cette méthode par défaut intègre directement toutes les caractéristiques dans l'équation. Aucune sélection de champ n'intervient dans la création du modèle.
- **Pas à pas.** Comme son nom l'indique, la méthode de sélection de champs pas à pas génère l'équation par étapes. Le modèle initial est le modèle le plus simple : son équation ne comporte aucune caractéristique de modèle (à l'exception de la constante). A chaque étape, les caractéristiques qui n'ont pas encore été intégrées au modèle sont évaluées et celles qui améliorent de manière significative la puissance de prévision du modèle sont alors ajoutées au modèle. De plus, les caractéristiques déjà intégrées au modèle sont réévaluées afin de déterminer si certaines d'entre elles peuvent être supprimées sans que cela affecte le fonctionnement du modèle. Si c'est le cas, elles sont supprimées. Le processus se répète et d'autres caractéristiques sont donc ajoutées et/ou supprimées. Lorsqu'aucune caractéristique ne peut être ajoutée au modèle pour l'améliorer, ou qu'aucune caractéristique ne peut être supprimée du modèle sans risquer de le dégrader, le modèle final est généré.
- **Ascendante.** Cette méthode de sélection des champs est similaire à la méthode Pas à pas dans la mesure où les modèles sont également générés par étapes. Cependant, avec cette méthode, le modèle initial est le modèle le plus simple, et seules la constante et les caractéristiques

peuvent être ajoutées à ce modèle. A chaque étape, les caractéristiques non encore intégrées au modèle sont testées : le système évalue leur contribution à l'amélioration du modèle et les meilleures d'entre elles sont ajoutées au modèle final. Lorsqu'il est impossible d'ajouter des caractéristiques supplémentaires, ou que le meilleur candidat lui-même n'améliore pas le modèle de façon significative, le modèle final est généré.

- **Descendante.** Cette méthode est le contraire de la méthode Ascendante. Avec cette méthode, le modèle initial utilise toutes les caractéristiques en tant que valeurs prédites. Il est donc seulement possible de supprimer des caractéristiques du modèle. Les caractéristiques de modèle contribuant peu à l'amélioration du modèle sont supprimées une à une : lorsque la suppression d'une autre caractéristique ne peut que dégrader le modèle, le modèle final est généré.
- **Pas à pas descendante.** Cette méthode est le contraire de la méthode Pas à pas. Avec cette méthode, le modèle initial utilise toutes les caractéristiques en tant que valeurs prédites. A chaque étape, les caractéristiques du modèle sont évaluées et celles dont la suppression n'a aucune incidence sur le fonctionnement du modèle sont supprimées. De plus, les caractéristiques précédemment supprimées sont réévaluées afin de déterminer si la meilleure d'entre elles améliore de manière significative la puissance de prévision du modèle. Si c'est le cas, elle est rajoutée au modèle. Lorsqu'aucune caractéristique ne peut être supprimée du modèle sans risquer de le dégrader, et qu'aucune caractéristique ne peut être ajoutée au modèle pour l'améliorer, le modèle final est généré.

Remarque : les méthodes de sélection de champs automatiques (Pas à pas, Ascendante, et Descendante) sont des méthodes d'apprentissage extrêmement flexibles qui ont une forte tendance à surajuster les données d'apprentissage. Lorsque vous utilisez ces méthodes, il est important de vérifier la validité du modèle généré, soit à l'aide de nouvelles données, soit à l'aide d'un échantillon de test retenu créé à l'aide du noeud Partitionner. [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.](#)

Catégorie de base de la cible. Indique le mode de détermination de la catégorie de référence. Il s'agit de la référence par rapport à laquelle les équations de régression de toutes les autres catégories de la cible sont estimées. Sélectionnez Premiers afin d'utiliser la première catégorie pour le champ cible en cours (trié dans l'ordre alphabétique) ou choisissez Derniers pour utiliser la dernière catégorie. Vous pouvez également sélectionner Spécifier pour choisir une catégorie spécifique, puis sélectionner la valeur souhaitée dans la liste. Les valeurs disponibles peuvent être définies pour chaque champ d'un noeud Typer. [Pour plus d'informations, reportez-vous à la section Utilisation de la boîte de dialogue Valeurs dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

La catégorie spécifiée est souvent celle qui, en tant que catégorie de base, vous intéresse le moins, par exemple un produit d'appel. Les autres catégories sont alors liées à cette catégorie de base de façon relative afin d'identifier ce qui les rend plus susceptibles d'être dans leur propre catégorie. Vous pouvez alors éventuellement identifier les champs et valeurs dont la probabilité de fournir une réponse spécifique est plus élevée.

La catégorie de base apparaît dans la sortie sous la forme 0,0. En effet, la comparaison de la catégorie de base avec elle-même engendre un résultat nul. Toutes les autres catégories sont présentées sous forme d'équations en rapport avec la catégorie de base. [Pour plus d'informations, reportez-vous à la section Modèle de nugget logistique - Détails sur p. 303.](#)

Type de modèle. Deux options permettent de définir les caractéristiques de votre modèle. Les modèles **Effets principaux** incluent les champs d'entrée individuellement et ne testent pas les interactions (effets multiplicateurs) entre les champs d'entrée. Les modèles **Factoriel complet** incluent toutes les interactions, ainsi que les effets principaux des champs d'entrée. Plus performants dans la capture de relations complexes, les modèles Factoriel complet sont également beaucoup plus difficiles à interpréter et davantage sujets au phénomène de surajustement. En raison du nombre potentiellement élevé de combinaisons possibles, les méthodes de sélection automatique des champs (autres que la méthode Entrée) sont désactivées pour les modèles Factoriel complet. Les modèles **personnalisés** comprennent uniquement les caractéristiques (effets principaux et interactions) que vous indiquez. Lorsque cette option est sélectionnée, utilisez la liste Caractéristiques du modèle pour ajouter des caractéristiques au modèle ou pour en supprimer.

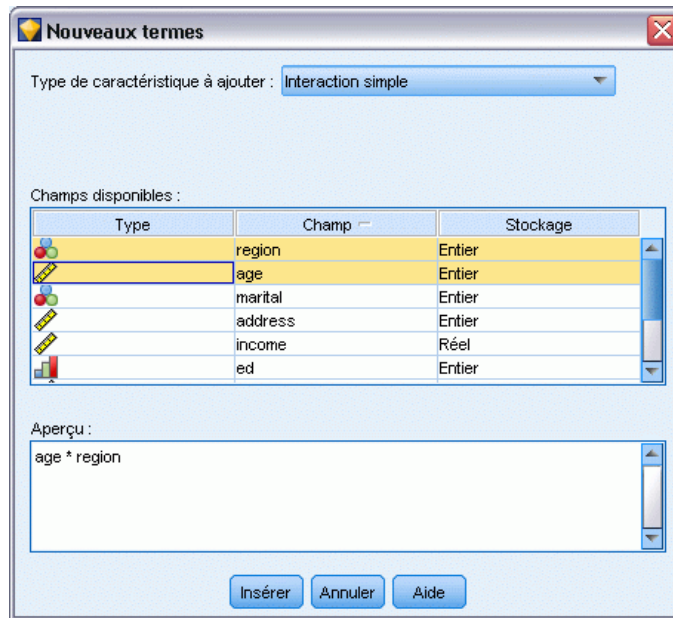
Caractéristiques du modèle. Lorsque vous créez un modèle personnalisé, vous devez indiquer explicitement ses caractéristiques. La liste répertorie toutes les caractéristiques actuelles du modèle. Les boutons situés à droite de la liste Caractéristiques du modèle permettent d'ajouter et de supprimer des caractéristiques de modèle.

- ▶ Pour ajouter des caractéristiques au modèle, cliquez sur le bouton *Ajouter de nouvelles caractéristiques au modèle*.
- ▶ Pour supprimer des caractéristiques, sélectionnez-les, puis cliquez sur le bouton *Supprimer les caractéristiques du modèle sélectionné*.

Ajout de caractéristiques à un modèle de régression logistique

Lorsque vous demandez un modèle de régression logistique personnalisé, vous pouvez ajouter des caractéristiques à ce modèle en cliquant sur le bouton *Ajouter de nouvelles caractéristiques au modèle* de l'onglet Modèle de régression logistique. Dans la boîte de dialogue qui apparaît, vous pouvez indiquer les caractéristiques souhaitées.

Figure 10-23
Boîte de dialogue Nouveaux termes d'un modèle de régression logistique



Type de caractéristique à ajouter. Il existe plusieurs méthodes pour ajouter des caractéristiques au modèle. Ces méthodes dépendent de la sélection des champs d'entrée dans la liste Champs disponibles.

- **Interaction simple.** Insère la caractéristique représentant l'interaction de tous les champs sélectionnés.
- **Effets principaux :** Insère une caractéristique effet principal (le champ lui-même) pour chaque champ d'entrée sélectionné.
- **Toutes les interactions de second ordre.** Insère une interaction de second ordre (le produit des champs d'entrée) pour chaque paire possible de champs d'entrée sélectionnés. Par exemple, si vous avez sélectionné les champs d'entrée A , B et C dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B$, $A * C$ et $B * C$.
- **Toutes les interactions de troisième ordre.** Insère une interaction de troisième ordre (le produit des champs d'entrée) pour chaque combinaison possible de champs d'entrée sélectionnés (trois à la fois). Par exemple, si vous avez sélectionné les champs d'entrée A , B , C et D dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B * C$, $A * B * D$, $A * C * D$ et $B * C * D$.
- **Toutes les interactions de quatrième ordre.** Insère une interaction de quatrième ordre (le produit des champs d'entrée) pour chaque combinaison possible de champs d'entrée sélectionnés (quatre à la fois). Par exemple, si vous avez sélectionné les champs d'entrée A , B , C , D et E dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B * C * D$, $A * B * C * E$, $A * B * D * E$, $A * C * D * E$ et $B * C * D * E$.

Champs disponibles. Répertoire les champs d'entrée disponibles à utiliser lors de la création des caractéristiques de modèle.

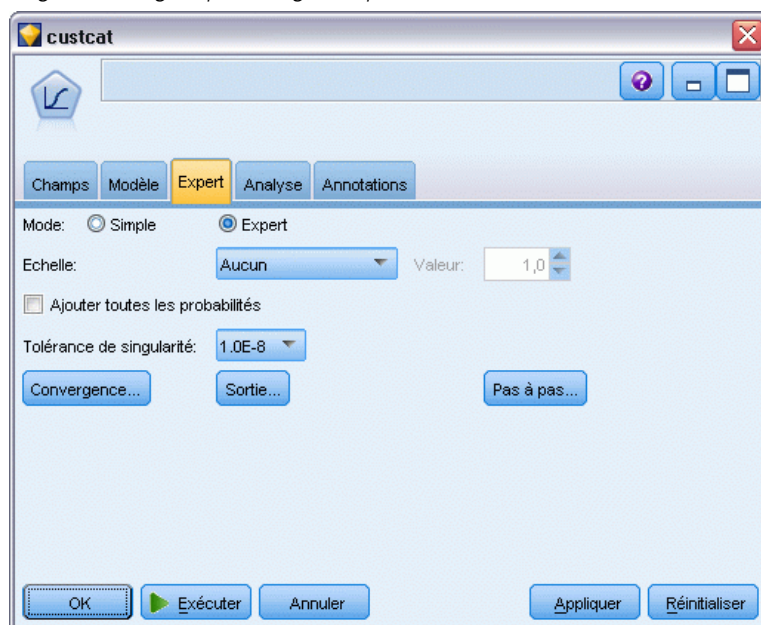
Aperçu. Affiche les caractéristiques qui seront ajoutées au modèle si vous cliquez sur Insérer, en fonction des champs sélectionnés et du type de caractéristique.

Insérer. Insère les caractéristiques du modèle (en fonction des champs et du type de caractéristique sélectionnés), puis ferme la boîte de dialogue.

Noeud Logistique - Options expert

Si vous êtes familier de la régression logistique, utilisez les options expert pour affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 10-24
Régression logistique - Onglet Expert



Echelle (modèles multinomiaux uniquement). Indiquez une valeur d'échelle de dispersion qui servira à corriger l'estimation de la matrice de covariances des paramètres. Pearson estime la valeur de l'échelle à l'aide des statistiques Pearson du Chi-deux. L'option Déviance estime la valeur de l'échelle à l'aide de la fonction de déviance (Chi-deux du maximum de vraisemblance). Vous pouvez également spécifier votre propre valeur d'échelle. Il doit s'agir d'une valeur numérique positive.

Ajouter toutes les probabilités. Lorsque cette option est sélectionnée, les probabilités de chacune des catégories du champ de sortie sont ajoutées à chacun des enregistrements traités par le noeud. Lorsqu'elle ne l'est pas, seule la probabilité de la catégorie prédite est ajoutée.

Par exemple, un tableau contenant les résultats d'un modèle multinomial avec trois catégories inclura cinq nouvelles colonnes. Une colonne indique la probabilité d'une prévision correcte du résultat ; la colonne suivante indique la probabilité de la prévision d'être une correspondance ou une absence de correspondance et trois colonnes supplémentaires indiquent la probabilité de la

prévision de chaque catégorie d'être une correspondance ou une absence de correspondance. [Pour plus d'informations, reportez-vous à la section Nugget de modèle logistique sur p. 302.](#)

Remarque : cette option est toujours sélectionnée pour les modèles binomiaux.

Tolérance singularité : Indiquez la tolérance utilisée pour la vérification des singularités.

Convergence. Ces options vous permettent de contrôler les paramètres de convergence des modèles. Lorsque vous exécutez le modèle, les paramètres de convergence contrôlent le nombre d'exécutions répétées des différents paramètres pour déterminer leur degré d'adéquation. Plus les essais de paramètres sont nombreux, plus les résultats sont proches (en d'autres termes, plus les résultats convergent). [Pour plus d'informations, reportez-vous à la section Options de convergence de la régression logistique sur p. 298.](#)

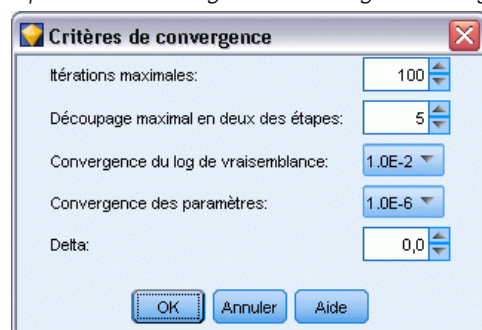
Résultats. Ces options vous permettent de demander des statistiques supplémentaires, qui apparaîtront dans les sorties avancées du nugget du modèle créé par le noeud. [Pour plus d'informations, reportez-vous à la section Sorties avancées du noeud Régression logistique sur p. 299.](#)

Pas à pas. Ces options vous permettent de contrôler les critères d'ajout et de suppression des champs avec les méthodes Pas à pas, Ascendante, Descendante ou Pas à pas descendante. (Le bouton est désactivé si la méthode Entrée est sélectionnée.) [Pour plus d'informations, reportez-vous à la section Régression logistique - Options pas à pas sur p. 301.](#)

Options de convergence de la régression logistique

Vous pouvez définir les paramètres de convergence pour l'estimation du modèle de régression logistique.

Figure 10-25
Options de convergence de la régression logistique



Nombre maximum d'itérations : Indiquez le nombre maximal d'itérations à utiliser pour l'évaluation du modèle.

Nombre maximum de dichotomie : Cette technique de régression logistique permet de traiter les complexités du processus d'estimation. Dans des circonstances normales, il est conseillé d'utiliser le paramètre par défaut.

Convergence de log-vraisemblance : Les itérations sont interrompues lorsque la modification relative du log de vraisemblance est inférieure à la valeur indiquée. Le critère n'est pas utilisé si la valeur est 0.

Convergence des paramètres : Les itérations sont interrompues lorsque la modification absolue ou relative des estimations des paramètres est inférieure à la valeur indiquée. Le critère n'est pas utilisé si la valeur est 0.

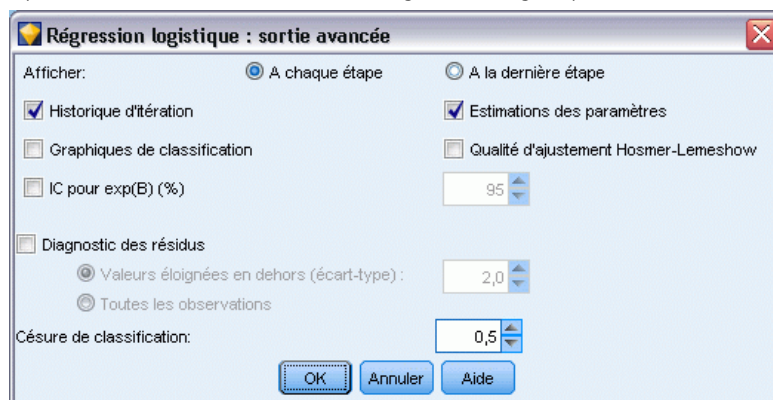
Delta (modèles multinomiaux uniquement). Spécifiez une valeur comprise entre 0 et 1 qui sera ajoutée à chacune des cellules vides (combinaison des valeurs des champs d'entrée et de sortie). L'algorithme d'estimation peut ainsi traiter des données dans lesquelles un grand nombre de combinaisons de valeurs de champ sont possibles (lorsqu'il existe un grand nombre d'enregistrements). La valeur par défaut est 0.

Sorties avancées du nœud Régression logistique

Sélectionnez le résultat facultatif à afficher dans la sortie avancée du nugget de modèle de régression. Pour visualiser les sorties avancées, parcourez le nugget du modèle et cliquez sur l'onglet Options avancées. [Pour plus d'informations, reportez-vous à la section Nugget de modèle Logistique - Sorties avancées sur p. 308.](#)

Options binomiales

Figure 10-26
Options de sortie binomiales de la régression logistique



Sélectionnez les types de sortie à générer pour le modèle. [Pour plus d'informations, reportez-vous à la section Nugget de modèle Logistique - Sorties avancées sur p. 308.](#)

Afficher : Choisissez d'afficher les résultats à chaque étape ou à la fin de toutes les étapes.

IC pour exp(B). Sélectionnez les intervalles de confiance pour chaque coefficient (indiqué sous la forme Bêta) de l'expression. Indiquez le niveau de l'intervalle de confiance (la valeur par défaut étant 95 %).

Diagnostic des résidus. Demande un tableau de diagnostic des observations des résidus.

- **Valeurs éloignées en dehors (écart-type).** Indique uniquement les observations résiduelles pour lesquelles la valeur standardisée absolue de la variable répertoriée est au moins aussi grande que la valeur spécifiée. La valeur par défaut est 2.
- **Toutes les observations.** Inclut toutes les observations dans le tableau de diagnostic des observations des résidus.

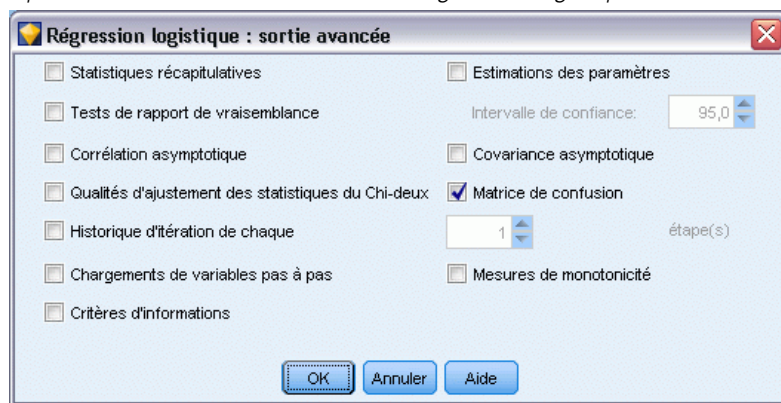
Remarque : étant donné que cette option répertorie chaque enregistrement d'entrée, elle peut aboutir à un tableau exceptionnellement grand dans le rapport, avec une ligne pour chaque enregistrement.

Césure du classement. Permet de déterminer le point de césure pour la classification des observations. Les observations avec des prévisions qui excèdent la limite de classification sont classées positives tandis que celles dont les prévisions sont inférieures à la limite sont classées négatives. Pour modifier la valeur par défaut, entrez une valeur entre 0.01 et 0.99.

Options multinomiales

Figure 10-27

Options de sortie multinomiales de la régression logistique



Sélectionnez les types de sortie à générer pour le modèle. [Pour plus d'informations, reportez-vous à la section Nugget de modèle Logistique - Sorties avancées sur p. 308.](#)

Remarque : la sélection de l'option de test des taux de vraisemblance accroît considérablement la durée de création d'un modèle de régression logistique. Si la création du modèle s'avère trop longue, désactivez cette option ou utilisez les statistiques de Wald et de score. [Pour plus d'informations, reportez-vous à la section Régression logistique - Options pas à pas sur p. 301.](#)

Historique d'itération de chaque. Sélectionnez l'intervalle d'étape devant être utilisé pour l'impression de l'état de l'itération dans la sortie avancée.

Intervalle de confiance : Intervalles de confiance des coefficients dans les équations. Indiquez le niveau de l'intervalle de confiance (la valeur par défaut étant 95 %).

Régression logistique - Options pas à pas

Figure 10-28
Régression logistique : critères de l'analyse pas à pas

Régression logistique : critères de l'analyse pas à pas

Nombre de caractéristiques du modèle :

Minimum: 0

Spécifier le maximum

Maximum: 10

Critère d'introduction: Rapport de vraisemblance Score

Critère de suppression: Rapport de vraisemblance Wald

Seuils de signification des critères LR :

Entrée: 0,05

Suppression: 0,1

Conditions requises de saisie ou de suppression :

Hiérarchie des effets discrets

Hiérarchie de tous les effets

Imbrication de tous les effets

Aucun

OK Annuler Aide

Nombre de termes du modèle (modèles multinomiaux uniquement). Vous pouvez indiquer le nombre minimal de caractéristiques pour les modèles Descendante et Pas à pas descendante, ainsi que le nombre maximal de caractéristiques pour les modèles Ascendante et Pas à pas. Si vous indiquez une valeur minimale supérieure à 0, le modèle contiendra le nombre de caractéristiques indiqué alors que certaines d'entre elles auraient été supprimées sur la base des critères statistiques. La valeur minimale est ignorée pour les modèles Ascendante, Pas à pas et Entrée. Si vous indiquez une valeur maximale, certaines caractéristiques peuvent être supprimées du modèle, alors qu'elles auraient été sélectionnées sur la base des critères statistiques. Le paramètre Indiquer valeur maximale est ignoré pour les modèles Descendante, Pas à pas descendante et Entrée.

Critère d'entrée (modèles multinomiaux uniquement). Sélectionnez Score pour optimiser la vitesse du traitement. L'option Rapport de vraisemblance peut fournir des estimations un peu plus fiables, mais les calculs effectués sont plus longs. Par défaut, il convient d'utiliser les statistiques de score.

Critère de suppression. Sélectionnez Rapport de vraisemblance pour obtenir un modèle plus fiable. Pour réduire la durée de création de ce modèle, vous pouvez essayer de sélectionner Wald. Toutefois, si les données sont totalement ou quasiment séparées (ce que vous pouvez déterminer dans l'onglet Options avancées du nugget de modèle), les statistiques de Wald ne sont pas fiables et ne doivent pas être utilisées. Par défaut, il convient d'utiliser les statistiques de rapport de vraisemblance. Pour les modèles binomiaux, il existe l'option supplémentaire Conditionnel. Elle fournit un test de suppression basé sur la probabilité de la statistique du rapport de vraisemblance, en fonction des estimations des paramètres conditionnels.

Seuils de signification des critères. Cette option vous permet de spécifier des critères de sélection basés sur la probabilité statistique (la valeur p) associée à chaque champ. Les champs sont ajoutés au modèle uniquement si la valeur p associée est inférieure à la valeur spécifiée dans le champ Entrée ; ils sont supprimés uniquement si la valeur p est supérieure à la valeur spécifiée dans le champ Suppression. La valeur du champ Entrée doit être inférieure à la valeur du champ Suppression.

Conditions requises de saisie ou de suppression (modèles multinomiaux uniquement). Pour certaines applications, l'ajout de caractéristiques d'interaction au modèle n'a aucun sens mathématique, sauf si le modèle contient également les caractéristiques d'ordre inférieur pour les champs concernés par la caractéristique d'interaction. Par exemple, l'ajout de $A * B$ au modèle peut n'avoir aucun sens, sauf si A et B font également partie du modèle. Les options suivantes vous permettent de déterminer le mode de traitement de ces dépendances au cours de la sélection des caractéristiques pas à pas.

- **Hiérarchie des effets discrets.** Les effets d'ordre supérieur (interactions impliquant plus de champs) seront ajoutés au modèle uniquement si tous les effets d'ordre inférieur (effets principaux ou interactions impliquant moins de champs) des champs correspondants se trouvent déjà dans le modèle. De plus, les effets d'ordre inférieur ne seront pas supprimés si les effets d'ordre supérieur impliquant les mêmes champs se trouvent dans le modèle. Cette option s'applique uniquement aux champs catégoriels. [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*.](#)
- **Hiérarchie de tous les effets.** Cette option fonctionne de la même façon que l'option précédente, sauf qu'elle s'applique à tous les champs d'entrée.
- **Imbrication de tous les effets.** Les effets peuvent être inclus dans le modèle uniquement si tous les effets contenus dans un effet sont également inclus dans le modèle. Cette option est identique à l'option Hiérarchie de tous les effets, sauf que les champs continus sont traités de manière quelque peu différente. Pour un effet qui doit en contenir un autre, l'effet (d'ordre inférieur) contenu doit comprendre *tous* les champs continus de l'effet (d'ordre supérieur) contenant et tous les champs catégoriels de l'effet contenu doivent représenter un sous-ensemble de ceux qui se trouvent dans l'effet contenant. Par exemple, si A et B sont des champs catégoriels et que X est un champ continu, la caractéristique $A * B * X$ contient les caractéristiques $A * X$ et $B * X$.
- **Aucune.** Aucune relation n'est appliquée : les caractéristiques sont ajoutées et supprimées séparément dans le modèle.

Nugget de modèle logistique

Un nugget de modèle logistique représente l'équation estimée par un noeud Logistique. Il contient toutes les informations rassemblées par le modèle de régression logistique, ainsi que des informations sur la structure et les performances du modèle. Ce type d'équation peut également être généré par d'autres modèles comme Oracle SVM.

Lorsque vous exécutez un flux contenant un nugget de modèle logistique, le noeud crée deux champs contenant la prévision du modèle et la probabilité associée. Les noms des nouveaux champs sont formés du nom du champ de sortie sur lequel porte la prévision, auquel est ajouté le préfixe $\$L$ - pour désigner la catégorie de prévision et $\$LP$ - pour désigner le champ de la probabilité associée. Par exemple, si le nom du champ de sortie est *prefcoul*, les nouveaux champs

seront alors intitulés SL -*prefcoul* et SLP -*prefcoul*. De plus, si vous avez sélectionné l'option Ajouter toutes les probabilités dans le noeud Logistique, un champ supplémentaire sera créé pour chaque catégorie du champ de sortie et affichera la probabilité associée à la catégorie. Les noms de ces champs supplémentaires sont basés sur les valeurs du champ de sortie auxquelles le préfixe SLP - est ajouté. Par exemple, si les valeurs valides du champ *prefcoul* sont *Rouge*, *Vert* et *Bleu*, les trois nouveaux champs suivants sont créés : SLP -*Rouge*, SLP -*Vert* et SLP -*Bleu*.

Génération d'un noeud Filtrer. Le menu Générer vous permet de créer un noeud Filtrer pour transmettre des champs d'entrée en fonction des résultats du modèle. Les champs supprimés du modèle à cause de la multicollinéarité sont filtrés par le noeud généré, ainsi que les champs qui ne sont pas utilisés dans le modèle.

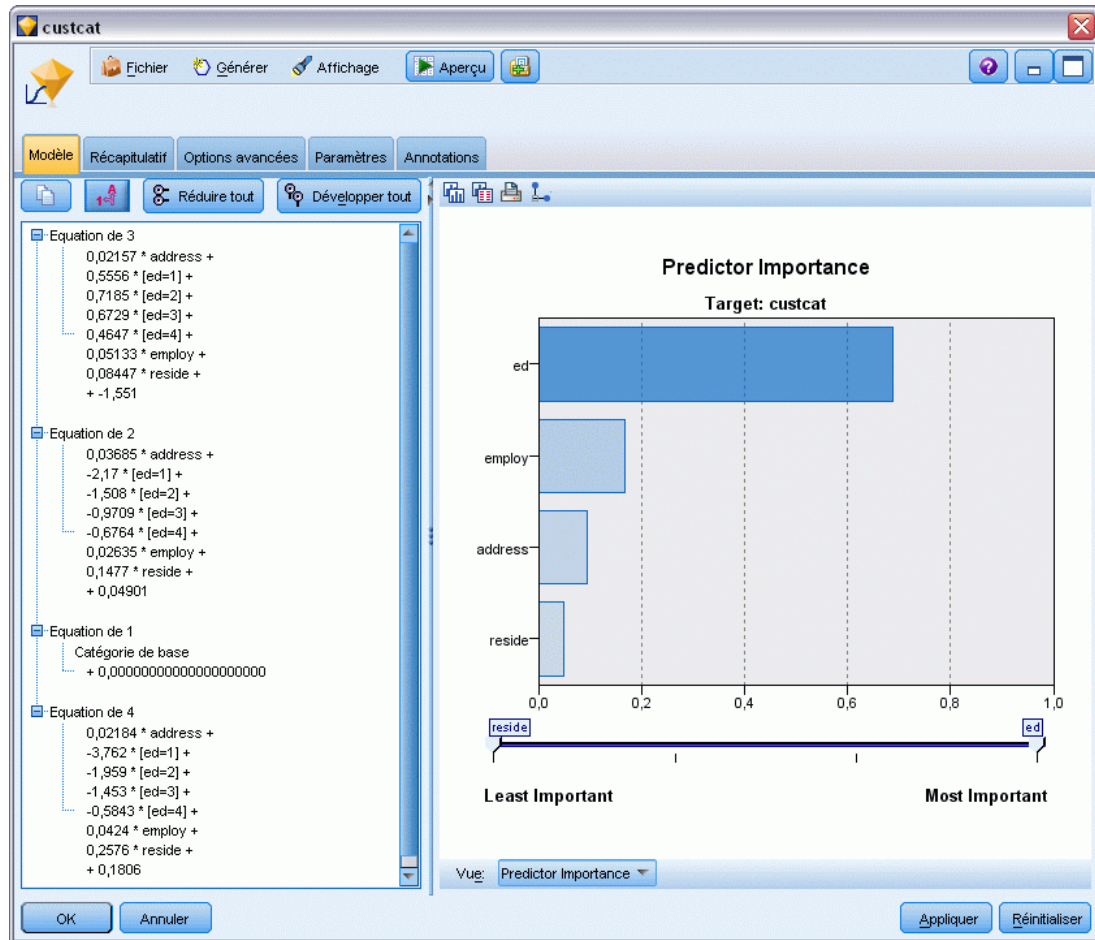
Modèle de nugget logistique - Détails

Pour des modèles multinomiaux, l'onglet Modèle d'un nugget de modèle logistique possède un affichage divisé avec les équations du modèle dans le panneau gauche et l'importance des valeurs prédites dans le panneau droit. Pour des modèles binomiaux, l'onglet affiche uniquement l'importance des valeurs prédites. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Equations de modèle

Pour des modèles multinomiaux, le panneau de gauche affiche les équations réelles estimées pour le modèle de régression logistique. Il y a une équation par catégorie dans le champ cible, à l'exception de la catégorie de référence. Les équations apparaissent sous forme d'arbres. Ce type d'équation peut également être généré par certains autres modèles tels qu'Oracle SVM.

Figure 10-29
 Détails d'un modèle de nugget logistique avec affichage de l'importance des valeurs prédites



Equation de. Affiche les équations de régression utilisées pour calculer les probabilités de catégorie cible, selon un ensemble de valeurs prédites. La dernière catégorie du champ cible est considérée comme la **catégorie de référence**. Les équations affichées donnent les log-odds des autres catégories cible par rapport à la référence, pour un ensemble particulier de valeurs prédites. La probabilité prédite de chaque catégorie du modèle de prévision indiqué provient de ces valeurs de log-odds.

Comment les probabilités sont calculées

Chaque équation calcule les log-odds d'une catégorie cible donnée, par rapport à la catégorie de référence. Les **log-odds**, ou **logits**, sont le rapport de la probabilité de la catégorie cible spécifiée sur celui de la catégorie de référence, la fonction de logarithme naturel étant appliquée au résultat. Pour la catégorie de référence, l'incidence de la catégorie par rapport à elle-même est de 1,0. Les log-odds ont donc pour valeur 0. Pour simplifier, imaginez qu'il s'agit d'une équation implicite de la catégorie de référence où tous les coefficients ont pour valeur 0.

Pour calculer la probabilité à partir des log-odds d'une catégorie cible précise, reprenez la valeur de logit calculée par l'équation de cette catégorie et appliquez la formule suivante :

$$P(\text{groupe}_i) = \exp(g_i) / \sum_k \exp(g_k)$$

g correspondant aux log-odds calculés, i à l'index de catégorie et k étant compris entre 1 et le nombre de catégories cible.

Importance des valeurs prédites

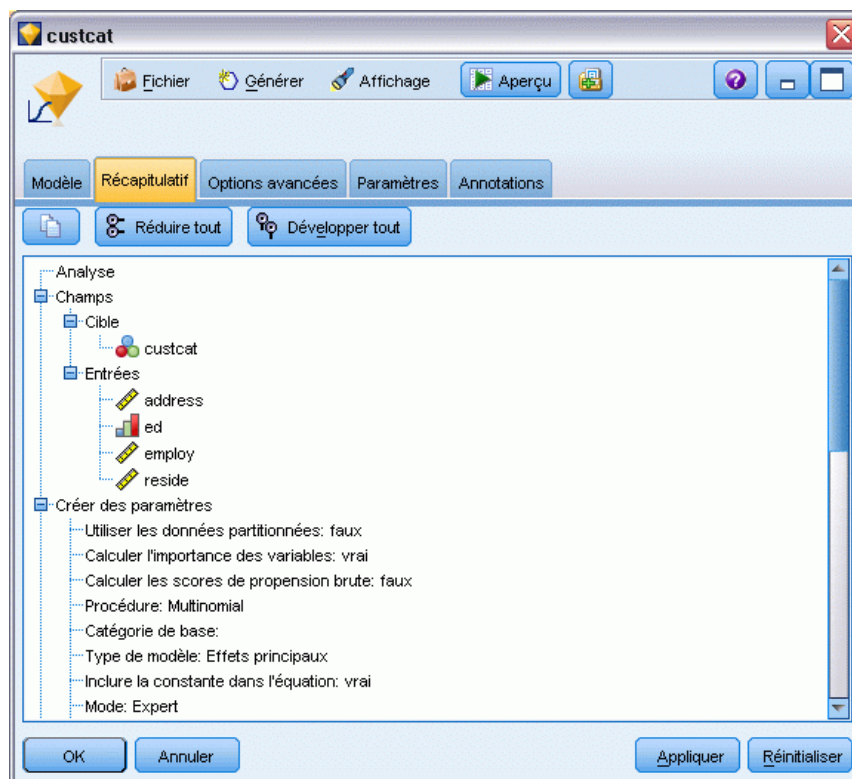
Un graphique illustrant l'importance relative de chaque valeur prédite dans l'estimation du modèle peut également être affiché dans l'onglet Modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique n'est disponible que si Calculer l'importance des valeurs prédites a été sélectionné dans l'onglet Analyse avant la génération du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Remarque : Le calcul de l'importance des valeurs prédites peut prendre davantage de temps pour une régression logistique que pour d'autres types de modèles et il n'est pas sélectionné par défaut dans l'onglet Analyse. Le choix de cette option risque de ralentir les performances, surtout avec de grands ensembles de données.

Nugget de modèle logistique - Récapitulatif

Le récapitulatif d'un modèle de régression logistique affiche les champs et paramètres utilisés pour la génération du modèle. En outre, si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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*. Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à \[Navigation dans les nuggets de modèle sur p. 55.\]\(#\)](#)

Figure 10-30
Nugget de modèle de régression logistique - Onglet Récapitulatif



Nugget de modèle logistique - Paramètres

L'onglet Paramètres d'un nugget de modèle logistique propose des options pour les confiances, les probabilités, les scores de propension et la génération SQL lors du scoring du modèle. Cet onglet est uniquement disponible une fois que le nugget de modèle a été ajouté à un flux et affiche différentes options en fonction du type de modèle et de cible.

Modèles multinomiaux

Pour les modèles multinomiaux, les options suivantes sont disponibles :

Calculer les confiances. Indique si les confiances sont calculées lors de la détermination des scores.

Calculer les scores de propension brute (cibles de type booléen uniquement). Pour les modèles possédant des cibles de type booléen uniquement, vous pouvez demander des scores de propension brute qui indiquent la vraisemblance de la sortie *true* (*vrai*) spécifiée pour le champ cible. Ceux-ci s'ajoutent aux valeurs standard de prédiction et de confiance. Les scores de propension ajustée ne sont pas disponibles. [Pour plus d'informations, reportez-vous à la section Options d'analyse des noeuds de modélisation dans le chapitre 3 sur p. 44.](#)

Ajouter toutes les probabilités. Spécifie l'ajout ou le non-ajout des probabilités de chacune des catégories du champ de sortie à chacun des enregistrements traités par le noeud. Lorsqu'elle ne l'est pas, seule la probabilité de la catégorie prédite est ajoutée. Pour une cible nominale comportant trois catégories, par exemple, la sortie du scoring comprend une colonne pour chacune des trois catégories, plus une quatrième colonne qui indique la probabilité pour toute catégorie prédite. Par exemple, si les probabilités des catégories *Rouge*, *Vert* et *Bleu* valent respectivement 0,6, 0,3, et 0,1, la catégorie prédite est *Rouge*, avec une probabilité de 0,6.

Evaluer en convertissant en SQL natif. Si cette option est sélectionnée, le SQL est généré pour évaluer le modèle de manière native dans l'application.

Remarque : Pour des modèles multinomiaux, la génération SQL n'est pas disponible si l'option Ajouter toutes les probabilités a été sélectionnée, ou — pour des modèles comportant des cibles nominales — si l'option Calculer les confiances a été sélectionnée. La génération SQL avec calculs de confiance est prise en charge uniquement pour les modèles multinomiaux comportant des cibles de type booléen. La génération SQL n'est pas disponible pour les modèles binomiaux.

Modèles binomiaux

Pour les modèles binomiaux, les confiances et les probabilités sont toujours activées et les paramètres qui vous permettraient de désactiver ces options ne sont pas disponibles. La génération SQL n'est pas disponible pour les modèles binomiaux. Le seul paramètre modifiable pour les modèles binomiaux est la capacité à calculer des scores de propension brute. Comme indiqué précédemment pour les modèles multinomiaux, cela ne s'applique qu'aux modèles comportant des cibles de type booléen. [Pour plus d'informations, reportez-vous à la section Options d'analyse des noeuds de modélisation dans le chapitre 3 sur p. 44.](#)

Nugget de modèle Logistique - Sorties avancées

Figure 10-31

Exemple : noeud Equation de régression logistique – Onglet Options avancées

Nominal Regression

Case Processing Summary

		N	Marginal Percentage
custcat	Basic service	266	26.6%
	E-service	217	21.7%
	Plus service	281	28.1%
	Total service	236	23.6%
region	Zone 1	322	32.2%
	Zone 2	334	33.4%
	Zone 3	344	34.4%
marital	Unmarried	505	50.5%
	Married	495	49.5%
ed	Did not complete high school	204	20.4%
	High school degree	287	28.7%
	Some college	209	20.9%
	College degree	234	23.4%
	Post-undergraduate degree	66	6.6%
retire	No	953	95.3%
	Yes	47	4.7%

La sortie avancée de la régression logistique (également appelée **régression nominale**) fournit des informations détaillées sur le modèle évalué et ses performances. Dans la mesure où la plupart des informations contenues dans la sortie avancée ont un caractère technique, il est nécessaire d'avoir une bonne connaissance de l'analyse de régression logistique pour pouvoir les interpréter correctement.

Avertissements. Indique tout avertissement ou problème potentiel relatif aux résultats.

Récapitulatif du traitement des observations. Répertorie le nombre d'enregistrements traités pour chaque champ symbolique que contient le modèle.

Récapitulatif des étapes (facultatif). Répertorie les effets ajoutés ou supprimés à chaque étape de la création du modèle, lorsque la sélection automatique des champs est utilisée.

Remarque : affiché uniquement pour les méthodes Pas à pas, Ascendante, Descendante et Pas à pas descendante.

Historique d'itération (facultatif). Affiche l'historique d'itération des estimations des paramètres toutes les n itérations, en commençant par les estimations initiales, où n est la valeur de l'intervalle d'impression. La valeur par défaut consiste à imprimer chaque itération ($n=1$).

Informations sur l'ajustement du modèle (modèles multinomiaux). Affiche le test de rapport de vraisemblance de votre modèle final comparé à un test dans lequel tous les coefficients des paramètres sont égaux à 0 (seulement la constante).

Classification (facultatif). Affiche la matrice des valeurs prédites et réelles des champs de sortie, ainsi que des pourcentages.

Qualités d'ajustement des statistiques du Chi-deux (facultatif). Affiche les statistiques de Pearson et de vraisemblance du Chi-deux. Ces statistiques permettent de tester l'ajustement global du modèle aux données d'apprentissage.

Qualité d'ajustement Hosmer-Lemeshow (facultatif). Affiche les résultats du groupement des observations en déciles de risque et de la comparaison entre la probabilité observée et la probabilité attendue au sein de chaque décile. Cette statistique de qualité d'ajustement est plus fiable que la statistique de qualité d'ajustement traditionnelle utilisée dans les modèles multinomiaux, notamment pour les modèles avec des covariables continues et les études avec des échantillons de petite taille.

Pseudo R-deux (facultatif). Affiche les mesures R -deux d'ajustement du modèle de Cox et Snell, Nagelkerke et McFadden. Ces statistiques sont quasi-identiques à R -deux de la régression linéaire.

Mesures de monotonie (facultatif). Affiche le nombre de paires concordantes, de paires discordantes et de paires reliées dans les données, ainsi que le pourcentage du nombre total de paires que chaque catégorie représente. Le D de Somers, le Γ de Goodman et Kruskal, le τ -a de Kendall et l'Indice de concordance C apparaissent également dans ce tableau.

Critères d'informations (facultatif). Affiche les critères d'informations d'Akaike (AIC) et les critères d'informations bayésien de Schwarz (BIC).

Tests de rapport de vraisemblance (facultatif). Affiche les statistiques permettant de vérifier que les coefficients des effets du modèle sont statistiquement différents de 0. Les champs d'entrée sont ceux dont les niveaux de signification de sortie sont peu importants (indiqués par *Sig.*).

Estimations des paramètres (facultatif). Affiche des estimations des coefficients de l'équation, les tests de ces coefficients, les ratios d'incidence approchés obtenus à partir des coefficients (intitulés $Exp(B)$) et les intervalles de confiance de ces ratios d'incidence approchés.

Covariance asymptotique/Matrice de corrélations (facultatif). Affiche les covariances et/ou corrélations asymptotiques des estimations des coefficients.

Fréquences observées et prédites (facultatif). Pour chaque paramètre de covariable, indique les fréquences observées et prédites de chaque valeur de champ de sortie. Ce tableau peut être d'une taille relativement importante, notamment pour les modèles comportant des champs d'entrée

numériques. Si la taille de ce tableau est trop importante, il n'est pas généré et un avertissement apparaît.

Noeud ACP/Analyse factorielle

Le noeud ACP/Analyse factorielle propose des techniques de factorisation puissantes qui vous permettent de réduire la complexité de vos données. Deux approches similaires mais distinctes sont disponibles.

- **L'analyse en composantes principales (ACP)** recherche les combinaisons linéaires des champs d'entrée qui permettent de capturer au mieux la variance dans l'ensemble de champs, où les composantes sont orthogonales (perpendiculaires) les unes par rapport aux autres. La technique ACP se concentre sur tous les types de variance, y compris les variances partagée et unique.
- **L'analyse factorielle** a pour but d'identifier les concepts sous-jacents, appelés **facteurs**, qui expliquent la tendance des corrélations dans un ensemble de champs observés. L'analyse factorielle se concentre uniquement sur la variance partagée. Une variance unique propre à certains champs n'est pas prise en compte dans l'estimation du modèle. Le noeud Analyse factorielle/ACP comporte plusieurs méthodes d'analyse factorielle.

Quelle que soit l'approche choisie, le but consiste à trouver un nombre limité de champs dérivés récapitulant les informations contenues dans l'ensemble de champs d'origine.

Conditions requises. Seuls les champs numériques peuvent être utilisés dans un modèle factoriel-ACP. L'estimation d'une analyse factorielle ou d'une ACP requiert l'utilisation d'au moins un champ avec le rôle défini sur les champs d'*entrée*. Les champs avec le rôle défini sur *Cible*, *Les deux* ou *Aucun* sont ignorés, tout comme les champs non numériques.

Puissance. L'analyse factorielle et l'analyse ACP permettent de réduire la complexité de vos données sans pour autant sacrifier leur contenu informatif. Elles génèrent des modèles robustes, dont l'exécution est plus rapide qu'avec des champs d'entrée bruts.

Noeud ACP/Analyse factorielle – Options du modèle

Figure 10-32
Onglet Modèles ACP/Analyse factorielle



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 d'extraction. Indiquez la méthode de factorisation.

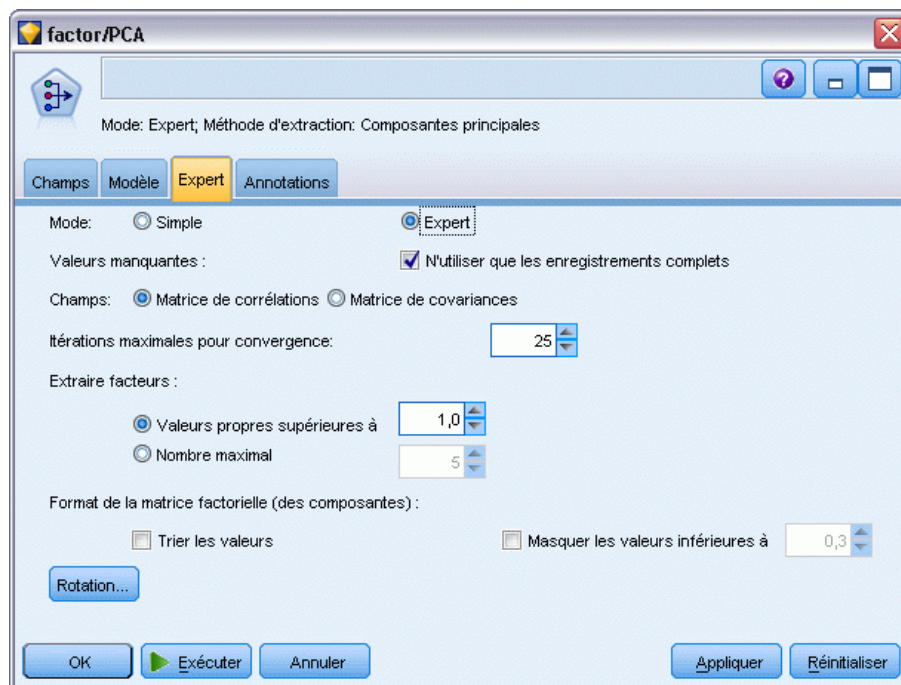
- **Composantes principales.** Il s'agit de la méthode par défaut : l'analyse en composantes principales recherche les composantes qui récapitulent les champs d'entrée.
- **Moindres carrés non pondérés.** Cette méthode d'analyse factorielle recherche l'ensemble de facteurs pouvant reproduire le modèle des relations (corrélations) entre les champs d'entrée.
- **Moindres carrés généralisés.** Cette méthode d'analyse factorielle, similaire à celle des moindres carrés non pondérés, utilise une technique de pondération afin de donner moins de valeur aux champs présentant une variance unique (non partagée).
- **Maximum de vraisemblance.** Cette méthode d'analyse factorielle génère les équations factorielles qui sont les plus susceptibles d'avoir produit le modèle de relations (corrélations) observé dans les champs d'entrée, en se basant sur des hypothèses relatives à la forme de ces relations. Le principe de départ de cette méthode est que les données d'apprentissage ont une distribution multivariable normale.
- **Factorisation en axes principaux.** Cette méthode d'analyse factorielle, très proche de l'analyse en composantes principales, se concentre exclusivement sur la variance partagée.

- **Alpha-maximisation.** Cette méthode d'analyse factorielle considère les champs à analyser comme un échantillon des champs d'entrée potentiels. Elle maximise la fiabilité statistique des facteurs.
- **Factorisation en projections.** Cette méthode d'analyse factorielle utilise la technique d'estimation des données afin d'isoler la variance commune, ainsi que ses facteurs descriptifs.

Noeud ACP/Analyse factorielle – Options expert

Si vous êtes familier des analyses factorielles et en composantes principales, utilisez Les options expert pour affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 10-33
Analyse factorielle/ACP - Onglet Expert



Valeurs manquantes : Par défaut, IBM® SPSS® Modeler utilise exclusivement les enregistrements dont les valeurs sont valides pour tous les champs du modèle. (cette stratégie est parfois appelée **suppression des observations incomplètes** des valeurs manquantes). S'il manque un grand nombre de données, cette stratégie risque d'éliminer trop d'enregistrements ; par conséquent, la quantité de données disponibles peut être insuffisante pour générer un modèle pertinent. Dans ce cas, vous pouvez désélectionner l'option N'utiliser que les enregistrements complets. SPSS Modeler essaie alors d'utiliser autant d'informations que possible pour estimer le modèle, y compris des enregistrements dont certains champs auxquels il manque des valeurs. (Cette stratégie est parfois appelée **suppression des valeurs manquantes appariées**.) Toutefois dans certains cas, l'utilisation d'enregistrements incomplets peut entraîner des difficultés pour le calcul de l'estimation du modèle.

Champs. Choisissez la matrice de corrélation (option par défaut) ou la matrice de covariances des champs d'entrée pour l'estimation du modèle.

Itérations maximales pour convergence. Indiquez le nombre maximal d'itérations à utiliser pour l'évaluation du modèle.

Extraire facteurs. Vous pouvez sélectionner le nombre de facteurs à extraire des champs d'entrée de deux manières différentes.

- **Valeurs propres supérieures à.** Cette option permet de conserver tous les facteurs ou composantes dont les valeurs propres sont supérieures au critère spécifié. Les **valeurs propres** mesurent la capacité de chaque facteur ou composante à récapituler la variance dans l'ensemble des champs d'entrée. Si vous optez pour la matrice de corrélation, le modèle conservera tous les facteurs ou composantes dont les valeurs propres sont supérieures à la valeur spécifiée. Si vous préférez utiliser la matrice de covariances, le critère sera égal à la valeur spécifiée, multipliée par la valeur propre moyenne. Cette mise à l'échelle permet de donner des résultats similaires quelle que soit la matrice utilisée.
- **Nombre maximal.** Cette option permet de conserver le nombre spécifié de facteurs ou de composantes par ordre décroissant des valeurs propres. Autrement dit, les facteurs ou composantes correspondant aux n valeurs propres les plus élevées sont conservé (où n correspond au critère spécifié). Le critère d'extraction par défaut est de cinq facteurs/composantes.

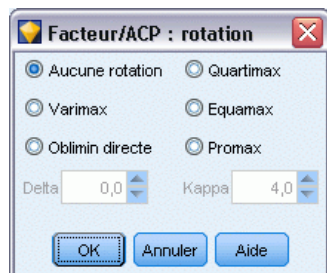
Format de la matrice factorielle (des composantes). Ces options contrôlent le format de la matrice factorielle (ou de la matrice des composantes pour les modèles ACP).

- **Trier les valeurs.** Lorsque cette option est sélectionnée, les facteurs sont triés selon leur pondération (valeur numérique) dans la sortie du modèle.
- **Masquer les valeurs inférieures à.** Si cette option est sélectionnée, les scores inférieurs au seuil spécifié sont masqués dans la matrice, ce qui permet de voir plus facilement les caractéristiques de la matrice.

Rotation. Ces options vous permettent de contrôler la méthode de rotation à appliquer au modèle. [Pour plus d'informations, reportez-vous à la section Noeud ACP/Analyse factorielle – Options de rotation sur p. 313.](#)

Noeud ACP/Analyse factorielle – Options de rotation

Figure 10-34
Options de rotation du noeud ACP/Analyse factorielle



De manière générale, le fait d'effectuer une rotation mathématique sur l'ensemble des facteurs conservés permet de faciliter leur interprétation. Sélectionnez une méthode de rotation :

- **Aucune rotation.** Il s'agit de l'option par défaut. Aucune rotation n'est effectuée.
- **Varimax.** Méthode de rotation orthogonale permettant de minimiser le nombre de champs qui pondèrent fortement chacun des facteurs. Cette méthode permet de simplifier l'interprétation des facteurs.
- **Oblimin directe.** Méthode de rotation oblique (non orthogonale). Lorsque le paramètre Delta est paramétré sur 0 (valeur par défaut), les solutions sont obliques. Plus la valeur de delta est négative, moins les facteurs sont obliques. Pour remplacer la valeur nulle par défaut de delta, entrez un nombre inférieur ou égal à 0,8.
- **Quartimax.** Méthode orthogonale qui minimise le nombre de facteurs requis pour expliquer chaque champ. Cette méthode permet de simplifier l'interprétation des champs observés.
- **Equamax.** Méthode de rotation qui combine la méthode Varimax, simplifiant les facteurs, et la méthode Quartimax, simplifiant les champs. Le nombre de champs pondérant fortement un facteur et le nombre de facteurs nécessaires pour expliquer un champ sont minimisés.
- **Promax.** Rotation oblique qui permet la corrélation des facteurs. Son calcul étant plus rapide que celui d'une rotation Oblimin directe, cette méthode s'avère particulièrement utile pour traiter des volumes importants de données. Kappa contrôle l'obliquité de la solution (dans quelle mesure les facteurs peuvent être corrélés).

Nugget de modèle ACP/Analyse factorielle

Un nugget de modèle ACP/Analyse factorielle représente le modèle d'analyse factorielle et d'analyse de la composante principale (ACP) créés par un noeud ACP/Analyse factorielle. Ils contiennent toutes les informations rassemblées par le réseau formé, ainsi que des informations sur les performances et les caractéristiques du modèle.

Lorsque vous exécutez un flux contenant un modèle Equation factorielle, le noeud ajoute un nouveau champ pour chaque facteur ou composant que comporte le modèle. Les noms attribués aux nouveaux champs sont constitués du nom du modèle auquel le préfixe *\$F-* et le suffixe *-n* sont ajoutés, *n* correspondant au numéro du facteur ou de la composante. Par exemple, si le nom de votre modèle est intitulé *Facteur* et si le modèle contient trois facteurs, les noms des nouveaux champs sont *\$F-Facteur-1*, *\$F-Facteur-2* et *\$F-Facteur-3*.

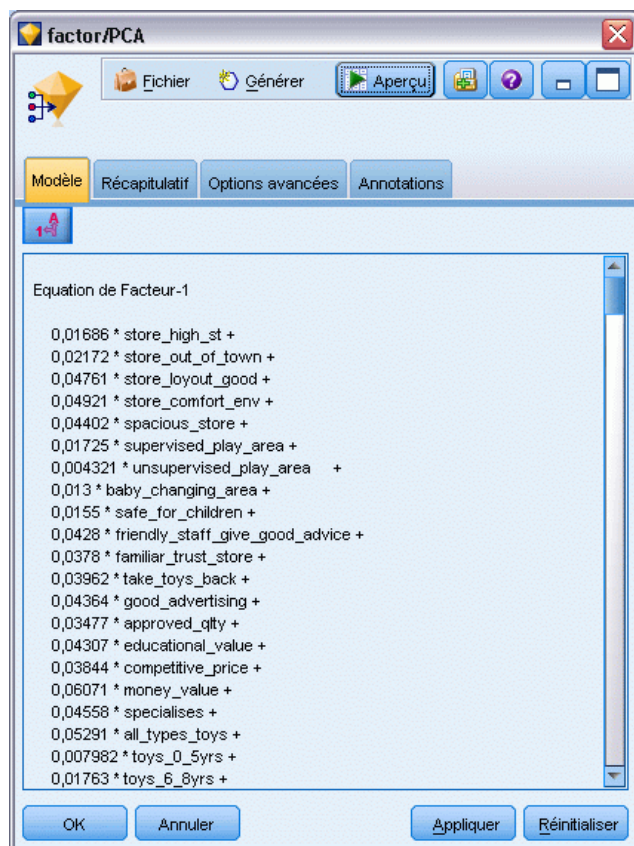
Pour mieux comprendre ce qui a été codé par le modèle factoriel, vous pouvez effectuer une analyse en aval approfondie. L'une des méthodes permettant d'évaluer un modèle factoriel consiste à afficher les corrélations entre les facteurs et les champs d'entrée à l'aide d'un noeud Statistiques. Cette méthode vous permet de voir quels champs d'entrée ont un poids important sur les facteurs et donc de déterminer si les facteurs ont une signification sous-jacente. [Pour plus d'informations, reportez-vous à la section Noeud Statistiques dans le chapitre 6 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)

Vous pouvez également évaluer le modèle factoriel en utilisant les informations fournies par la sortie avancée. Pour afficher les options de sortie avancées, cliquez sur l'onglet Options avancées du navigateur de nugget de modèle. La sortie avancée contient une grande quantité d'informations détaillées et est destinée aux utilisateurs ayant une connaissance approfondie de l'analyse factorielle et de l'analyse en composantes principales. [Pour plus d'informations, reportez-vous à la section Nugget de modèle ACP/Analyse factorielle - Sorties avancées sur p. 317.](#)

Equations de nugget de modèle ACP/Analyse factorielle

L'onglet Modèle d'un nugget de modèle factoriel affiche l'équation des scores factoriels de chaque facteur. Les scores des facteurs ou des composantes sont calculés en multipliant chaque champ d'entrée par son coefficient et en additionnant les résultats obtenus.

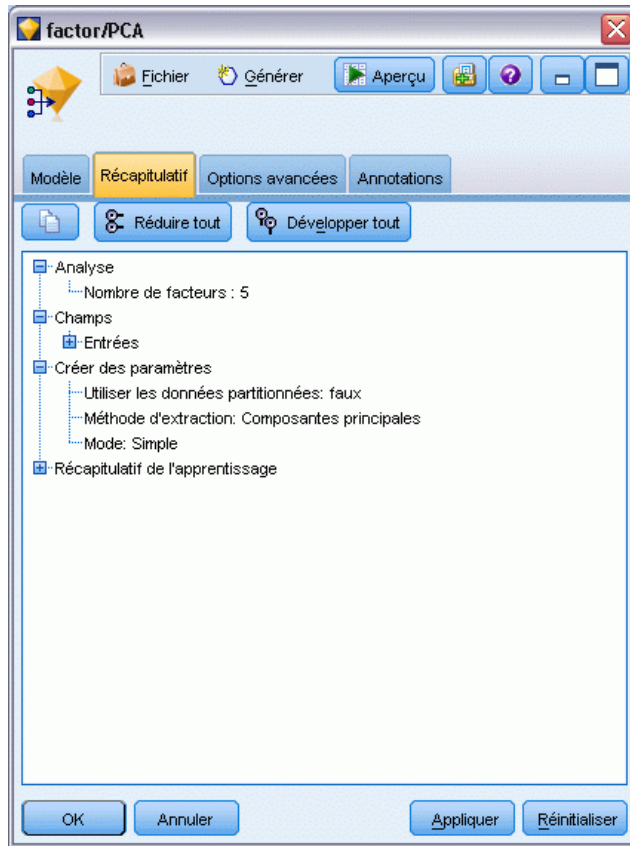
Figure 10-35
Nugget ACP/Analyse factorielle - Onglet Modèle



Nugget de modèle ACP/Analyse factorielle - Récapitulatif

L'onglet Récapitulatif d'un modèle factoriel affiche le nombre de facteurs conservés dans le modèle facteur/ACP, ainsi que des informations supplémentaires relatives aux champs et paramètres utilisés dans la génération du modèle. [Pour plus d'informations, reportez-vous à la section Navigation dans les nuggets de modèle dans le chapitre 3 sur p. 55.](#)

Figure 10-36
Exemple : noeud Equation factorielle Onglet Récapitulatif



Nugget de modèle ACP/Analyse factorielle - Sorties avancées

Figure 10-37

Exemple : onglet Options avancées du noeud Equation factorielle

	Initial	Extraction
store_high_st	1.000	.283
store_out_of_town	1.000	.307
store_layout_good	1.000	.434
store_comfort_env	1.000	.423
spacious_store	1.000	.467
supervised_play_area	1.000	.705
unsupervised_play_area	1.000	.275
baby_changing_area	1.000	.620
safe_for_children	1.000	.524
friendly_staff_give_good_advice	1.000	.409
familiar_trust_store	1.000	.334
take_tours_back	1.000	.209

La sortie avancée de l'analyse factorielle vous fournit des informations détaillées sur le modèle évalué et ses performances. Dans la mesure où la plupart des informations contenues dans la sortie avancée ont un caractère technique, il est nécessaire d'avoir une bonne connaissance de l'analyse factorielle pour pouvoir interpréter cette sortie correctement.

Avertissements. Indique tout avertissement ou problème potentiel relatif aux résultats.

Qualité de représentation. Affiche la proportion de la variance de chaque champ due aux facteurs ou composantes. La colonne *Initial* indique la qualité de représentation initiale obtenue avec la totalité des facteurs (au départ, le modèle utilise un nombre donné de facteurs comme champs d'entrée) et la colonne *Extraction* indique la qualité de représentation obtenue avec les facteurs retenus.

Variance totale expliquée. Affiche la variance totale due aux facteurs utilisés par le modèle. *Valeurs propres initiales* affiche la variance due aux facteurs initiaux. *Sommes des carrés des facteurs retenus* affiche la variance due aux facteurs retenus par le modèle. *Sommes des carrés pour la rotation* affiche la variance due à la rotation des facteurs. Notez que pour les rotations obliques, le paramètre *Sommes des carrés pour la rotation* indique uniquement les sommes des carrés et n'affiche pas les pourcentages des variances.

Matrice factorielle (des composantes). Affiche les corrélations entre les champs d'entrée et les facteurs avant rotation.

Matrice factorielle (des composantes) après rotation. Affiche pour les rotations orthogonales les corrélations entre les champs d'entrée et les facteurs ayant subi une rotation.

Matrice des types. Affiche pour les rotations obliques les corrélations partielles entre les champs d'entrée et les facteurs ayant subi une rotation.

Matrice de structure. Affiche pour les rotations obliques les corrélations simples entre les champs d'entrée et les facteurs ayant fait l'objet d'une rotation.

Matrice de corrélation factorielle. Affiche les corrélations entre les facteurs pour les rotations obliques.

Noeud discriminant

L'analyse discriminante crée un modèle prédictif pour l'affectation des groupes. Le modèle est composé d'une fonction discriminante (ou, pour plus de deux groupes, d'un ensemble de fonctions discriminantes) basée sur les combinaisons linéaires des variables de prévision offrant la meilleure discrimination entre les groupes. Les fonctions sont générées à partir d'un échantillon d'observations pour lesquelles l'affectation des groupes est connue. Les fonctions peuvent alors être appliquées à de nouvelles observations bénéficiant de mesures pour les variables de prévision, mais dont l'affectation des groupes est inconnue.

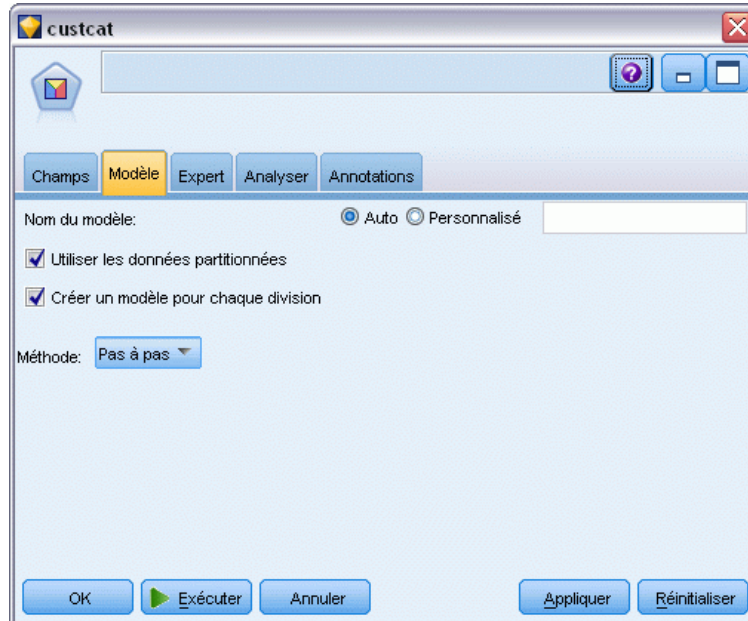
Exemple : Une société de télécommunications peut utiliser l'analyse discriminante pour classer les clients par groupes en fonction de l'utilisation des données. Ceci lui permet d'évaluer les clients potentiels et de cibler ceux qui sont le plus susceptibles d'appartenir à des groupes importants. [Pour plus d'informations, reportez-vous à la section Classification des clients de services de télécommunications \(analyse discriminante\) dans le chapitre 21 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Vous avez besoin d'un ou de plusieurs champs d'entrée et précisément d'un champ cible. La cible doit être un champ catégoriel (avec un niveau de mesure *Booléen* ou *Nominal*) avec stockage de chaîne ou d'entier. (Le stockage peut être converti à l'aide d'un noeud Remplacer ou Calculer si nécessaire. [Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)). Les champs paramétrés sur *Les deux* ou sur *Aucun* sont ignorés. Les types des champs utilisés dans ce modèle doivent être complètement instanciés.

Puissance. L'analyse discriminante et la régression logistique sont des modèles de classification supervisée. Toutefois, l'analyse discriminante formule plus d'hypothèses sur les champs d'entrée—par exemple, elle suppose qu'ils sont normalement distribués, continus, et qu'ils donnent de meilleurs résultats si ces conditions sont remplies, en particulier si la taille de l'échantillon est réduite.

Noeud discriminant - Options de modèle

Figure 10-38
Boîte de dialogue du noeud discriminant, 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Méthode : Les options suivantes sont disponibles pour la saisie de valeurs prédites dans le modèle :

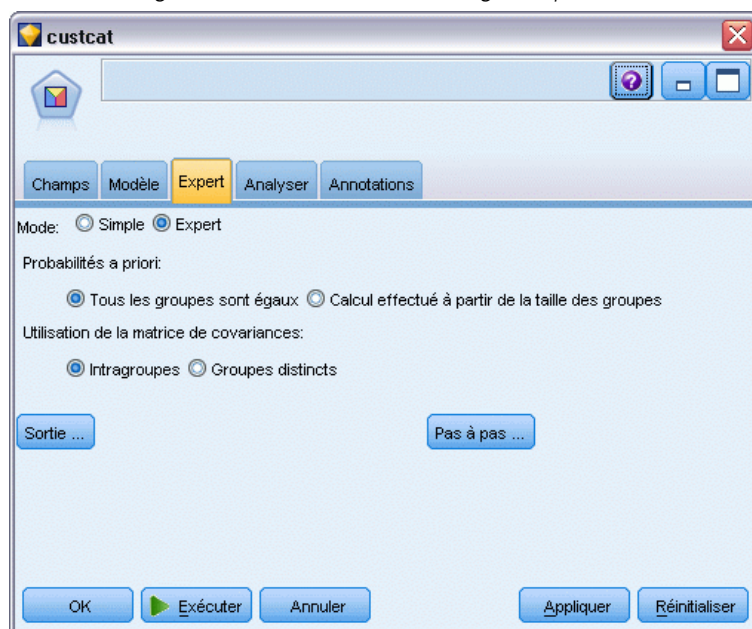
- **Entrée.** Cette méthode par défaut intègre directement toutes les caractéristiques dans l'équation. les termes qui n'améliorent pas la puissance de prévision du modèle de manière significative ne sont pas ajoutés.
- **Pas à pas.** Le modèle initial est le modèle le plus simple : son équation ne comporte aucune caractéristique de modèle (à l'exception de la constante). A chaque étape, les caractéristiques qui n'ont pas encore été intégrées au modèle sont évaluées et celles qui améliorent de manière significative la puissance de prévision du modèle sont alors ajoutées au modèle.

Remarque : La méthode Pas à pas a une forte tendance à surajuster les données d'apprentissage. Lorsque vous utilisez ces méthodes, il est important de vérifier la validité du modèle généré, soit à l'aide d'un échantillon de test retenu soit à l'aide de nouvelles données.

Noeud discriminant - Options Expert

Si vous êtes familier de l'analyse discriminante, utilisez les options expert pour affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 10-39
Boîte de dialogue du noeud discriminant, onglet Expert



Arbre CRT : probabilités a priori. Cette option détermine si les coefficients de classification supervisée sont ajustés pour une connaissance a priori de l'affectation des groupes.

- **Tous les groupes sont égaux.** Des probabilités a priori égales sont supposées pour tous les groupes et cela n'a aucun effet sur les coefficients.
- **Calcul effectué à partir de la taille des groupes.** La taille des groupes observée dans l'échantillon détermine les probabilités a priori de l'affectation des groupes. Par exemple, si 50 % des observations incluses dans l'analyse sont rattachées au premier groupe, 25 % au deuxième et 25 % au troisième, les coefficients de classification supervisée sont ajustés pour augmenter la probabilité d'appartenance au premier groupe par rapport aux deux autres.

Utilisation de la matrice de covariances. Vous pouvez choisir de classer les observations en utilisant une matrice de covariances intragroupes ou une matrice de covariances par groupes distincts.

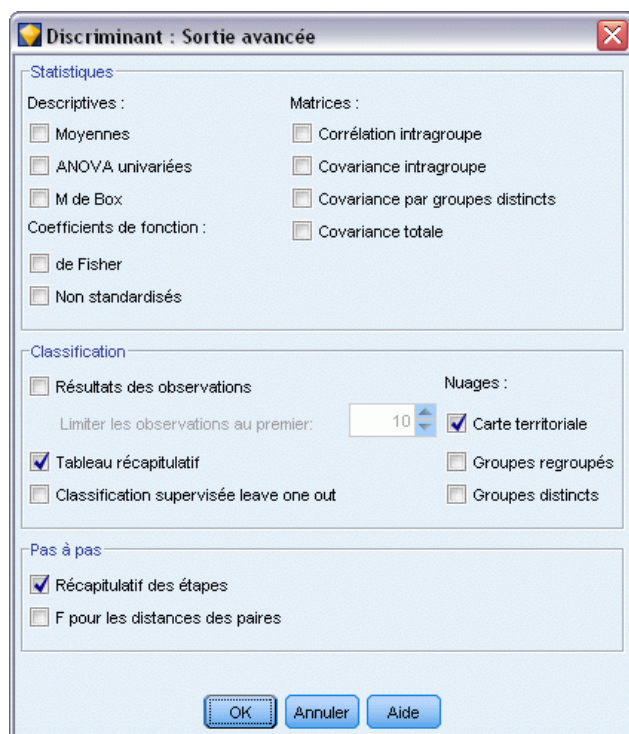
- **Intragroupes.** La matrice de covariances intra-classes globale est utilisée pour classer les observations.
- **Groupes distincts.** Les matrices de covariances de chaque groupe sont utilisées pour la classification. Comme la classification repose sur les fonctions discriminantes et pas sur les variables d'origine, cette option n'est pas toujours équivalente à la discrimination quadratique.

Résultats. Ces options vous permettent de demander des statistiques supplémentaires, qui apparaîtront dans les sorties avancées du nugget du modèle créé par le noeud. [Pour plus d'informations, reportez-vous à la section Noeud discriminant - Options de sortie sur p. 321.](#)

Pas à pas. Ces options vous permettent de contrôler les critères d'ajout et de suppression des champs avec la méthode d'estimation Pas à pas. (Le bouton est désactivé si la méthode Entrée est sélectionnée.) [Pour plus d'informations, reportez-vous à la section Noeud discriminant - Options pas à pas sur p. 323.](#)

Noeud discriminant - Options de sortie

Figure 10-40
Noeud discriminant - Options de résultat avancées



Sélectionnez la sortie facultative à afficher dans la sortie avancée du nugget du modèle de régression logistique. Pour visualiser les sorties avancées, parcourez le nugget du modèle et cliquez sur l'onglet Options avancées. [Pour plus d'informations, reportez-vous à la section Sortie avancée du nugget du modèle discriminant sur p. 325.](#)

Descriptives. Les options disponibles sont les moyennes (y compris les écarts-types), les ANOVA univariées et le test *M* de Box.

- **Moyennes.** Affiche le total et la moyenne de chaque groupe ainsi que l'écart-type des variables explicatives.
- **ANOVA à 1 facteur.** Effectue pour chacune des variables indépendantes une analyse de variance à 1 facteur pour tester l'égalité des moyennes de groupe.
- **M de Box.** Test d'égalité des matrices de covariance des classes. Pour les échantillons de taille suffisamment importante, une valeur *p* non significative indique qu'il n'est pas démontré que les matrices diffèrent. Ce test est sensible aux déviations par rapport à la distribution gaussienne multivariée.

Coefficients de fonction. Les options disponibles sont les coefficients de classification supervisée de Fisher et les coefficients non standardisés.

- **Fisher.** Affiche les coefficients de la fonction de classification de Fisher qui peuvent être directement utilisés pour la classification. Un groupe séparé de coefficients de fonctions de classification est obtenu pour chaque groupe et une observation est affectée au groupe qui a le plus grand score discriminant (valeur de fonction de classification).
- **Non standardisés.** Affiche les coefficients non normalisés de la fonction discriminante.

Matrices. Les matrices des coefficients disponibles pour les variables indépendantes sont la matrice de corrélations intragroupes, la matrice de covariances intragroupes, la matrice de covariances par groupes distincts et la matrice de covariances totales.

- **Corrélation intragroupe.** Affiche une matrice de corrélations intra-classes globale, en calculant la moyenne des matrices de covariance distinctes pour tous les groupes avant de calculer les corrélations.
- **Covariance intragroupe.** Affiche une matrice de covariances intra-classes globale, qui peut différer de la matrice de covariance totale. Cette matrice est obtenue en calculant la moyenne des matrices de covariances distinctes de tous les groupes.
- **Covariance par groupes distincts.** Affiche des matrices de covariances distinctes pour chaque groupe.
- **Covariance totale.** Affiche la matrice de covariance de toutes les observations comme si elles provenaient d'un seul échantillon.

Classification La sortie suivante appartient aux résultats de la classification supervisée.

- **Résultats par observation.** Les codes du groupe actuel, du groupe prévu, des probabilités a posteriori et des scores discriminants sont affichés pour chaque observation.
- **Récapitulatif.** Nombre d'observations correctement et incorrectement affectées à chacune des classes sur la base de l'analyse discriminante. Parfois appelés « matrice de confusion ».
- **Classification par élimination.** Classement de chaque observation de l'analyse par les fonctions dérivées de l'ensemble des observations autres que cette observation. Cette classification est également appelée « méthode U ».
- **Carte territoriale.** Diagramme des limites servant à classer les observations en fonction de valeurs de fonction. Les numéros correspondent aux groupes auxquels les observations ont été affectées. La moyenne de chaque groupe est indiquée par un astérisque à l'intérieur de ses limites. La carte n'est pas affichée s'il n'existe qu'une seule fonction discriminante.
- **Toutes classes combinées.** Crée un diagramme de dispersion de tous les groupes, des valeurs des deux premières fonctions discriminantes. S'il n'y a qu'une seule fonction, un histogramme est tracé à la place.
- **Classe par classe.** Crée des diagrammes de dispersion classe par classe pour les deux premières valeurs de fonction discriminante. Lorsqu'il n'y a qu'une seule fonction, des histogrammes sont affichés à la place.

Pas à pas. Le récapitulatif des étapes affiche des statistiques pour toutes les variables après chaque étape ; le rapport F concernant les distances des paires affiche une matrice des rapports F de paires pour chaque paire de groupes. Les rapports F peuvent être utilisés pour des tests de signification des distances de Mahalanobis entre les groupes.

Noeud discriminant - Options pas à pas

Figure 10-41
Noeud discriminant - Options Méthode pas à pas

Discriminant : Critères de l'analyse pas à pas

Méthode:

- Lambda de Wilk
- Variance inexpliquée
- Distance de Mahalanobis
- Plus petit rapport F
- V de Rao

V-à-entrer:

Critères:

- Utiliser la valeur F
 - Entrée: Suppression:
 - La valeur d'entrée doit être supérieure à la valeur de suppression.
- Utiliser la probabilité de F
 - Entrée: Suppression:
 - La valeur d'entrée doit être inférieure à la valeur de suppression.

OK Annuler Aide

Méthode. Sélectionnez le type de statistique à utiliser pour entrer ou supprimer de nouvelles variables. Les choix disponibles sont le Lambda de Wilk, la variance inexpliquée, la distance de Mahalanobis, le plus petit rapport F et le V de Rao. Avec le V de Rao, vous pouvez spécifier l'augmentation minimale dans V pour une variable à entrer.

- **Lambda de Wilk.** Méthode de sélection des variables pour une analyse discriminante pas à pas qui sélectionne les variables à entrer dans l'équation d'après leur capacité à faire baisser le lambda de Wilks. A chaque étape, les variables sont entrées dans l'analyse d'après leur capacité à faire baisser le lambda de Wilks.
- **Variance inexpliquée.** A chaque étape, la variable qui minimise la somme des variations résiduelles entre les groupes est saisie.
- **Distance de Mahalanobis.** Mesure de la distance entre les valeurs d'une observation et la moyenne de toutes les observations sur les variables indépendantes. Une distance de Mahalanobis importante identifie une observation qui a des valeurs extrêmes pour des variables indépendantes.
- **Plus petit rapport F.** Méthode de sélection des variables en analyse pas à pas, fondée sur la maximisation d'un rapport F calculé à partir de la distance de Mahalanobis entre des groupes.
- **V de Rao.** Mesure des différences entre des moyennes de groupes. Egalement appelée trace de Lawley-Hotelling. A chaque étape, la variable qui maximise l'augmentation du V de RAO est entrée. Après avoir sélectionné cette option, entrez la valeur minimale que doit avoir une variable pour entrer dans l'analyse.

Critères. Les choix disponibles sont Utiliser la valeur F et Utiliser la probabilité de F . Entrez des valeurs pour l'entrée et la suppression de variables.

- **Utiliser la valeur F .** Une variable est introduite dans un modèle si sa valeur F est supérieure à la valeur Entrée et elle est éliminée si la valeur F est inférieure à la valeur Elimination. La valeur Entrée doit être supérieure à la valeur Elimination et toutes deux doivent être positives. Pour

introduire davantage de variables dans le modèle, réduisez la valeur du champ Entrée. Pour éliminer davantage de variables dans le modèle, augmentez la valeur du champ Elimination.

- **Utiliser la probabilité de F.** Une variable est entrée dans le modèle si le seuil de signification de la valeur F est inférieur à la valeur Entrée ; la variable est éliminée si ce seuil est supérieur à la valeur Elimination. La valeur Entrée doit être inférieure à la valeur Elimination et toutes deux doivent être positives. Pour introduire davantage de variables dans le modèle, diminuez la valeur Entrée. Pour éliminer davantage de variables du modèle, réduisez la valeur Elimination.

Nugget du modèle discriminant

Les nuggets du modèle discriminants représentent les équations estimées par les noeuds discriminants. Ils contiennent toutes les informations rassemblées par le modèle discriminant, ainsi que des informations sur la structure et les performances du modèle.

Lorsque vous exécutez un flux contenant un nugget du modèle discriminant, le noeud ajoute deux nouveaux champs contenant la prévision du modèle et la probabilité associée. Les noms des nouveaux champs sont formés du nom du champ de sortie sur lequel porte la prévision, auquel est ajouté le préfixe *\$D-* pour désigner la catégorie de prévision et *\$DP-* pour désigner la probabilité associée. Par exemple, si le nom du champ de sortie est *prefcoul*, les nouveaux champs seront alors intitulés *\$D-prefcoul* et *\$DP-prefcoul*.

Génération d'un noeud Filtrer. Le menu Générer vous permet de créer un noeud Filtrer pour transmettre des champs d'entrée en fonction des résultats du modèle.

Importance des valeurs prédites

Un graphique illustrant l'importance relative de chaque valeur prédite dans l'estimation du modèle peut également être affiché dans l'onglet Modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique n'est disponible que si Calculer l'importance des valeurs prédites a été sélectionné dans l'onglet Analyse avant la génération du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Sortie avancée du nugget du modèle discriminant

Figure 10-42
Nugget du modèle discriminant, onglet Options avancées

Analysis Case Processing Summary			
Unweighted Cases		N	Percent
Valid		1000	100.0
Excluded	Missing or out-of-range group codes		0 .0
	At least one missing discriminating variable		0 .0
	Both missing or out-of-range group codes and at least one missing discriminating variable		0 .0
	Total		0 .0
Total		1000	100.0

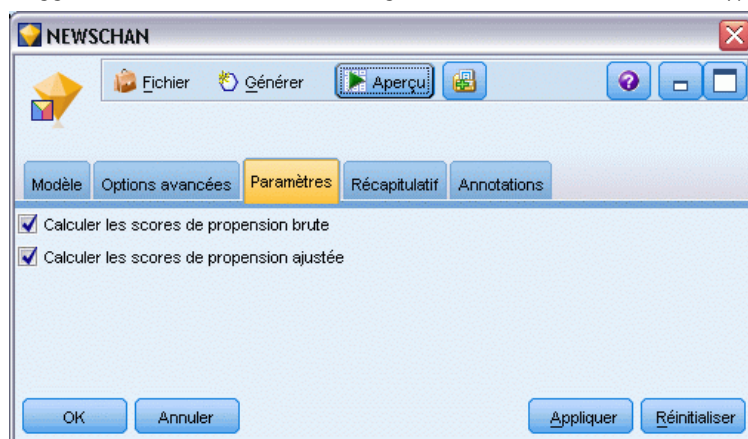
Group Statistics			
Customer category		Valid N (listwise)	
		Unweighted	Weighted
	Geographic indicator	266	266.000

La sortie avancée de l'analyse discriminante vous fournit des informations détaillées sur le modèle évalué et ses performances. Dans la mesure où la plupart des informations contenues dans la sortie avancée ont un caractère technique, il est nécessaire d'avoir une bonne connaissance de l'analyse discriminante pour pouvoir interpréter cette sortie correctement. [Pour plus d'informations, reportez-vous à la section Noeud discriminant - Options de sortie sur p. 321.](#)

Paramètres du nugget du modèle discriminant

L'onglet Paramètres d'un nugget du modèle discriminant vous permet d'obtenir les scores de propension lors de la détermination des scores du modèle. Cet onglet est disponible uniquement pour les modèles avec des cibles de type booléen, et seulement après que le nugget de modèle ait été ajouté à un flux.

Figure 10-43
Nugget du modèle discriminant, onglet Paramètres d'une cible de type booléen



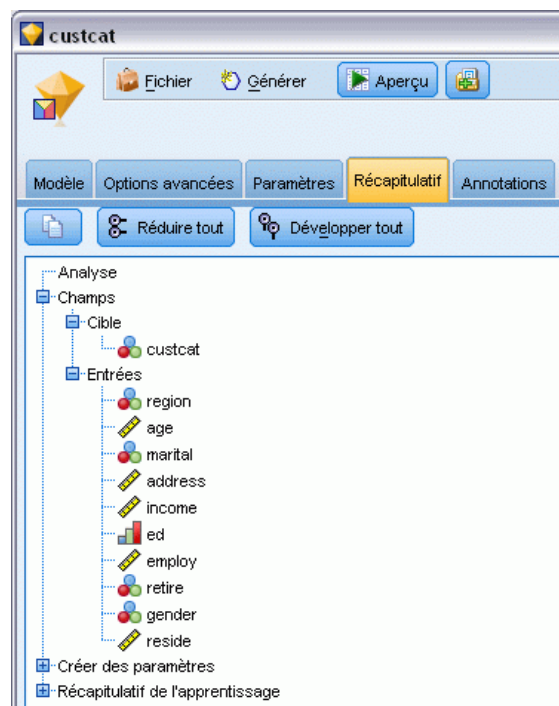
Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Récapitulatif du nugget du modèle discriminant

L'onglet Récapitulatif d'un nugget du modèle discriminant affiche les champs et paramètres utilisés pour la génération du modèle. En outre, si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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*](#). Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle sur p. 55](#).

Figure 10-44
Nugget du modèle discriminant, onglet Récapitulatif



Noeud Modèles linéaires généralisés

Le modèle linéaire généralisé développe le modèle linéaire général, et la variable dépendante est ainsi reliée aux facteurs et aux covariables via une fonction de lien spécifiée. En outre, le modèle offre à la variable dépendante une distribution non normale. Il couvre des modèles statistiques largement utilisés comme la régression linéaire pour les réponses normalement distribuées, les modèles logistiques pour les données binaires, les modèles log-linéaires pour les données de décompte, les modèles log-log complémentaires pour les données de survie avec censure par intervalle, ainsi que de nombreux autres modèles statistiques via sa formulation de modèles très générale.

Exemples : Une compagnie de navigation peut utiliser des modèles linéaires généralisés pour ajuster une régression de Poisson au nombre de détériorations subies par plusieurs types de bateaux construits à des périodes différentes. Le modèle qui en résulte peut permettre de déterminer quels types de bateaux sont plus enclins aux détériorations. [Pour plus d'informations, reportez-vous à la section Utilisation de la régression de Poisson pour analyser les taux de dommage aux navires \(modèles linéaires généralisés\) dans le chapitre 23 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Une compagnie d'assurance automobile peut utiliser des modèles linéaires généralisés pour ajuster une régression gamma à des actions en indemnisation pour des voitures. Le modèle qui en résulte peut permettre de déterminer les facteurs qui ont le plus d'influence sur le nombre de déclarations. [Pour plus d'informations, reportez-vous à la section Ajustement d'une régression gamma à des](#)

déclarations de sinistre automobile (modèles linéaires généralisés) dans le chapitre 24 dans *Guide d'applications de IBM SPSS Modeler 15*.

Des chercheurs en médecine peuvent utiliser des modèles linéaires généralisés pour ajuster une régression log-log complémentaire des données de survie censurées par intervalle pour prévoir la récurrence d'un état médical. [Pour plus d'informations, reportez-vous à la section Analyse de données de survie avec censure par intervalle \(modèles linéaires généralisés\) dans le chapitre 22 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Les modèles linéaires généralisés créent une équation qui lie les valeurs de champ d'entrée aux valeurs de champ de sortie. Une fois le modèle créé, il peut être utilisé pour générer des valeurs pour de nouvelles données. Pour chaque enregistrement, une probabilité d'appartenance est calculée pour chaque catégorie de sortie. La catégorie cible présentant la plus forte probabilité devient la valeur de sortie prédite de l'enregistrement.

Conditions requises. Vous avez besoin d'un ou de plusieurs champs d'entrée ou d'un champ cible (qui peut avoir un niveau de mesure *Continu* ou *Booléen*) doté de plusieurs catégories. Les types des champs utilisés dans ce modèle doivent être complètement instanciés.

Puissance. Le modèle linéaire généralisé est extrêmement souple, mais le processus consistant à choisir la structure du modèle n'est pas automatisé et requiert donc une connaissance des données plus importante qu'avec les algorithmes de type « boîte noire ».

Noeud Modèles linéaires généralisés - Options de champs

Figure 10-45

Boîte de dialogue du noeud Modèles linéaires généralisés - Onglet Champs

Outre les options personnalisées de cible, d'entrée et de partition généralement proposées dans les onglets Champs du noeud de modélisation (reportez-vous à [Options de champs des noeuds de modélisation](#) sur p. 39), le noeud Modèles linéaires généralisés offre les fonctionnalités suivantes.

Utiliser le champ de pondération. Le paramètre d'échelle est un paramètre de modèle estimé lié à la variance de la réponse. Les pondérations d'échelle sont des « valeurs connues », susceptibles de varier d'une observation à l'autre. Si la variable de pondération d'échelle est spécifiée, le paramètre d'échelle, qui est lié à la variance de la réponse, est divisé par cette variable pour chaque observation. Les enregistrements dont les valeurs pondérées sont inférieures ou égales à 0, ou sont manquantes, ne sont pas utilisés dans l'analyse.

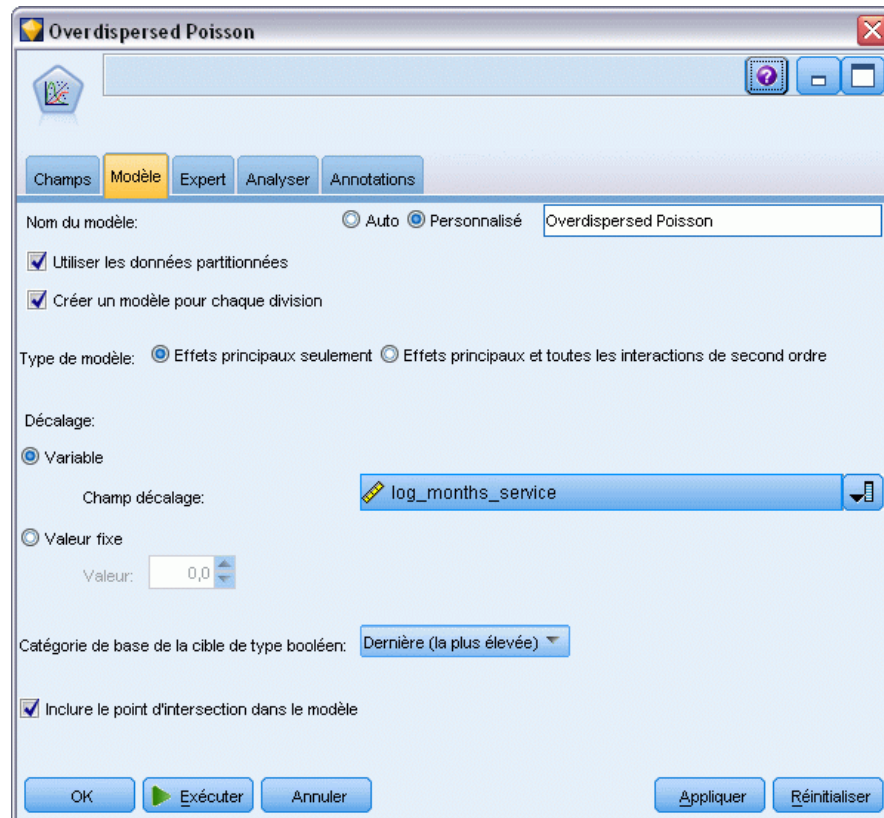
Le champ cible représente le nombre d'événements se produisant dans un ensemble d'essais. Lorsque la réponse est le nombre d'événements qui se produisent dans un ensemble d'essais, le champ cible contient le nombre d'événements. Vous pouvez sélectionner une variable supplémentaire qui contient le nombre d'essais. Si le nombre de tentatives est identique dans tous les sujets, vous pouvez également indiquer les tentatives à l'aide d'une valeur fixe. Le nombre d'essais doit être

supérieur ou égal au nombre d'événements pour chaque enregistrement. Les événements doivent être des entiers non négatifs et les tentatives, des entiers positifs.

Noeud Modèles linéaires généralisés - Options du modèle

Figure 10-46

Boîte de dialogue du noeud Modèles linéaires généralisés - 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Type de modèle. Il existe deux options pour le type de modèle à créer. Effets principaux uniquement permet au modèle d'inclure uniquement les champs d'entrée individuellement et ne teste pas les interactions (effets multiplicateurs) entre les champs d'entrée. Effets principaux et toutes les

interactions de second ordre inclut toutes les interactions de second ordre ainsi que les effets principaux du champ d'entrée.

Décalage. Le terme de décalage est une valeur prédite « structurelle ». Son coefficient n'est pas estimé par le modèle mais sa valeur est supposée être 1. Par conséquent, les valeurs du décalage sont simplement ajoutées à la valeur prédite linéaire de la cible. Cela s'avère particulièrement utile dans les modèles de régression de Poisson, dans lesquels chaque observation peut avoir différents niveaux d'exposition à l'événement étudié.

Par exemple, lors de la modélisation des taux d'accident pour des conducteurs, la différence est importante entre un conducteur avec trois ans de conduite qui est en tort dans un accident et un autre conducteur n'ayant été en tort qu'une fois en 25 ans ! Le nombre d'accidents peut être modélisé comme une réponse de type Poisson ou binominale négative avec un lien de journal si le logarithme naturel de l'expérience du lecteur est inclus comme un terme de décalage.

D'autres combinaisons de distribution et types de liens nécessiteraient d'autres transformations de la variable de décalage.

Remarque : Si un champ de décalage de variables est utilisé, le champ spécifié ne doit pas être également utilisé comme un champ d'entrée. Paramétrez le rôle du champ de décalage sur Aucun dans une source en amont ou dans le noeud Typer si nécessaire. [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*.](#)

Catégorie de base de la cible de type booléen.

Pour une réponse binaire, vous pouvez choisir la catégorie de référence pour la variable dépendante. Cela peut avoir des conséquences sur certaines sorties, comme les estimations de paramètre et les valeurs enregistrées, mais ne devrait pas modifier l'ajustement du modèle. Par exemple, si votre réponse binaire prend les valeurs 0 et 1 :

- Par défaut, la procédure fait de la dernière catégorie (ayant la valeur la plus élevée), ou 1, la catégorie de référence. Dans cette situation, les probabilités enregistrées dans le modèle évaluent la possibilité qu'une observation donnée prenne la valeur 0 et les estimations de paramètre doivent être interprétées comme étant liées à la vraisemblance de la catégorie 0.
- Si vous indiquez la première catégorie (ayant la valeur la moins élevée), ou 0, comme catégorie de référence, les probabilités enregistrées dans le modèle évaluent la probabilité qu'une observation donnée prenne la valeur 1.
- Si vous indiquez la catégorie personnalisée et que les étiquettes de votre variable sont définies, vous pouvez paramétrer la catégorie de référence en choisissant une valeur dans la liste. Cela peut s'avérer pratique lorsque, au cours de la spécification d'un modèle, vous ne vous rappelez pas du codage exact d'une variable spécifique.

Inclure une constante au modèle : L'ordonnée est généralement incluse dans le modèle. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante.

Noeud Modèles linéaires généralisés - Options expert

Si vous êtes familier des modèles linéaires généralisés, utilisez les options expert pour affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 10-47

Boîte de dialogue du noeud Modèles linéaires généralisés - Onglet Expert

The screenshot shows the 'Overdispersed Poisson' dialog box with the 'Expert' tab selected. The 'Mode' is set to 'Expert'. The 'Proportion' is set to 'Poisson' and the 'Fonction de lien' is set to 'Log'. The 'Paramètres' section includes 'Paramètre pour un binomial négatif' with 'Spécifier la valeur' selected and a value of 1,0, and 'Paramètre de Tweedie' with a value of 1,5. The 'Fonction de lien' is 'Log' with a 'Puissance' of 0,0. Below, it states 'Les paramètres de méthode et d'itération ne sont pas disponibles si la Proportion = Normale et Lien Fonction = Identité.' The 'Estimation des paramètres' section shows 'Méthode' set to 'Hybride', 'Méthode de paramètre d'échelle' set to 'Pearson du Chi-deux', and 'Matrice de covariances' set to 'Estimateur basé sur le modèle'. The 'Tolérance de singularité' is set to '1E-012' and 'Ordre des valeurs pour les entrées catégorielles' is set to 'Décroissant'. Buttons for 'Itérations...', 'Sortie...', 'OK', 'Exécuter', 'Annuler', 'Appliquer', and 'Réinitialiser' are visible.

Proportion appliquée au champ cible et fonction de lien

Distribution :

Cette sélection indique la proportion de la variable dépendante. La possibilité d'indiquer une proportion non normale et une fonction de lien rattaché à aucune identité constitue l'amélioration essentielle du modèle linéaire généralisé par rapport au modèle linéaire général. Il existe plusieurs combinaisons de fonctions lien-proportion possibles. Parmi elles, plusieurs peuvent convenir à un ensemble de données spécifique ; votre choix peut donc être guidé par des considérations théoriques a priori ou par les combinaisons qui semblent le mieux adaptées.

- **Binomial.** Cette proportion ne convient qu'aux variables qui représentent une réponse binaire ou un nombre d'événements.
- **Gamma.** Cette proportion ne convient qu'aux variables présentant des valeurs d'échelle positives biaisées vers des valeurs positives plus élevées. Si une valeur de données est inférieure ou égale à 0, ou bien manquante, l'observation correspondante n'est pas utilisée dans l'analyse.
- **Gaussienne inverse.** Cette proportion ne convient qu'aux variables présentant des valeurs d'échelle positives biaisées vers des valeurs positives plus élevées. Si une valeur de données est inférieure ou égale à 0, ou bien manquante, l'observation correspondante n'est pas utilisée dans l'analyse.
- **Binomiale négative.** Cette proportion peut être envisagée comme étant le nombre d'essais requis pour observer k succès ; elle convient aux variables présentant des valeurs d'entier non négatif. Si une valeur de données n'est pas un nombre entier, est inférieure à 0, ou bien manquante, l'observation correspondante n'est pas utilisée dans l'analyse. La valeur fixe du paramètre secondaire de la loi binomiale négative peut être tout nombre supérieur ou égal à 0. Lorsque le paramètre secondaire est défini à 0, l'utilisation de cette proportion équivaut à utiliser la distribution de Poisson.
- **Normal.** Cette proportion convient aux variables d'échelle dont les valeurs prennent la forme d'une proportion symétrique, en cloche sur une valeur (moyenne) centrale. La variable dépendante doit être numérique.
- **Poisson.** Cette proportion peut être envisagée comme étant le nombre d'occurrences d'un événement d'intérêt dans une période de temps fixe ; elle convient aux variables présentant des valeurs d'entier non négatif. Si une valeur de données n'est pas un nombre entier, est inférieure à 0, ou bien manquante, l'observation correspondante n'est pas utilisée dans l'analyse.
- **Tweedie.** Cette proportion ne convient qu'aux variables représentées par les mélanges Poisson de distributions gamma ; la distribution est "mêlée" dans le sens où elle combine les propriétés des distributions continues (elle prend des valeurs réelles non négatives) et discrètes (masse de probabilité positive à une seule valeur, 0). La variable dépendante doit être numérique, avec des valeurs de données supérieures ou égales à zéro. Si une valeur de données est inférieure à 0, ou bien manquante, alors l'observation correspondante n'est pas utilisée dans l'analyse. La valeur fixe du paramètre de distribution Tweedie peut être tout nombre supérieur à un et inférieur à deux.
- **Multinomial.** Cette proportion ne convient qu'aux variables qui représentent une réponse ordinale. La variable dépendante peut être un nombre ou une chaîne de caractères, et doit avoir au moins deux valeurs de données valides distinctes.

Fonctions de lien.

La fonction de lien est une transformation de la variable dépendante qui permet l'estimation du modèle. Les fonctions suivantes sont disponibles :

- **Identité.** $f(x)=x$. La variable dépendante n'est pas transformée. Ce lien peut être utilisé avec n'importe quelle distribution.
- **Log-log complémentaire.** $f(x)=\log(-\log(1-x))$. Ceci n'est approprié qu'avec la loi binomiale.
- **Cauchit cumulatif.** $f(x) = \tan(\pi(x - 0,5))$, appliqué à la probabilité cumulative de chaque catégorie de la réponse. Ceci n'est approprié qu'avec la loi multinomiale.

- **Log-log complémentaire cumulatif.** $f(x) = \ln(-\ln(1-x))$, appliqué à la probabilité cumulative de chaque catégorie de la réponse. Ceci n'est approprié qu'avec la loi multinomiale.
- **Logit cumulatif.** $f(x) = \ln(x / (1-x))$, appliqué à la probabilité cumulative de chaque catégorie de la réponse. Ceci n'est approprié qu'avec la loi multinomiale.
- **Log-log négatif cumulatif.** $f(x) = -\ln(-\ln(x))$, appliqué à la probabilité cumulative de chaque catégorie de la réponse. Ceci n'est approprié qu'avec la loi multinomiale.
- **Probit cumulatif.** $f(x) = \Phi^{-1}(x)$, appliqué à la probabilité cumulative de chaque catégorie de la réponse, où Φ^{-1} est la fonction de distribution cumulée normale standard inverse. Ceci n'est approprié qu'avec la loi multinomiale.
- **Log.** $f(x) = \log(x)$. Ce lien peut être utilisé avec n'importe quelle distribution.
- **Complément log.** $f(x) = \log(1-x)$. Ceci n'est approprié qu'avec la loi binomiale.
- **Logit.** $f(x) = \log(x / (1-x))$. Ceci n'est approprié qu'avec la loi binomiale.
- **Binomiale négative.** $f(x) = \log(x / (x+k^{-1}))$, où k est le paramètre secondaire de la loi binomiale négative. Ceci n'est approprié qu'avec la loi binomiale négative.
- **Log-log négatif.** $f(x) = -\log(-\log(x))$. Ceci n'est approprié qu'avec la loi binomiale.
- **Puissance des cotes.** $f(x) = [(x/(1-x))^\alpha - 1] / \alpha$, si $\alpha \neq 0$. $f(x) = \log(x)$, si $\alpha = 0$. α est la spécification numérique requise et doit être un nombre réel. Ceci n'est approprié qu'avec la loi binomiale.
- **Probit.** $f(x) = \Phi^{-1}(x)$, où Φ^{-1} est la fonction de distribution cumulée normale standard inverse. Ceci n'est approprié qu'avec la loi binomiale.
- **Puissance.** $f(x) = x^\alpha$, si $\alpha \neq 0$. $f(x) = \log(x)$, si $\alpha = 0$. α est la spécification numérique requise et doit être un nombre réel. Ce lien peut être utilisé avec n'importe quelle distribution.

Paramètres. Les commandes de ce groupe vous permettent de spécifier des valeurs de paramètres si certaines options de distribution sont sélectionnées.

- **Paramètre pour un binomial négatif.** Pour une distribution binomiale négative, choisissez de spécifier une valeur ou d'autoriser le système à fournir une valeur estimée.
- **Paramètre pour Tweedie.** Pour la distribution Tweedie, spécifiez un nombre compris entre 1,0 et 2,0 pour la valeur fixe.

Estimation des paramètres. Les commandes de ce groupe vous permettent d'indiquer des méthodes d'estimation et de fournir des valeurs initiales pour les estimations de paramètre.

- **Méthode.** Vous pouvez sélectionner une méthode d'estimation de paramètre. Choisissez la méthode Newton-Raphson, scoring de Fisher ou une méthode hybride dans laquelle les itérations de scoring de Fisher sont effectuées avant de passer à la méthode Newton-Raphson. Si la convergence est obtenue au cours de la phase de scoring de Fisher de la méthode hybride avant que le nombre maximal d'itérations Fischer soit atteint, l'algorithme poursuit avec la méthode Newton-Raphson.
- **Méthode de paramètre d'échelle.** Vous pouvez sélectionner la méthode d'estimation de paramètre d'échelle. Le maximum de vraisemblance estime conjointement le paramètre d'échelle et les effets de modèle ; notez que cette option n'est pas valide si la réponse comporte une loi binomiale négative, de Poisson ou binomiale. Les options de déviance et de

Pearson du Chi-deux estiment le paramètre d'échelle à partir de la valeur de ces statistiques. Vous pouvez aussi indiquer une valeur fixe pour le paramètre d'échelle.

- **Matrice de covariance.** L'estimateur basé sur un modèle est le négatif de l'inverse généralisé de la matrice de Hess. L'estimateur fiable (ou Huber/White/sandwich) est un estimateur « corrigé » basé sur un modèle qui fournit une estimation cohérente de la covariance, même lorsque les spécifications de la variance et des fonctions de lien est incorrecte.

Itérations : Ces options vous permettent de contrôler les paramètres de convergence des modèles. [Pour plus d'informations, reportez-vous à la section Modèles linéaires généralisés - Itérations sur p. 335.](#)

Résultats. Ces options vous permettent de demander des statistiques supplémentaires, qui apparaîtront dans les sorties avancées du nugget du modèle créé par le noeud. [Pour plus d'informations, reportez-vous à la section Modèles linéaires généralisés - Sorties avancées sur p. 337.](#)

Tolérance de singularité. Les matrices singulières (ou non réversibles) présentent des colonnes dépendantes linéairement, ce qui peut provoquer de graves problèmes pour l'algorithme d'estimation. Les matrices quasi-singulières étant également susceptibles de générer de faibles résultats, la procédure traitera une matrice dont le déterminant est inférieur à la tolérance singulière. Indiquez une valeur positive.

Modèles linéaires généralisés - Itérations

Vous pouvez définir les paramètres de convergence pour l'estimation du modèle linéaire généralisé.

Figure 10-48
Modèles linéaires généralisés - Options d'itération

Modèles linéaires généralisés : Itérations

Itérations maximales: 100

Découpage maximal en deux des étapes: 5

Rechercher la séparation des points de données (seulement les cibles de type booléen)

Démarrage de l'itération: 20

Critères de convergence

Au moins un critère de convergence doit être spécifié avec un minimum supérieur à 0.

	Minimum :	Type
<input checked="" type="checkbox"/> Modification dans les estimations des paramètres	0,0000010000	Absolus
<input type="checkbox"/> Modification dans le log de vraisemblance	0,0000010000	Absolus
<input type="checkbox"/> Convergence hessienne	0,0000010000	Absolus

OK Annuler Aide

Itérations.

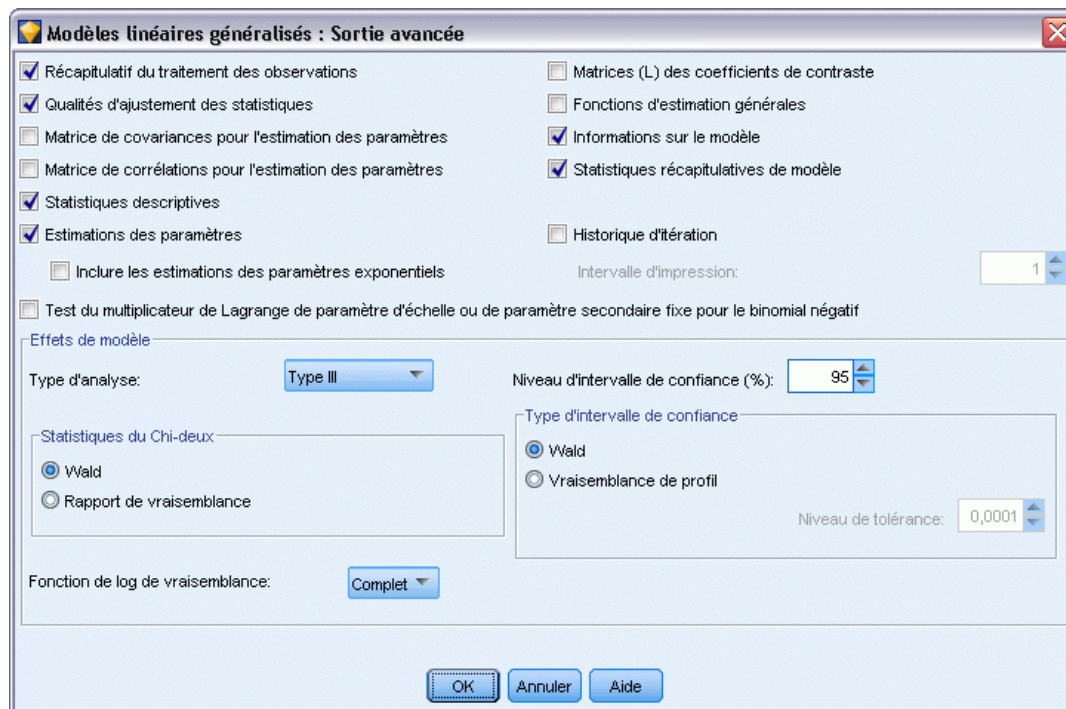
- **Itérations maximales.** Nombre maximal d'itérations que l'algorithme exécutera. Spécifiez un nombre entier non négatif.
- **Découpage maximal en deux des étapes.** A chaque itération, la taille de pas est réduite d'un facteur de 0,5 jusqu'à ce que le log de vraisemblance augmente ou qu'un découpage maximal en deux des étapes soit atteint. Indiquez un entier positif.
- **Rechercher une séparation des points de données.** Lorsque cette option est sélectionnée, l'algorithme effectue des tests pour vérifier que les estimations des paramètres ont des valeurs uniques. La séparation intervient lorsque la procédure peut produire un modèle qui classe correctement chaque observation. Cette option est disponible pour des réponses binomiales avec un format binaire.

Critères de convergence.

- **Convergence des paramètres.** Lorsque cette option est sélectionnée, l'algorithme s'arrête après une itération dans laquelle la modification absolue ou relative des estimations de paramètres est inférieure à la valeur indiquée, qui doit être positive.
- **Convergence du log de vraisemblance.** Lorsque cette option est sélectionnée, l'algorithme s'arrête après une itération dans laquelle la modification absolue ou relative de la fonction du log de vraisemblance est inférieure à la valeur indiquée, qui doit être positive.
- **Convergence hessienne.** Pour la spécification absolue, la convergence est supposée si une statistique basée sur la convergence hessienne est inférieure à la valeur positive indiquée. Pour la spécification relative, la convergence est supposée si la statistique est inférieure au produit de la valeur positive indiquée et de la valeur absolue du log de vraisemblance.

Modèles linéaires généralisés - Sorties avancées

Figure 10-49
Modèles linéaires généralisés - Options de sortie avancées



Sélectionnez le résultat facultatif à afficher dans la sortie avancée du nugget du modèle linéaire généralisé. Pour visualiser les sorties avancées, parcourez le nugget du modèle et cliquez sur l'onglet Options avancées. [Pour plus d'informations, reportez-vous à la section Nugget de modèle Modèles linéaires généralisés - Sorties avancées sur p. 340.](#)

La sortie suivante est disponible :

- **Récapitulatif du traitement des observations.** Affiche le nombre et le pourcentage d'observations incluses dans l'analyse et exclues de celle-ci, et le tableau récapitulatif des données corrélées.
- **Statistiques descriptives.** Affiche les statistiques descriptives et les informations récapitulatives concernant la variable dépendante, les covariables et les facteurs.
- **Informations sur le modèle.** Affiche le nom de l'ensemble de données, la variable dépendante ou les variables d'événement et d'essai, la variable de décalage, la variable de pondération, la proportion des probabilités et la fonction de lien.
- **Qualités d'ajustement des statistiques.** Affiche la déviance et la déviance mise à l'échelle, le Pearson du Chi-deux et le Pearson du Chi-deux mis à l'échelle, le log de vraisemblance, le critère d'Akaike (AIC), le critère d'Akaike corrigé pour un échantillon fini (AICC), le critère de Bayes (BIC) et le critère d'Akaike cohérent (CAIC).
- **Statistiques descriptives du modèle.** Affiche les tests de qualité d'ajustement, y compris les statistiques de rapport de vraisemblance pour le test composite de qualité d'ajustement et les statistiques des contrastes de type I ou III pour chaque effet.

- **Estimations des paramètres.** Affiche les estimations des paramètres, ainsi que les statistiques et les intervalles de confiance des tests correspondants. Outre les estimations de paramètre brutes, vous pouvez également afficher les estimations de paramètre élevées à une puissance.
- **Matrice de covariances pour l'estimation des paramètres.** Affiche la matrice de covariances du paramètre estimé.
- **Matrice de corrélations pour l'estimation des paramètres.** Affiche la matrice de corrélations des paramètres estimés.
- **Matrices (L) des coefficients de contraste.** Affiche les coefficients de contraste pour les effets par défaut et pour les moyennes marginales estimées, si cela est requis au niveau de l'onglet Moyennes EM.
- **Fonctions d'estimation générales.** Affiche les matrices pour la génération des matrices (L) des coefficients de contraste.
- **Historique d'itération.** Affiche l'historique d'itération des estimations de paramètre, ainsi que le log de vraisemblance, et imprime la dernière évaluation du vecteur gradient et de la matrice hessienne. La table d'historique d'itération affiche les estimations de paramètre toutes les $n^{\text{ièmes}}$ itérations, en commençant par l'itération 0 (les estimations initiales), n étant la valeur de l'intervalle d'impression. Si l'historique d'itération est demandé, la dernière itération est toujours affichée quelle que soit la valeur de n .
- **Test multiplicateur Lagrange.** Affiche les statistiques obtenues par le test du multiplicateur de Lagrange afin d'évaluer la validité d'un paramètre d'échelle calculé à l'aide de la déviance ou des statistiques Pearson du Chi-deux, ou défini comme un nombre fixe, pour les courbes de fréquence de la loi normale et de la loi gamma et pour la distribution gaussienne inverse. Pour la distribution binomiale négative, cette option teste le paramètre secondaire fixe.

Effets de modèle.

- **Type d'analyse.** Spécifiez le type d'analyse à générer. Une analyse de type I est généralement appropriée lorsque vous avez des raisons a priori d'organiser des valeurs prédites dans le modèle, tandis que le type III est applicable de manière plus générale. Les statistiques de Wald ou de rapport de vraisemblance sont calculées en fonction de la sélection dans le groupe Statistiques du Chi-deux.
- **Intervalles de confiance.** Spécifiez un niveau de confiance supérieur à 50 et inférieur à 100. Les intervalles de Wald sont basés sur l'hypothèse selon laquelle les paramètres ont une distribution normale asymptotique ; les intervalles par vraisemblance de profil sont plus précis mais peuvent être coûteux en calcul. Le niveau de tolérance des intervalles par vraisemblance de profil est le critère utilisé pour interrompre l'algorithme itératif utilisé pour calculer les intervalles.
- **Fonction de log de vraisemblance.** Cette option permet de contrôler le format d'affichage de la fonction du log de vraisemblance. La fonction complète inclut un terme supplémentaire qui est constant par rapport aux estimations des paramètres ; elle n'a aucun effet sur l'estimation des paramètres et est exclue de l'affichage dans certains logiciels.

Modèle de nugget Modèles linéaires généralisés

Les modèles de nugget Modèles linéaires généralisés représentent les équations estimées par un noeud Modèles linéaires généralisés. Ils contiennent toutes les informations rassemblées par le modèle, ainsi que des informations sur la structure et les performances du modèle.

Lorsque vous exécutez un flux qui contient un nugget de modèle Modèles linéaires généralisés, le noeud ajoute de nouveaux champs dont le contenu dépend de la nature du champ cible :

- **Champ cible de type booléen.** Ajoute des champs contenant la catégorie prédite et la probabilité associée, ainsi que les probabilités de chaque catégorie. Les noms des deux premiers champs sont formés du nom du champ de sortie sur lequel porte la prévision, auquel est ajouté le préfixe *\$G-* pour désigner la catégorie de prévision et *\$GP-* pour désigner le champ de la probabilité associée. Par exemple, si un champ de sortie est intitulé *défaut*, les nouveaux champs seront alors intitulés *\$G-défaut* et *\$GP-défaut*. Les noms de ces deux derniers champs supplémentaires sont basés sur les valeurs du champ de sortie auxquelles le préfixe *\$GP-* est ajouté. Par exemple, si les valeurs valides du champ *défaut* sont *Oui* et *Non*, les nouveaux champs sont nommés *\$GP-Oui* et *\$GP-Non*.
- **Cible continue.** Ajoute des champs contenant la moyenne prédite et l'erreur standard.
- **Cible continue, indiquant le nombre d'événements d'une série d'essais.** Ajoute des champs contenant la moyenne prédite et l'erreur standard.

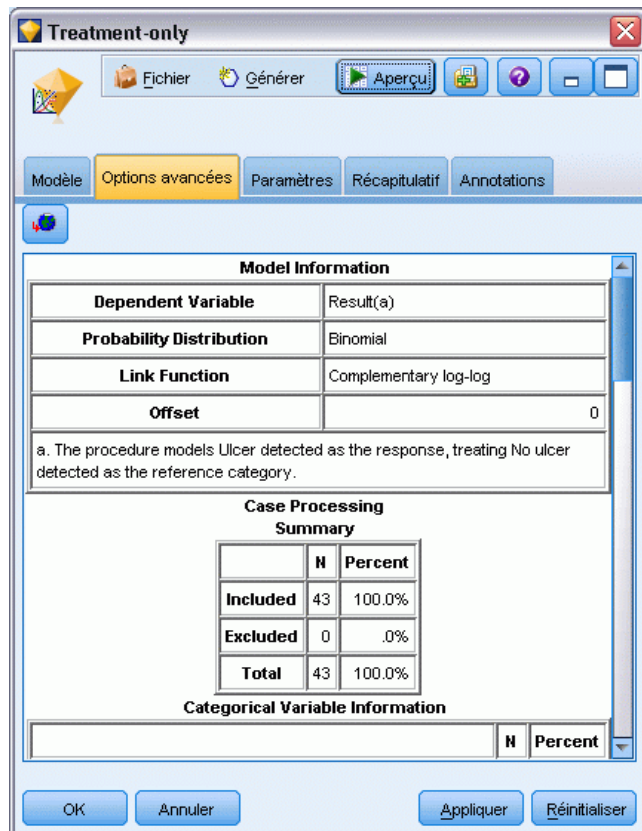
Génération d'un noeud Filtrer. Le menu Générer vous permet de créer un noeud Filtrer pour transmettre des champs d'entrée en fonction des résultats du modèle.

Importance des valeurs prédites

Un graphique illustrant l'importance relative de chaque valeur prédite dans l'estimation du modèle peut également être affiché dans l'onglet Modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique n'est disponible que si Calculer l'importance des valeurs prédites a été sélectionné dans l'onglet Analyse avant la génération du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Nugget de modèle Modèles linéaires généralisés - Sorties avancées

Figure 10-50
Nugget de modèle Modèles linéaires généralisés - Onglet Options avancées

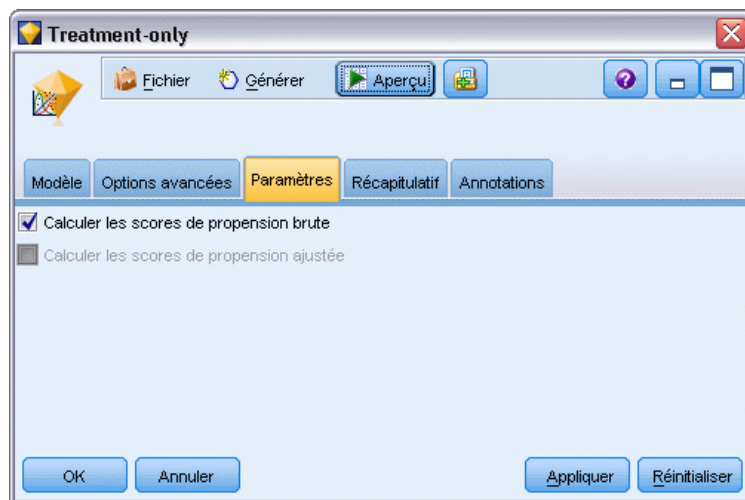


La sortie avancée du modèle linéaire généralisé vous fournit des informations détaillées sur le modèle évalué et ses performances. Dans la mesure où la plupart des informations contenues dans la sortie avancée ont un caractère technique, il est nécessaire d'avoir une bonne connaissance de ce type d'analyse pour pouvoir interpréter cette sortie correctement. [Pour plus d'informations, reportez-vous à la section Modèles linéaires généralisés - Sorties avancées sur p. 337.](#)

Nugget de modèle Modèles linéaires généralisés - Paramètres

L'onglet Paramètres d'un nugget de modèle Modèles linéaires généralisés vous permet d'obtenir les scores de propension lors de l'évaluation du modèle. Cet onglet est disponible uniquement pour les modèles avec des cibles de type booléen, et seulement après que le nugget de modèle ait été ajouté à un flux.

Figure 10-51
Nugget de modèle Modèles linéaires généralisés - Onglet Paramètres pour cibles de type booléen



Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Modèle de nugget Modèles linéaires généralisés - Récapitulatif

L'onglet Récapitulatif d'un nugget de modèle Modèles linéaires généralisés affiche les champs et paramètres utilisés pour la génération du modèle. En outre, si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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.](#) Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle sur p. 55.](#)

Figure 10-52
Nugget de modèle Modèles linéaires généralisés - Onglet Récapitulatif



Noeud MMLG

Ce noeud sert à créer un modèle mixte linéaire généralisé (MMLG).

Modèles mixtes linéaires généralisés

Les modèles mixtes linéaires généralisés élargissent le modèle linéaire de sorte que :

- La cible est liée de manière linéaire aux facteurs et covariables par une fonction de lien spécifiée.
- La cible peut avoir une distribution non normale
- Les observations peuvent être corrélées.

Les modèles mixtes linéaires généralisés couvrent une large variété de modèles, depuis les modèles de régression linéaire simple aux modèles multi-niveaux complexes destinés aux données longitudinales non normales.

Exemples : La commission scolaire régionale peut utiliser un modèle mixte linéaire généralisé pour déterminer si une méthode d'apprentissage expérimentale est efficace pour l'amélioration des notes en mathématiques. Les élèves d'une même classe devraient être corrélés puisque le même enseignant leur dispense les cours, et les classes d'une même école devraient aussi être

corrélées donc nous pouvons inclure des effets aléatoires aux niveaux de l'école et de la classe pour prendre en compte les différentes sources de variabilité.

Les chercheurs dans le domaine médical peuvent utiliser un modèle mixte linéaire généralisé pour déterminer si un nouveau médicament anti-convulsions peut réduire le taux de crises épileptiques chez un patient. Les mesures répétées chez le même patient sont normalement corrélées positivement, donc un modèle mixte avec des effets aléatoires devrait être adéquat. Le champ cible, en l'occurrence le nombre de crises, prend des valeurs entières positives, donc un modèle mixte linéaire généralisé avec une distribution de Poisson et un lien log peut être approprié.

Les dirigeants d'une société de prestation de services de télévision par câble, de téléphonie et Internet peuvent utiliser un modèle mixte linéaire généralisé pour mieux connaître leurs clients potentiels. Les réponses possibles ayant des niveaux de mesure nominaux, l'analyste de la société utilise un modèle mixte logit généralisé avec une constante aléatoire, pour capturer la corrélation entre les réponses aux questions sur l'utilisation du service selon le type de service (tv, téléphone, internet) au sein des réponses à une enquête donnée.

Figure 10-53
Onglet Structure des données

Structure de données: Champs et effets Options de construction Options de modèle

De quelle manière vos données sont-elles structurées ?

Cette procédure part du principe que des enregistrements multiples représentent des mesures répétées pour un sujet unique.

Champs: Trier: Aucun

- Center size
- Gender
- Date of birth
- Treatment received

Tous

Plus

Définir les groupes de covariance par:

Type de covariance répété: Autorégressif de premier ordre (AR1)

Canevas:

Sujets			Mesures répétées
Center ID	Attending Physician ID	Patient ID	Week
		Patient ID	Week
	Attending Physician ID	Patient ID	Week
		Patient ID	Week

L'onglet Structure des données vous permet de spécifier les relations structurelles entre les enregistrements de votre ensemble de données, lorsque les observations sont corrélées. Si les enregistrements de l'ensemble de données représentent des observations indépendantes, vous n'avez pas besoin de spécifier quoi que ce soit sur cet onglet.

Sujets. La combinaison des valeurs des champs catégoriels spécifiés doit définir de manière unique les sujets à l'intérieur de l'ensemble de données. Par exemple, un seul champ *ID du patient* devrait suffire à définir les sujets d'un hôpital donné, mais la combinaison de l'*ID de l'hôpital* et de l'*ID du patient* peut être nécessaire si les numéros d'identification des patients ne sont pas uniques entre les hôpitaux. Dans le cas de mesures répétées, plusieurs observations sont enregistrées pour chaque sujet, de sorte que chaque sujet peut occuper plusieurs enregistrements dans l'ensemble de données.

Un **sujet** est une unité d'observation qui peut être considérée indépendante des autres sujets. Par exemple, la mesure de pression artérielle d'un patient au sein d'une étude médicale peut être considérée indépendante des mesures prise sur les autres patients. La définition des sujets devient particulièrement importante lorsqu'il existe deux mesures répétées par sujet et que vous souhaitez modéliser la corrélation entre ces observations. Par exemple, vous pouvez vous attendre à ce que les mesures de la pression artérielle d'un patient donné soient corrélées, lors de visites consécutives chez le médecin.

Tous les champs spécifiés comme Sujets dans l'onglet Structure des données sont utilisés pour définir les sujets dans le cadre de la structure de covariance des résidus, et constituent la liste des champs possibles à utiliser lors de la définition des sujets pour les structures de covariance à effets aléatoires sur le [Bloc des effets aléatoires](#).

Mesures répétées. Les champs spécifiés ici permettent d'identifier les observations répétées. Par exemple, une variable unique *Semaine* peut identifier les 10 semaines d'observation au cours d'une étude médicale, ou les variables *Mois* et *Jour* peuvent être utilisés ensemble pour identifier les observations quotidiennes sur une période d'un an.

Définir les classes de covariance par. Les champs spécifiés ici définissent des ensembles indépendants de paramètres de covariance d'effets répétés ; un pour chaque catégorie définie par la classification croisée des champs de regroupement. Tous les sujets ont le même type de covariance ; les sujets faisant partie du même regroupement de covariance auront les mêmes valeurs de paramètres.

Type de covariance répétée. Spécifie la structure de covariance des résidus. Les structures disponibles sont les suivantes :

- Autorégressive de premier ordre (AR1)
- Moyenne mobile autorégressive (1,1) (ARMA11)
- Symétrie composée
- Diagonale
- Identité mise à l'échelle
- Toeplitz
- Sans structure
- Composantes de variance

Cible

Figure 10-55
Paramètres de la cible

Ces paramètres définissent la cible, sa distribution et sa relation avec les variables prédites via la fonction de lien.

Cible. La cible est requise. Elle peut avoir n'importe quel niveau de mesure, et le niveau de mesure de la cible limite les distributions et fonctions de lien appropriées.

- **Utiliser le nombre d'essais comme dénominateur.** Lorsque la réponse est le nombre d'événements qui se produisent dans un ensemble d'essais, le champ cible contient le nombre d'événements et vous pouvez sélectionner un champ supplémentaire qui 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. Si le nombre de fourmis est identique dans tous les échantillons, alors le nombre d'essais peut être indiqué à l'aide d'une valeur fixe.

Le nombre d'essais doit être supérieur ou égal au nombre d'événements pour chaque enregistrement. Les événements doivent être des entiers non négatifs et les essais des entiers positifs.

- **Personnaliser la catégorie de référence.** Pour une cible catégorielle, vous pouvez choisir la catégorie de référence. Cela peut avoir des conséquences sur certains résultats, comme les estimations de paramètre, mais ne devrait pas modifier l'ajustement du modèle. Par exemple, si votre cible prend les valeurs 0, 1 et 2, par défaut, la procédure fait de la dernière catégorie (la plus élevée), soit 2, la catégorie de référence. Dans cette situation, les estimations de paramètres doivent être interprétées comme étant relatives à la vraisemblance de la catégorie 0 ou 1 *par rapport* à la vraisemblance de la catégorie 2. Si vous spécifiez une catégorie personnalisée et que votre cible possède des étiquettes définies, vous pouvez définir la catégorie de référence en choisissant une valeur dans la liste. Ceci peut être pratique, si à mi-chemin de préciser un modèle, vous n'êtes pas sûr de la méthode de codage d'un champ spécifique.

Distribution et relation (lien) de la cible avec le modèle linéaire. D'après les valeurs des variables prédites, le modèle s'attend à ce que la distribution des valeurs de la cible revête la forme spécifiée, et à ce que les valeurs cibles soient liées linéairement aux variables prédites via la fonction de lien spécifiée. Des raccourcis sont fournis pour plusieurs modèles communs, ou vous pouvez choisir un paramètre personnalisé si vous souhaitez ajuster une distribution particulière et une combinaison de fonction de lien qui n'apparaît pas sur la liste des raccourcis.

- **Modèle linéaire.** Spécifie une distribution normale avec un lien d'identité, ce qui est utile lorsque la cible peut être prédite à l'aide d'une régression linéaire ou d'un modèle ANOVA.
- **Régression gamma.** Spécifie une distribution gamma avec un lien log, qui doit être utilisée lorsque la cible ne contient que des valeurs positives et est arrondie à des valeurs supérieures.
- **Log linéaire.** Spécifie une distribution de Poisson avec un lien log, qui doit être utilisée lorsque la cible représente un nombre d'occurrences sur une période fixe.
- **Régression binomiale négative.** Spécifie une distribution binomiale négative avec un lien log, qui doit être utilisée lorsque la cible et le dénominateur représentent le nombre d'essais requis pour observer k succès.
- **Régression logistique multinomiale.** Spécifie une distribution multinomiale qui doit être utilisée lorsque la cible est une réponse à plusieurs catégories. Elle utilise soit un lien logit cumulatif (résultats ordinaux), soit un lien logit généralisé (réponses nominales à plusieurs catégories).
- **Régression logistique binaire.** Spécifie une distribution binomiale avec un lien logit, qui doit être utilisée lorsque la cible est une réponse binaire prédite par un modèle de régression logistique.
- **Probit binaire.** Spécifie une distribution binomiale avec un lien probit, qui doit être utilisée lorsque la cible est une réponse binaire avec une distribution normale sous-jacente.
- **Données de survie censurées par intervalle.** Spécifie une distribution binomiale avec un lien log-log complémentaire, ce qui est utile dans l'analyse de survie lorsque certaines observations n'ont pas d'événement d'arrêt.

Proportion

Cette sélection spécifie la distribution de la cible. La possibilité de spécifier une distribution non normale et une fonction de lien sans identité constitue la principale amélioration du modèle mixte linéaire généralisé par rapport au modèle mixte linéaire. Il existe de nombreuses combinaisons de distribution/fonction de lien possibles, et plusieurs peuvent convenir à un ensemble de données particulier. Votre choix peut être guidé par des considérations théoriques a priori ou en fonction de la combinaison qui vous semble la plus adaptée.

- **Binomiale.** Cette distribution convient uniquement à une cible représentant une réponse binaire ou un nombre d'événements.
- **Gamma.** Cette distribution convient à une cible avec des valeurs d'échelle positives arrondies à des valeurs positives supérieures. Si une valeur de données est inférieure ou égale à 0, ou manquante, l'observation correspondante n'est pas utilisée dans l'analyse.
- **Gaussienne inverse.** Cette distribution convient à une cible avec des valeurs d'échelle positives arrondies à des valeurs positives supérieures. Si une valeur de données est inférieure ou égale à 0, ou manquante, l'observation correspondante n'est pas utilisée dans l'analyse.
- **Multinomiale.** Cette distribution convient à une cible représentant une réponse à plusieurs catégories. La forme du modèle dépendra du niveau de mesure de la cible.

Une cible **nominale** engendrera un modèle multinomial nominal dans lequel un ensemble de paramètres de modèle distinct est estimé pour chaque catégorie de la cible (à l'exception de la catégorie de référence). Les estimations de paramètre d'une variable prédite indiquent la relation entre cette variable prédite et la probabilité de chaque catégorie de la cible, par rapport à la catégorie de référence.

Une cible **ordinaire** engendrera un modèle multinomial ordinal dans lequel le terme de constante traditionnelle est remplacé par un ensemble de paramètres de **seuil** relatif à la probabilité cumulée des catégories de la cible.

- **Binomiale négative.** La régression binomiale négative utilise une distribution binomiale négative avec un lien log, qui doit être utilisée lorsque la cible représente un nombre d'occurrences de variance élevée.
- **Normale.** Cette distribution convient à une cible continue dont les valeurs suivent une distribution symétrique, en cloche, autour d'une valeur centrale (moyenne).
- **Poisson.** Vous pouvez considérer cette distribution comme étant le nombre d'occurrences d'un événement d'intérêt au cours d'une période de temps fixe. Elle convient aux variables présentant des valeurs d'entier non négatif. Si une valeur de données n'est pas un entier, est inférieure à 0 ou est manquante, l'observation correspondante n'est pas utilisée dans l'analyse.

Fonctions de lien

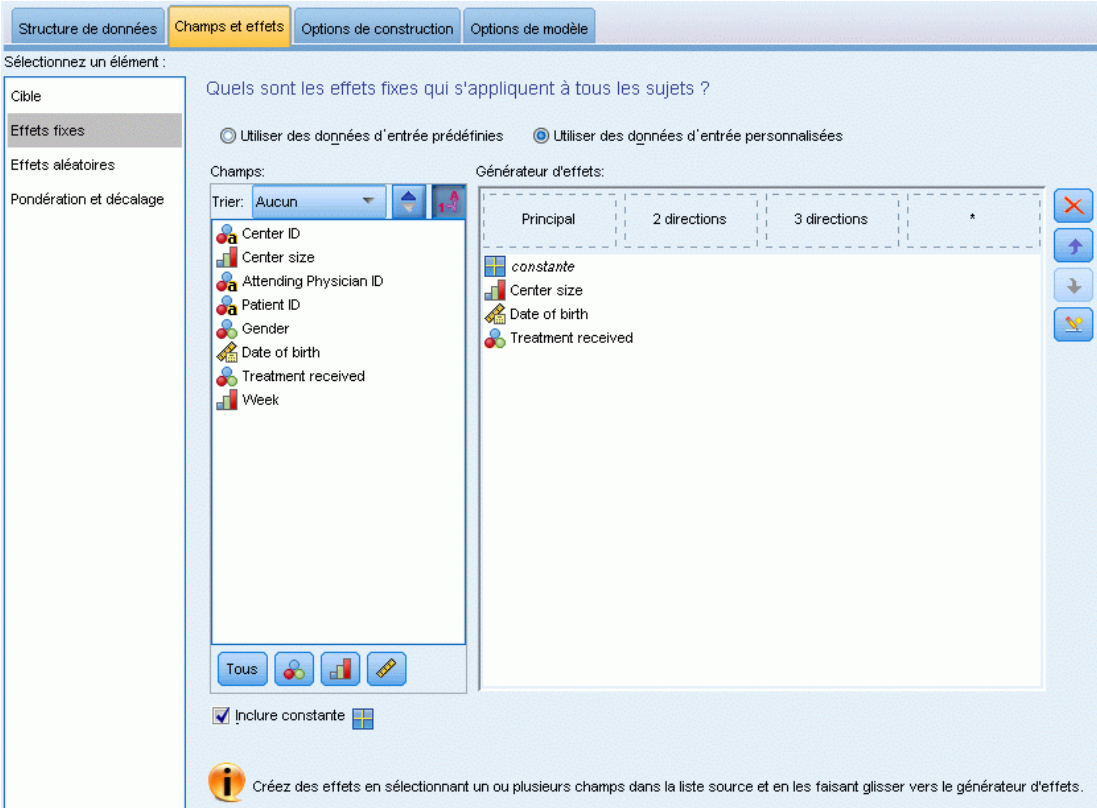
La fonction de lien consiste en une transformation de la cible permettant d'estimer le modèle. Les fonctions suivantes sont disponibles :

- **Identité.** $f(x)=x$. La cible n'est pas transformée. Ce lien peut être utilisé avec n'importe quelle distribution, excepté la distribution multinomiale.
- **Log-log complémentaire.** $f(x)=\log(-\log(1-x))$. Cette fonction convient uniquement à la distribution binomiale ou multinomiale.

- **Cauchit.** $f(x) = \tan(\pi (x - 0.5))$. Cette fonction convient uniquement à la distribution binomiale ou multinomiale.
- **Log.** $f(x) = \log(x)$. Ce lien peut être utilisé avec n'importe quelle distribution, excepté la distribution multinomiale.
- **Log complémentaire.** $f(x) = \log(1-x)$. Cette fonction convient uniquement à la distribution binomiale.
- **Logit.** $f(x) = \log(x / (1-x))$. Cette fonction convient uniquement à la distribution binomiale ou multinomiale.
- **Log-log négatif.** $f(x) = -\log(-\log(x))$. Cette fonction convient uniquement à la distribution binomiale ou multinomiale.
- **Probit.** $f(x) = \Phi^{-1}(x)$, où Φ^{-1} est la fonction de distribution cumulée normale standard inverse. Cette fonction convient uniquement à la distribution binomiale ou multinomiale.
- **Puissance.** $f(x) = x^\alpha$, si $\alpha \neq 0$. $f(x) = \log(x)$, si $\alpha = 0$. α est la spécification numérique requise et doit être un nombre réel. Ce lien peut être utilisé avec n'importe quelle distribution, excepté la distribution multinomiale.

Effets fixes

Figure 10-56
Paramètres des effets fixes



Sélectionnez un élément :

Cible

Effets fixes

Effets aléatoires

Pondération et décalage

Structure de données | **Champs et effets** | Options de construction | Options de modèle

Quels sont les effets fixes qui s'appliquent à tous les sujets ?

Utiliser des données d'entrée prédéfinies Utiliser des données d'entrée personnalisées

Champs: Trier: Aucun

- Center ID
- Center size
- Attending Physician ID
- Patient ID
- Gender
- Date of birth
- Treatment received
- Week

Générateur d'effets:

Principal	2 directions	3 directions	*
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Inclure constante

Créez des effets en sélectionnant un ou plusieurs champs dans la liste source et en les faisant glisser vers le générateur d'effets.

Les facteurs d'effets fixes sont généralement considérés comme des champs dont les valeurs d'intérêt sont toutes représentées dans l'ensemble de données, et pouvant être utilisés pour le scoring. Par défaut, les champs avec un rôle d'entrée prédéfini qui ne sont pas spécifiés ailleurs dans la boîte de dialogue sont entrés dans la partie effets fixes du modèle. Les champs catégoriels (booléens, nominaux et ordinaux) sont utilisés comme facteurs dans le modèle et les champs continus sont utilisés comme covariables.

Entrez les effets dans le modèle en sélectionnant un ou plusieurs champs dans la liste source et en les faisant glisser vers la liste des effets. Le type d'effet créé dépend de l'endroit où vous déposez la sélection.

- **Principaux.** Les champs déposés apparaissent sous forme d'effets principaux distincts au bas de la liste des effets.
- **Bidirectionnels.** Toutes les paires possibles des champs déposés apparaissent sous forme d'interactions bidirectionnelles au bas de la liste des effets.
- **Tridirectionnels.** Tous les triplets possibles des champs déposés apparaissent sous forme d'interactions tridirectionnelles au bas de la liste des effets.
- *****. La combinaison de tous les champs déposés apparaît sous forme d'une unique interaction au bas de la liste des effets.

Les boutons situés à droite du Générateur d'effets vous permettent de :



Supprimer des termes du modèle à effets fixes, en sélectionnant ceux que vous souhaitez supprimer, puis en cliquant sur le bouton de suppression.



Réorganiser les termes dans le modèle à effets fixes, en sélectionnant ceux que vous souhaitez réorganiser, puis en cliquant sur les flèches vers le haut ou vers le bas, et

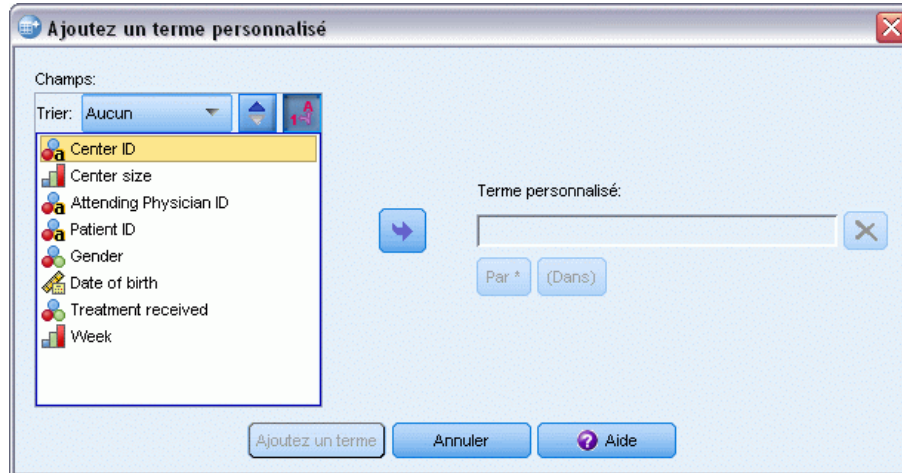


Ajouter des termes imbriqués au modèle à l'aide de la boîte de dialogue [Ajouter un terme personnalisé](#), en cliquant sur le bouton Ajouter un terme personnalisé.

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 10-57
Boîte de dialogue Ajouter un Terme personnalisé



Dans cette procédure, vous pouvez générer des termes imbriqués pour votre modèle. Les termes imbriqués sont utiles pour modéliser l'effet d'un facteur ou d'une covariable dont les valeurs n'interagissent pas avec les niveaux d'un autre facteur. Par exemple, une chaîne d'épiceries peut suivre les habitudes d'achat de ses clients dans divers magasins. Puisque chaque client ne fréquente qu'un seul de ces magasins, l'effet *Client* peut être considéré comme étant **imbriqué dans** l'effet *Emplacement du magasin*.

En outre, vous pouvez inclure des effets d'interaction, tels que des termes polynomiaux impliquant la même covariable, ou ajouter plusieurs niveaux d'imbrication au terme imbriqué.

Limites. Les termes imbriqués comportent les restrictions suivantes :

- Tous les facteurs d'une interaction doivent être uniques. Ainsi, si A est un facteur, la spécification $A * A$ n'est pas valide.
- Tous les facteurs au sein d'un effet imbriqué doivent être uniques. Ainsi, si A est un facteur, la spécification $A(A)$ n'est pas valide.
- Aucun effet ne peut être imbriqué dans une covariable. Ainsi, si A est un facteur et X une covariable, la spécification $A(X)$ n'est pas valide.

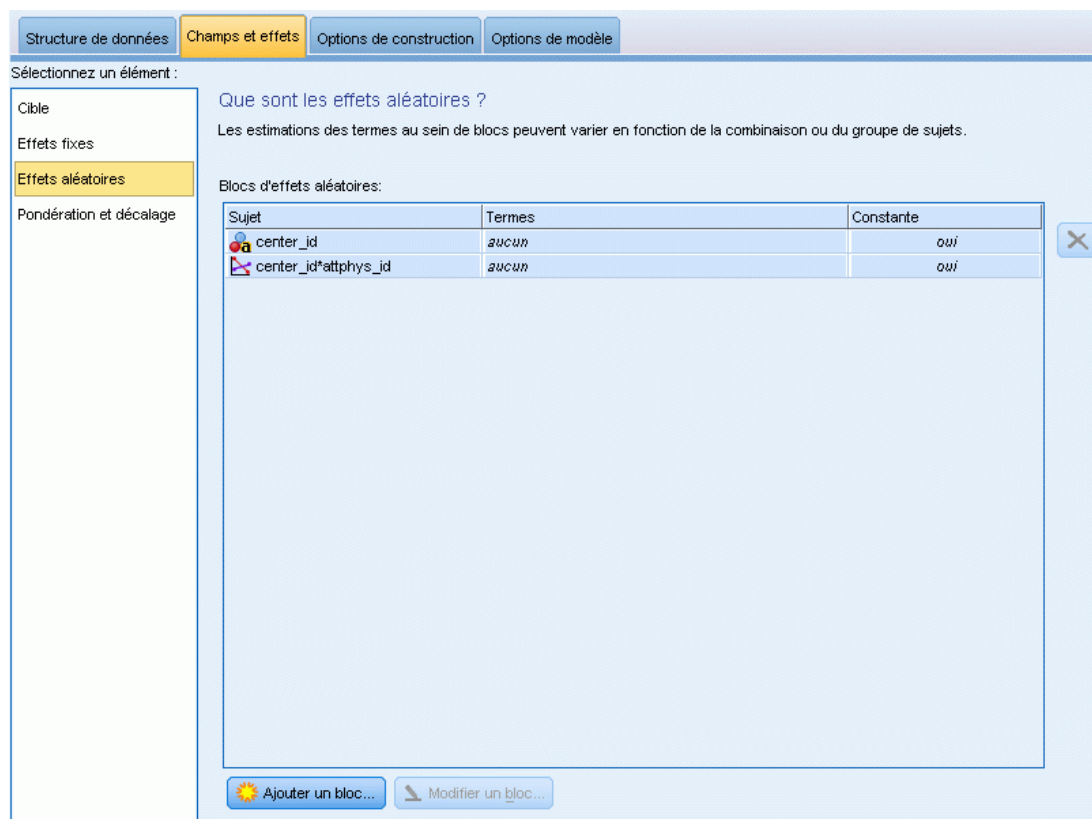
Construction d'un terme imbriqué

- ▶ Sélectionnez une covariable ou un facteur imbriqué dans un autre facteur, puis cliquez sur le bouton fléché.
- ▶ Cliquez sur (Dans).
- ▶ Sélectionnez le facteur dans lequel la covariable ou le facteur précédent est imbriqué, puis cliquez sur le bouton fléché.
- ▶ Cliquez sur Ajouter un terme.

Vous pouvez éventuellement inclure des effets d'interaction ou ajouter plusieurs niveaux d'imbrication au terme imbriqué.

Effets aléatoires

Figure 10-58
Paramètres des effets aléatoires



Les facteurs d'effets aléatoires sont des champs dont les valeurs dans le fichier de données peuvent être considérées comme un échantillon aléatoire d'une population plus large. Ils sont utiles pour expliquer la variabilité des excès dans la cible. Par défaut, si vous avez sélectionné plus d'un sujet dans l'onglet Structure des données, un bloc Effet aléatoire sera créé pour chaque sujet au-delà du sujet le plus interne. Par exemple, si vous avez sélectionné École, Classe et Élève comme sujets sur l'onglet Structure des données, les blocs d'effets aléatoires suivants sont automatiquement créés :

- Effet aléatoire 1 : le sujet est l'école (sans effets, constante uniquement)
- Effet aléatoire 2 : le sujet est l'école * classe (pas d'effets, constante uniquement)

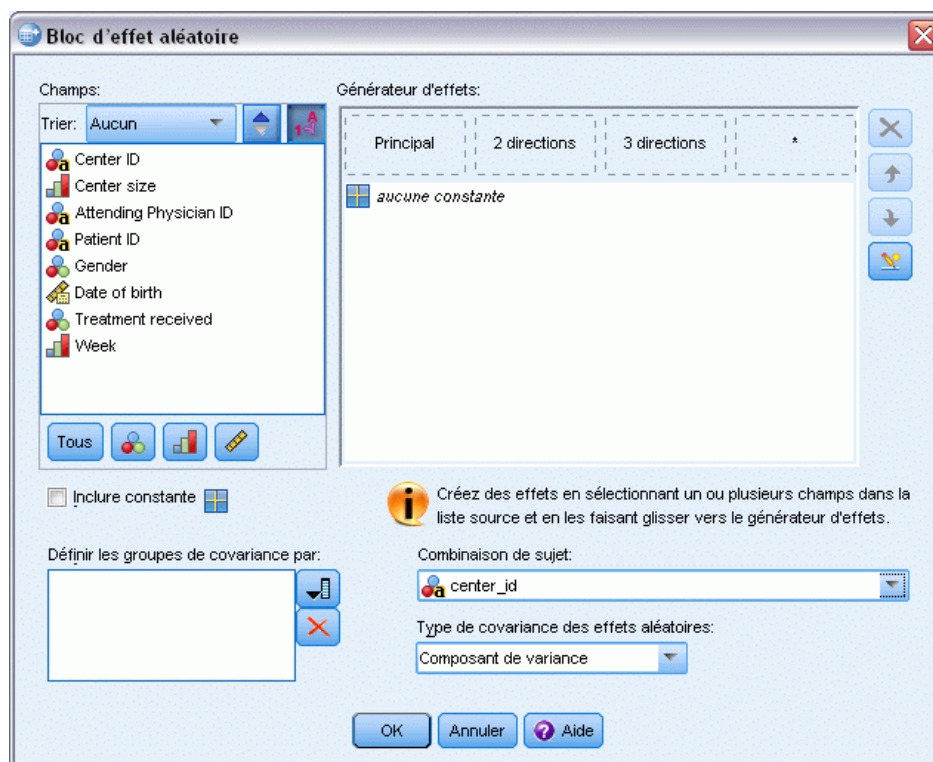
Vous pouvez utiliser les blocs d'effets aléatoires de la manière suivante :

- ▶ Pour ajouter un nouveau bloc, cliquez sur Ajouter un bloc... Cela ouvre la boîte de dialogue [Bloc d'effets aléatoires](#) .
- ▶ Pour modifier un bloc existant, sélectionnez le bloc que vous souhaitez modifier et cliquez sur Modifier le bloc... Cela ouvre la boîte de dialogue [Bloc d'effets aléatoires](#) .

- Pour supprimer un ou plusieurs blocs, sélectionnez les blocs que vous souhaitez supprimer et cliquez sur le bouton de suppression.

Bloc d'effets aléatoires

Figure 10-59
Boîte de dialogue Bloc d'effets aléatoires



Entrez les effets dans le modèle en sélectionnant un ou plusieurs champs dans la liste source et en les faisant glisser vers la liste des effets. Le type d'effet créé dépend de l'endroit où vous déposez la sélection. Les champs catégoriels (booléens, nominaux et ordinaux) sont utilisés comme facteurs dans le modèle et les champs continus sont utilisés comme covariables.

- **Principaux.** Les champs déposés apparaissent sous forme d'effets principaux distincts au bas de la liste des effets.
- **Bidirectionnels.** Toutes les paires possibles des champs déposés apparaissent sous forme d'interactions bidirectionnelles au bas de la liste des effets.
- **Tridirectionnels.** Tous les triplets possibles des champs déposés apparaissent sous forme d'interactions tridirectionnelles au bas de la liste des effets.
- *****. La combinaison de tous les champs déposés apparaît sous forme d'une unique interaction au bas de la liste des effets.

Les boutons situés à droite du Générateur d'effets vous permettent de :



Supprimer des termes du modèle à effets fixes, en sélectionnant ceux que vous souhaitez supprimer, puis en cliquant sur le bouton de suppression.



Réorganiser les termes dans le modèle à effets fixes, en sélectionnant ceux que vous souhaitez réorganiser, puis en cliquant sur les flèches vers le haut ou vers le bas, et



Ajouter des termes imbriqués au modèle à l'aide de la boîte de dialogue [Ajouter un terme personnalisé](#), en cliquant sur le bouton Ajouter un terme personnalisé.

Inclure la constante. La constante n'est pas incluse dans le modèle à effets aléatoires par défaut. Si vous partez du principe que les données passent par l'origine, vous pouvez exclure la constante.

Définir les classes de covariance par. Les champs spécifiés ici définissent des ensembles indépendants de paramètres de covariance d'effets aléatoires, un ensemble étant défini pour chaque catégorie par la classification croisée des champs de regroupement. Un ensemble de champs de regroupement différent peut être spécifié pour chaque bloc d'effet aléatoire. Tous les sujets ont le même type de covariance ; les sujets faisant partie du même regroupement de covariance auront les mêmes valeurs de paramètres.

Combinaison de sujets. Cette option vous permet de spécifier des sujets à effets aléatoires à partir de combinaisons prédéfinies de sujets depuis l'onglet Structure des données. Par exemple, si *École*, *Classe* et *Elève* sont définis comme sujets sur l'onglet Structure des données, dans cet ordre, alors la liste déroulante Combinaison de sujets comprendra les options Aucun, École, École * Classe, et École * Classe * Elève.

Type de covariance des effets aléatoires. Spécifie la structure de covariance des résidus. Les structures disponibles sont les suivantes :

- Autorégressive de premier ordre (AR1)
- Moyenne mobile autorégressive (1,1) (ARMA11)
- Symétrie composée
- Diagonale
- Identité mise à l'échelle
- Toeplitz
- Sans structure
- Composantes de variance

Pondération et décalage

Figure 10-60
Paramètres de pondération et de décalage

Structure de données | Champs et effets | Options de construction | Options de modèle

Sélectionnez un élément :

- Cible
- Effets fixes
- Effets aléatoires
- Pondération et décalage**

Pondération d'analyse :
(aucun)

Décalage

- Ce modèle ne nécessite pas de décalage
- Utiliser la valeur de décalage
Valeur de décalage: 0.0
- Utiliser le champ de décalage
Champ de décalage: (aucun)

Pondération d'analyse. Le paramètre d'échelle est un paramètre de modèle estimé lié à la variance de la réponse. Les pondérations d'analyse sont des valeurs « connues », susceptibles de varier d'une observation à l'autre. Si le champ de pondération d'analyse est spécifié, le paramètre d'échelle, qui est lié à la variance de la réponse, est divisé par les valeurs de pondération d'analyse pour chaque observation. Les enregistrements avec des valeurs de pondération d'analyse inférieures ou égales à 0, ou manquantes, ne sont pas utilisées dans l'analyse.

Décalage. Le terme de décalage est une valeur prédite « structurelle ». Son coefficient n'est pas estimé par le modèle mais sa valeur est supposée être 1. Par conséquent, les valeurs du décalage sont simplement ajoutées à la valeur prédite linéaire de la cible. Cela s'avère particulièrement utile dans les modèles de régression de Poisson, dans lesquels chaque observation peut avoir différents niveaux d'exposition à l'événement étudié.

Par exemple, lors de la modélisation des taux d'accident pour des conducteurs, la différence est importante entre un conducteur avec trois ans de conduite qui est en tort dans un accident et un autre conducteur n'ayant été en tort qu'une fois en 25 ans ! Le nombre d'accidents peut être modélisé comme une réponse de type Poisson ou binominale négative avec un lien de journal si le logarithme naturel de l'expérience du lecteur est inclus comme un terme de décalage.

D'autres combinaisons de distribution et types de liens nécessiteraient d'autres transformations de la variable de décalage.

Options de création

Figure 10-61
Paramètres des options de création

The screenshot shows the 'Options de construction' dialog box with the following settings:

- Ordre de tri:**
 - Ordre de tri des cibles qualitatives : Croissant
 - Ordre de tri des variables indépendantes qualitatives : Croissant
- Règles d'arrêt:**
 - Nombre maximum d'itérations : 100
- Paramètres post-estimation:**
 - Niveau de confiance (%) : 95.0
 - Degrees of freedom:
 - Fixes pour tous les tests (méthode résiduelle)
Utile si la taille de l'échantillon est plus importante, si les données sont équilibrées ou si elles utilisent un type de covariance simple (par exemple, identité mise à l'échelle ou diagonale)
 - Variables selon les tests (approximation de Satterthwaite)
Utile si la taille de l'échantillon est plus petite, si les données ne sont pas équilibrées ou si elles ont des types de covariance complexes (par exemple, non structurés)
 - Test des effets fixes et des coefficients:
 - Supposez que les hypothèses de modèle sont correctes.
(covariances basées sur un modèle)
 - Utilisez une estimation robuste pour traiter les violations des hypothèses de modèle
(covariances robustes)

Ces sélections spécifient des critères plus avancés utilisés pour créer le modèle.

Ordre de tri. Ces commandes déterminent l'ordre des catégories de la cible et des facteurs (entrées catégorielles) afin de déterminer la « dernière » catégorie. Le paramètre ordre de tri de la cible est ignoré si la cible n'est pas catégorielle ou si une catégorie de référence personnalisée est spécifiée sur les paramètres [Cible](#) .

Règles d'arrêt. Vous pouvez spécifier le nombre maximal d'itérations exécutées par l'algorithme. Spécifiez un nombre entier non négatif. La valeur par défaut est 100.

Paramètres post-estimation. Ces paramètres déterminent la façon dont une partie du résultat du modèle est calculée pour l'affichage.

- **Niveau de confiance.** Il s'agit du niveau de confiance utilisé pour calculer les estimations d'intervalle des coefficients du modèle. Définissez une valeur supérieure à 0 et inférieure à 100. La valeur par défaut est 95.

- **Degrés de liberté.** Spécifie la façon dont les degrés de liberté sont calculés pour les tests de signification. Choisissez Fixes pour tous les tests (méthode des résidus) si la taille de votre échantillon est suffisamment grande, si les données sont équilibrées ou si le modèle utilise un type de covariance plus simple ; par exemple, identité mise à l'échelle ou diagonale. Il s'agit de la valeur par défaut. Choisissez Différents selon les tests (approximation Satterthwaite) si la taille de votre échantillon est petite, si les données sont déséquilibrées ou si le modèle utilise un type de covariance compliqué ; par exemple, non structuré.
- **Tests des effets fixes et coefficients.** C'est la méthode de calcul de la matrice de covariance des estimations de paramètres. Choisissez l'estimation robuste si vous craignez de ne pas respecter les hypothèses du modèle.

Général

Figure 10-62
Paramètres généraux

Sélectionnez un élément :

General
Moyennes estimées

Nom de modèle: Automatique Personnalisé

Rendre disponible pour le scoring

La valeur prédite est toujours disponible pour le scoring. La confiance est disponible pour le scoring uniquement s'il s'agit de cibles catégorielles.

La confiance est basée sur:

La probabilité de la valeur prédite
 L'augmentation dans la probabilité de la valeur la plus probable suivante

Prévision de probabilité pour les cibles qualitatives

Modalités maximales à enregistrer:

Scores de propension pour les cibles de type booléen

Nom de modèle. Vous pouvez générer automatiquement le nom du modèle en fonction des champs cible ou spécifier un nom personnalisé. Le nom généré automatiquement est le nom du champ cible. S'il existe plusieurs cibles, le nom du modèle correspond aux noms des champs reliés par des esperluettes. Par exemple, si *champ1 champ2 champ3* sont des cibles, alors le nom du modèle est : *champ1 & champ2 & champ3*.

Rendre disponible pour le scoring. Lorsque le modèle est évalué, les éléments sélectionnés dans ce groupe doivent être générés. La valeur prédite (pour toutes les cibles) et la confiance (pour les cibles catégorielles) sont toujours calculées lorsque le modèle est évalué. La confiance calculée peut être basée sur la probabilité de la valeur prédite (la probabilité prédite la plus élevée) ou sur la différence entre la probabilité prédite la plus élevée et la deuxième probabilité prédite la plus élevée.

- **Probabilité prédite pour les cibles catégorielles.** Ceci génère les probabilités prédites pour les cibles catégorielles. Un champ est créé pour chaque modalité.
- **Scores de propension pour des variables cibles booléennes.** Pour les modèles disposant d'une variable cible booléenne (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Le modèle produit des scores de propension brute. Si les partitions sont activées, le modèle produit également des scores de propension ajustée basés sur la partition de test. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Moyennes estimées

Figure 10-63
Paramètres des moyennes estimées

Sélectionnez un élément :

Général
Moyennes estimées

Souhaitez-vous estimer la cible ?
Spécifiez des contrastes et des moyennes estimées personnalisés.

Termes:

Termes qualitatifs :	Moyennes estimées	Type de contraste	Champ de contraste
Center size	<input checked="" type="checkbox"/>	Aucun	Center size
Treatment received	<input type="checkbox"/>	Aucun	Treatment received

La constance des champs continus sera maintenue lors de l'estimation de la cible.

Champs:

Champs continus	Constante	Valeur
Date of birth	Moyenne	

Affichez les moyennes estimées en termes de:
 Échelle cible d'origine
 Transformation de la fonction de lien

Apportez des ajustements pour des comparaisons multiples à l'aide de:
 Différence la moins significative

Cet onglet vous permet d'afficher les moyennes marginales estimées des niveaux et interactions de facteurs. Les moyennes marginales estimées ne sont pas disponibles pour les modèles multinomiaux.

Termes. Les termes du modèle entièrement composés de champs catégoriels dans les effets fixes sont répertoriés ici. Cochez chaque terme pour lequel vous souhaitez que le modèle produise des moyennes marginales estimées.

- **Type de contraste.** Cela spécifie le type de contraste à utiliser pour les niveaux du champ contraste. Si Aucun est sélectionné, aucun contraste n'est produit. L'option Par paire produit des comparaisons par paire pour toutes les combinaisons de niveaux des facteurs spécifiés. C'est le seul contraste disponible pour les interactions de facteurs. Les contrastes Écart comparent chaque niveau du facteur à la moyenne générale. Les contrastes simples comparent chaque niveau du facteur, sauf le dernier, au dernier niveau. Le « dernier » niveau est

déterminé par l'ordre de tri des facteurs spécifiés dans les options de création. Notez que tous ces types de contraste ne sont pas orthogonaux.

- **Champ Contraste.** Spécifie un facteur, dont les niveaux sont comparés à l'aide du type de contraste sélectionné. Si le type de contraste sélectionné est Aucun, aucun champ de contraste ne peut (ou ne doit) être sélectionné.

Champs continus. Les champs continus répertoriés sont extraits des termes situés dans les effets fixes utilisant des champs continus. Lors du calcul des moyennes marginales estimées, les covariables sont fixées aux valeurs spécifiées. Sélectionnez la moyenne ou spécifiez une valeur personnalisée.

Afficher les moyennes estimées en termes de. Spécifie s'il faut calculer les moyennes marginales estimées d'après l'échelle originale de la cible ou d'après la transformation de la fonction de lien. L'Échelle de cible originale calcule les moyennes marginales estimées de la cible. Notez que lorsque la cible est spécifiée à l'aide de l'option événements/essais, il en résulte les moyennes marginales estimées de la proportion événements/essais plutôt que celles du nombre d'événements. La Transformation de la fonction de lien calcule les moyennes marginales estimées de la variable prédite linéaire.

Ajuster pour les comparaisons multiples à l'aide de. Lors de l'exécution de tests d'hypothèse avec plusieurs contrastes, vous pouvez ajuster le niveau de signification globale à partir des niveaux de signification des contrastes inclus. Cela vous permet de choisir la méthode d'ajustement.

- **Différence la moins significative.** Cette méthode ne contrôle pas l'intégralité de la probabilité de rejet des hypothèses qui supposent que certains contrastes linéaires sont différents des valeurs d'hypothèse nulles.
- **Bonferroni séquentiel.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.
- **Procédure de Sidak Séquentielle.** Il s'agit d'une procédure descendante de rejet séquentiel de Bonferroni beaucoup moins stricte en ce qui concerne le rejet des différentes hypothèses mais qui conserve le même niveau global de signification.

La méthode de différence la moins significative est moins stricte que la méthode de Sidak séquentielle, qui elle-même est moins stricte que la méthode séquentielle de Bonferroni. En d'autres termes, la différence la moins significative va rejeter au moins autant d'hypothèses individuelles que la méthode séquentielle de Sidak, qui elle-même va rejeter au moins autant d'hypothèses individuelles que la méthode séquentielle de Bonferroni.

Vue du modèle

Par défaut, la vue Récapitulatif du modèle apparaît. Pour voir une autre vue de modèle, sélectionnez-la parmi les miniatures des vues.

Récapitulatif du modèle

Figure 10-64
Vue récapitulative du modèle

Récapitulatif des modèles

Cible : Number of convulsions

Cible	Number of convulsions	
Distribution de probabilité	Poisson	
Fonction de lien	Log	
Critères d'informations	Akaike corrigé	6,983.683
	Bayésien	7,008.180

Les critères d'informations sont basés sur le pseudo log-2 vraisemblance (6,975.671) et sont utilisés pour comparer les modèles. Les modèles dont les valeurs de critères d'informations sont inférieures, sont plus appropriés. Soyez prudents lors de la comparaison des modèles à l'aide des valeurs de pseudo vraisemblance, car différentes transformations de données peuvent être utilisées selon les modèles.

Cette vue du modèle est un instantané, permettant de consulter en un coup d'œil le modèle et son ajustement.

Tableau. Le tableau identifie la cible, la distribution des probabilités et la fonction de lien spécifiées dans les [Paramètres de la cible](#). Si la cible est définie par des événements et des essais, la cellule est divisée de façon à montrer le champ des événements et le champ des essais ou un nombre fixe d'essais. En outre, le critère d'information Akaike corrigé (AICC) et le critère d'information bayésien (BIC) de l'échantillon fini sont affichés.

- **Akaike corrigé.** Mesure de sélection et de comparaison des modèles mixtes basée sur la valeur $-2 \log$ de vraisemblance (restreinte). Les petites valeurs indiquent de meilleurs modèles. L'AICC « corrige » l'AIC pour obtenir des tailles d'échantillon plus petites. Plus la taille de l'échantillon augmente, plus l'AICC converge vers l'AIC.
- **Bayésien.** Mesure de sélection et de comparaison des modèles basée sur la valeur $-2 \log$ de vraisemblance. Les petites valeurs indiquent de meilleurs modèles. Le critère BIC pénalise les modèles sur-paramétrés, mais de manière plus stricte que le critère AIC.

Diagramme. Si la cible est catégorielle, un diagramme affiche la précision du modèle final, soit le pourcentage de classifications correctes.

Structure des données

Figure 10-65
 Vue Structure des données

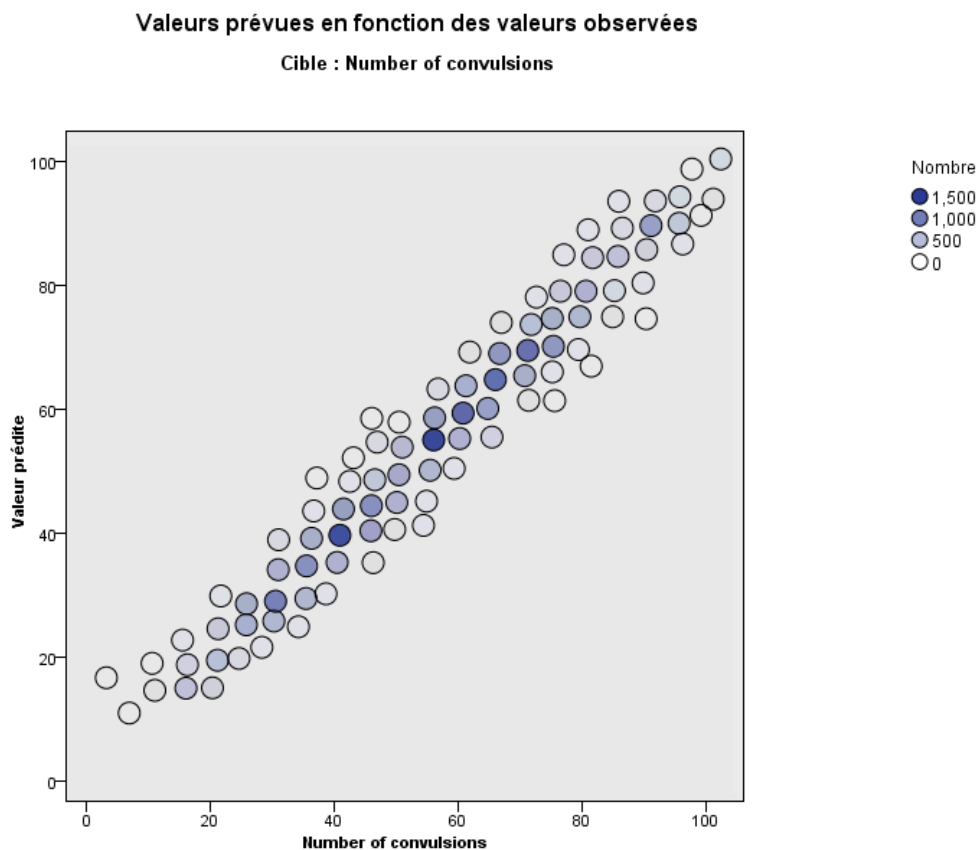
Structure de données
 Cible : Number of convulsions

	Sujets			Mesures répétées	Cible
	Center ID	Attending Physician ID	Patient ID	Week	Number of convulsions
Données du premier sujet	07057	COMX	1FSL	0	2
	07057	COMX	1FSL	1	6
	07057	COMX	1FSL	2	4
	07057	COMX	1FSL	3	4
	07057	COMX	1FSL	4	6
	07057	COMX	1FSL	5	3
Nombre de niveaux total	8	49	565	6	

Cette vue offre un récapitulatif de la structure des données que vous avez spécifiée et vous aide à vérifier que les sujets et les mesures répétées ont été correctement spécifiés. Les informations observées pour le premier sujet sont affichées pour chaque champ de sujet, chaque champ de mesures répétées et la cible. En outre, le nombre de niveaux de chaque champ de sujet et champ de mesures répétées est affiché.

Valeurs prédites en fonction des valeurs observées

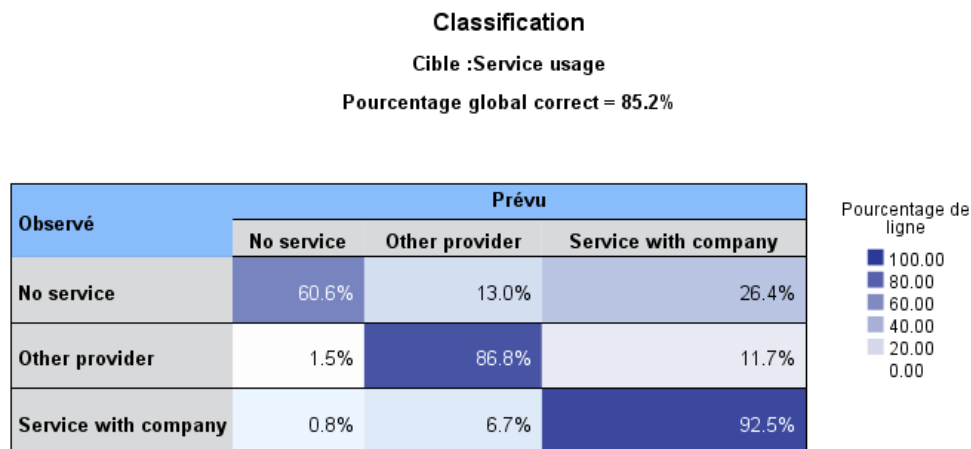
Figure 10-66

Vue des valeurs prédites en fonction des valeurs observées

Pour les cibles continues, y compris les cibles spécifiées comme événements/essais, cette option affiche un diagramme de dispersion mis en intervalles des valeurs prédites sur l'axe vertical en fonction des valeurs observées sur l'axe horizontal. Idéalement, les points devraient se trouver sur une ligne de 45 degrés ; cette vue peut vous indiquer si des enregistrements sont particulièrement mal prédits par le modèle.

Classification

Figure 10-67
Vue Classification



Pour les cibles catégorielles, cette option affiche la classification croisée des valeurs observées par rapport aux valeurs prédites sur une carte des zones de chaleur, ainsi que le pourcentage global correct.

Styles de tableaux. Il existe différents styles d’affichage, accessibles depuis la liste déroulante Style .

- **Pourcentages de ligne.** Affiche les pourcentages de ligne (les effectifs de cellules exprimés sous forme de pourcentage du nombre total de lignes) dans les cellules. Il s’agit de la valeur par défaut.
- **Effectifs de cellules.** Affiche les effectifs de cellules dans les cellules. L’ombrage de la carte des zones de chaleur reste basé sur les pourcentages de ligne.
- **Zones de chaleur.** N’affiche aucune valeur dans les cellules, uniquement l’ombrage.
- **Compressé.** N’affiche aucun en-tête de ligne ou de colonne ni de valeur dans les cellules. Ceci peut être utile lorsque la cible possède de nombreuses catégories.

Manquantes. Si des enregistrements ont des valeurs manquantes sur la cible, elles sont affichées sur une ligne (Manquantes) en dessous des lignes valides. Les enregistrements avec des valeurs manquantes ne contribuent pas au pourcentage général correct.

Cibles multiples. S’il existe plusieurs cibles catégorielles, chacune est affichée dans un tableau distinct et une liste déroulante Cible contrôle la cible à afficher.

Grands tableaux. Si la cible affichée comporte plus de 100 catégories, aucun tableau n’est affiché.

Effets fixes

Figure 10-68
Vue Effets fixes, style de diagramme

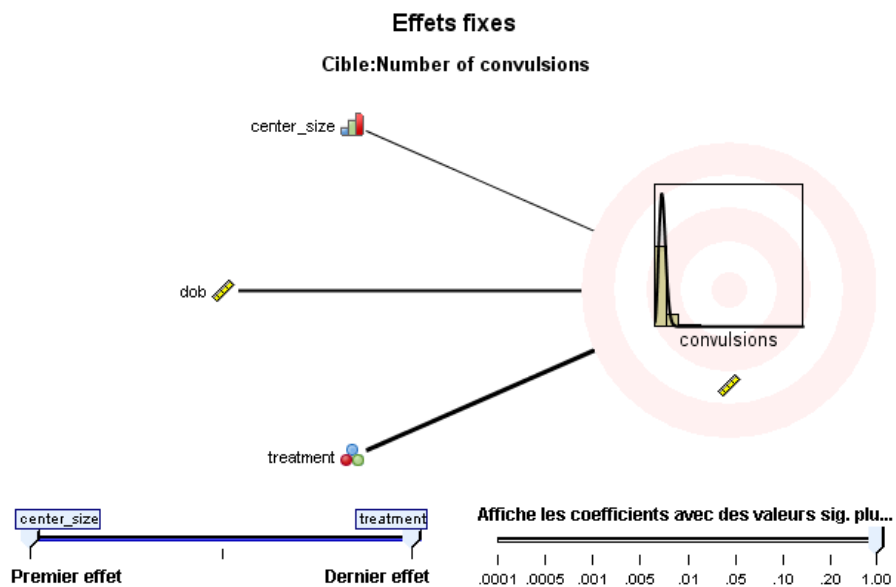


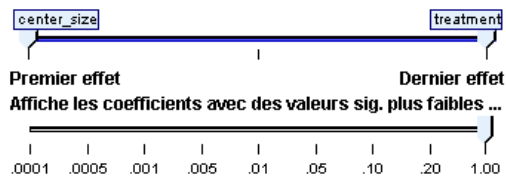
Figure 10-69
 Vue Effets fixes, style de tableau

Coefficients fixes
 Cible: Number of convulsions

Caractéristiques du modèle	Coefficient ►	Sig.
Constante	1.745	.000
center_size=1	0.078	.610
center_size=2	-0.157	.201
center_size=3	0 ^a	
dob	-0.000	.185
treatment=0	-0.138	.053
treatment=1	0 ^a	

Distribution de probabilité :Poisson
 Fonction de lien :Log

^aCe coefficient est défini sur zéro car il est redondant.



Cette vue affiche la taille de chaque effet fixe dans le modèle.

Styles. Il existe différents styles d’affichage accessibles depuis la liste déroulante Style.

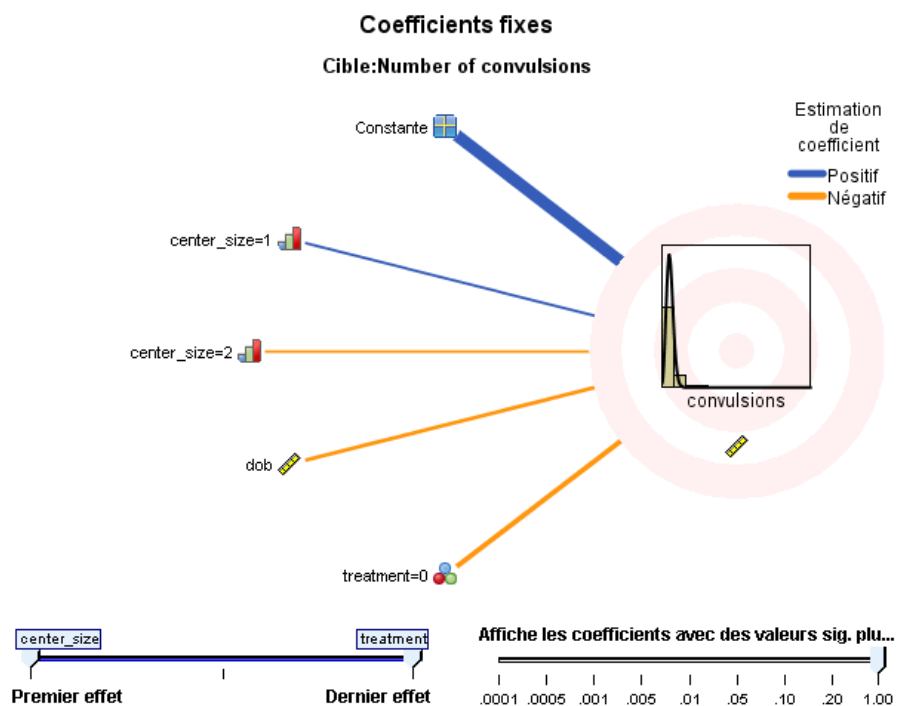
- **Diagramme.** Il s’agit d’un diagramme dans lequel les effets sont triés de haut en bas dans l’ordre où ils ont été spécifiés sur les paramètres des effets fixes. Les lignes de connexion du diagramme sont pondérées en fonction de la signification de l’effet, une largeur de ligne plus importante correspondant à des effets plus importants (valeurs p plus petites). Il s’agit de la valeur par défaut.
- **Tableau.** Il s’agit d’un tableau ANOVA pour le modèle général et les effets de modèle individuels. Les effets individuels sont triés de haut en bas dans l’ordre où ils ont été spécifiés sur les paramètres des effets fixes.

Signification. Il existe un curseur de signification qui contrôle les effets affichés dans la vue. Les effets ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les effets les plus

importants. Par défaut, la valeur est de 1,00, de sorte qu'aucun effet n'est filtré en fonction de la signification.

Coefficients fixes

Figure 10-70
Vue Coefficients fixes, style de diagramme



sein des effets contenant des facteurs, les coefficients sont triés dans l'ordre croissant de la valeur des données.

Multinomiale. Si la distribution multinomiale est activée, la liste déroulante Multinomiale contrôle la catégorie de cible à afficher. L'ordre de tri des valeurs dans la liste est déterminé par la spécification des paramètres des options de création.

Exponentiel. Cela affiche les estimations de coefficient exponentiel et les intervalles de confiance pour certains types de modèles, y compris la régression logistique binaire (distribution binomiale et lien logit), la régression logistique nominale (distribution multinomiale et lien logit), la régression binomiale négative (distribution binomiale négative et lien log) et le modèle Log-linéaire (distribution de Poisson et lien log).

Signification. Il existe un curseur de signification qui contrôle les coefficients affichés dans la vue. Les coefficients ayant des valeurs de signification plus grandes que la valeur du curseur sont masqués. Ceci ne modifie pas le modèle, mais vous permet simplement de vous concentrer sur les coefficients les plus importants. Par défaut, la valeur est de 1,00, de sorte qu'aucun coefficient n'est filtré en fonction de la signification.

Covariances des effets aléatoires

Cette vue affiche la matrice de covariance des effets aléatoires (**G**).

Styles. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style.

- **Valeurs de covariance.** Il s'agit d'une carte des zones de chaleur de la matrice de covariance dans laquelle les effets sont triés de haut en bas dans l'ordre où ils ont été spécifiés sur les paramètres des effets fixes. Les couleurs du corrélogramme correspondent aux valeurs des cellules telles qu'affichées dans la clé. Il s'agit de la valeur par défaut.
- **Corrélogramme.** Il s'agit d'une carte des zones de chaleur de la matrice de covariance.
- **Compressé.** Il s'agit d'une carte des zones de chaleur de la matrice de covariance sans les en-têtes de ligne et de colonne.

Blocs. S'il existe plusieurs blocs d'effets aléatoires, une liste déroulante Bloc permet de sélectionner le bloc à afficher.

Groupes. Si un bloc d'effets aléatoires possède une spécification de groupe, une liste déroulante Groupe permet de sélectionner le niveau de groupe à afficher.

Multinomiale. Si la distribution multinomiale est activée, la liste déroulante Multinomiale contrôle la catégorie de cible à afficher. L'ordre de tri des valeurs dans la liste est déterminé par la spécification des paramètres des options de création.

Paramètres de covariance

Figure 10-72
Vue des paramètres de covariance

Paramètres de covariance

Cible: Number of convulsions

Paramètres de covariance	Effet résiduel	2
	Effets aléatoires	2
Colonnes de matrice de plan	Effets fixes	7
	Effets aléatoires	10 ^a
Sujets communs		8

Les sujets communs sont basés sur les spécifications de sujet des effets aléatoires et résiduels, et sont utilisés pour segmenter les données en vue d'obtenir de meilleures performances.

^a Ceci est le nombre de colonnes par sujet commun.

Effet résiduel	Estimation	Erreur std.	Z	Sig.	Intervalle de confiance 95%	
					Inférieur	Supérieur
AR1 Diagonale	5.379	0.243	22.151	.000	4.923	5.876
AR1 Rho	0.805	0.010	82.823	.000	0.785	0.824

Structure de covariance : Autorégressif de premier ordre
Spécification de sujet : center_id*atphys_id*patient_id

Cette vue affiche les estimations des paramètres de covariance et les statistiques associées aux effets résiduels et aléatoires. Il s'agit de résultats élaborés, mais fondamentaux, qui indiquent si la structure de covariance est appropriée ou pas.

Récapitulatif. Il s'agit d'un aide-mémoire pour le nombre de paramètres dans les matrices de covariance des effets résiduels (**R**) et aléatoires (**G**), le rang (nombre de colonnes) dans les matrices de conception des effets fixes (**X**) et aléatoires (**Z**) et le nombre de sujets définis par les champs de sujet définissant la structure des données.

Tableau des paramètres de covariance. Pour l'effet sélectionné, l'estimation, l'erreur standard et l'intervalle de confiance sont affichés pour chaque paramètre de covariance. Le nombre de paramètres affiché dépend de la structure de covariance de l'effet et, pour les blocs d'effets aléatoires, du nombre d'effets dans le bloc. Si vous voyez que les paramètres hors diagonale ne sont pas significatifs, vous pouvez peut-être utiliser une structure de covariance plus simple.

Effets. S'il y a plusieurs blocs d'effets aléatoires, une liste déroulante Effet permet de sélectionner le bloc d'effets résiduels ou aléatoires à afficher. L'effet résiduel est toujours disponible.

Groupes. Si un bloc d'effets résiduels ou aléatoires possède une spécification de groupe, une liste déroulante Groupe permet de sélectionner le niveau de groupe à afficher.

Multinomiale. Si la distribution multinomiale est activée, la liste déroulante Multinomiale contrôle la catégorie de cible à afficher. L'ordre de tri des valeurs dans la liste est déterminé par la spécification des paramètres des options de création.

Moyennes estimées : Effets significatifs

Il s'agit des diagrammes affichés pour les 10 effets « les plus significatifs » de tous les facteurs fixes, avec d'abord les interactions tridirectionnelles, puis les interactions bidirectionnelles et enfin les effets principaux. Le diagramme affiche la valeur de la cible estimée par le modèle sur l'axe vertical pour chaque valeur de l'effet principal (ou l'effet listé en premier dans une interaction) sur l'axe horizontal ; une ligne séparée est produite pour chaque valeur du deuxième effet listé dans une interaction ; un diagramme séparé est produit pour chaque valeur du troisième effet listé dans une interaction tridirectionnelle ; toutes les autres variables prédites restent constantes. Il offre une visualisation pratique des effets des coefficients de chaque variable prédite sur la cible. Notez que si aucune variable prédite n'est significative, aucune moyenne estimée n'est générée.

Confiance. Affiche les limites de confiance supérieure et inférieure pour les moyennes marginales, en utilisant le niveau de confiance spécifié dans le cadre des Options de création.

Moyennes estimées : Effets personnalisés

Il s'agit de tableaux et de diagrammes relatifs aux effets de tous les facteurs fixes demandés par l'utilisateur.

Styles. Il existe différents styles d'affichage accessibles depuis la liste déroulante Style.

- **Diagramme.** Ce style affiche un diagramme curviligne de la valeur de la cible estimée par le modèle sur l'axe vertical pour chaque valeur de l'effet principal (ou l'effet listé en premier dans une interaction) sur l'axe horizontal ; une ligne séparée est produite pour chaque valeur du deuxième effet listé dans une interaction ; un diagramme séparé est produit pour chaque valeur du troisième effet listé dans une interaction tridirectionnelle ; toutes les autres variables prédites restent constantes.

Si des contrastes sont demandés, un autre diagramme est affiché pour comparer les niveaux du champ contraste ; pour les interactions, un diagramme est affiché pour chaque combinaison de niveau des effets autres que le champ contraste. Pour les contrastes **par paire**, il s'agit d'un diagramme de réseau des distances, soit une représentation graphique du tableau des comparaisons dans lequel les distances entre les nœuds du réseau correspondent aux différences entre les échantillons. Les lignes jaunes correspondent aux différences statistiquement significatives, alors que les lignes noires correspondent aux différences non significatives. Lorsque vous passez la souris sur une ligne du réseau, la signification ajustée de la différence entre les nœuds connectés par la ligne s'affiche dans une info-bulle.

Pour les contrastes d'**écart**, un diagramme en bâtons est affiché avec la valeur de la cible estimée par le modèle sur l'axe vertical et les valeurs du champ contraste sur l'axe horizontal ; pour les interactions, un diagramme est affiché pour chaque combinaison de niveau des effets autres que le champ contraste. Les bâtons montrent la différence entre chaque niveau du champ contraste et la moyenne générale, qui est représentée par une ligne horizontale noire.

Pour les contrastes **simples**, un diagramme en bâtons est affiché avec la valeur de la cible estimée par le modèle sur l'axe vertical et les valeurs du champ contraste sur l'axe horizontal ; pour les interactions, un diagramme est affiché pour chaque combinaison de niveau des effets autres que le champ contraste. Les bâtons montrent la différence entre chaque niveau du champ contraste (sauf le dernier) et le dernier niveau, qui est représenté par une ligne horizontale noire.

- **Tableau.** Ce style affiche un tableau de la valeur de la cible estimée par le modèle, son erreur standard et l'intervalle de confiance de chaque combinaison de niveau des champs de l'effet ; toutes les autres variables prédites restent constantes.

Si des contrastes sont demandés, un autre tableau est affiché avec l'estimation, l'erreur standard, le test de signification et l'intervalle de confiance de chaque contraste ; pour les interactions, il y a un ensemble séparé de lignes pour chaque combinaison de niveau des effets autres que le champ contraste. En outre, un tableau avec les résultats des tests globaux est affiché ; pour les interactions, il y a un test global séparé pour chaque combinaison de niveau des effets autres que le champ contraste.

Confiance. Cela affiche ou masque les limites de confiance supérieure et inférieure des moyennes marginales, en utilisant le niveau de confiance spécifié dans le cadre des Options de création.

Présentation Affiche ou masque la présentation du diagramme des contrastes par paire. La présentation en cercle est moins révélatrice des contrastes que la présentation en réseau mais évite que les lignes ne se chevauchent.

Paramètres

Figure 10-73
Paramètres de modèle

Modèle Paramètres Annotations

La valeur prédite et la confiance sont toujours disponibles pour le scoring.

La confiance est basée sur:

La probabilité de la valeur prédite

L'augmentation dans la probabilité de la valeur la plus probable suivante

Probabilité prédite pour les cibles catégorielles

Nombre maximal de catégories à enregistrer: 25

Scores de propension pour les cibles de type booléen

Lorsque le modèle est évalué, les éléments sélectionnés dans cet onglet doivent être générés. La valeur prédite (pour toutes les cibles) et la confiance (pour les cibles catégorielles) sont toujours calculées lorsque le modèle est évalué. La confiance calculée peut être basée sur la probabilité de la valeur prédite (la probabilité prédite la plus élevée) ou sur la différence entre la probabilité prédite la plus élevée et la deuxième probabilité prédite la plus élevée.

- **Probabilité prédite pour les cibles catégorielles.** Ceci génère les probabilités prédites pour les cibles catégorielles. Un champ est créé pour chaque modalité.
- **Scores de propension pour des variables cibles booléennes.** Pour les modèles disposant d'une variable cible booléenne (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Le modèle produit des scores de propension brute. Si les partitions sont activées, le modèle produit également des scores de propension ajustée basés sur la partition de test. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Noeud de Cox

La régression Cox crée un modèle prédictif pour les données concernant le temps écoulé jusqu'à un événement. Le modèle permet d'obtenir une fonction de survie qui prédit la probabilité que l'événement en question se soit produit à un moment donné t pour des valeurs données des variables de prévision. La forme de la fonction de survie et les coefficients de régression des valeurs prédites sont estimés à partir de sujets observés ; le modèle peut alors s'appliquer à de nouvelles observations ayant des mesures pour les variables de prévision. Notez que les informations provenant de sujets censurés, autrement dit celles qui ne font pas état de l'événement en question pendant la durée de l'observation, contribuent utilement à l'estimation du modèle.

Exemple. Dans ses efforts pour réduire l'attrition de la clientèle, une entreprise de télécommunications s'intéresse à la modélisation de la "durée d'attrition" afin de déterminer les facteurs associés aux clients qui changent rapidement de service. A ces fins, un échantillon de clients est sélectionné au hasard, et la durée passé comme client (qu'ils soient des clients actifs ou non) et différents champs démographiques sont extraits de la base de données. [Pour plus d'informations, reportez-vous à la section Utilisation de la régression de Cox pour modéliser la durée jusqu'à l'attrition de la clientèle dans le chapitre 26 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises. Vous devez avoir un ou plusieurs champs d'entrée, exactement un champ cible et vous devez préciser un champ de durée de survie dans le noeud de Cox. Le champ cible doit être codé afin que la valeur « fausse » indique la survie et que la valeur « vraie » indique que l'évènement donné s'est produit. Celui-ci doit comporter un niveau de mesure *Booléen* avec stockage de chaîne ou d'entier. (Le stockage peut être converti à l'aide d'un noeud Remplacer ou Calculer si nécessaire. [Pour plus d'informations, reportez-vous à la section Conversion du stockage à l'aide du noeud Remplacer dans le chapitre 4 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#)). Les champs paramétrés sur *Les deux* ou sur *Aucun* sont ignorés. Les types des champs utilisés dans ce modèle doivent être complètement instanciés. La durée de survie peut être tout champ numérique.

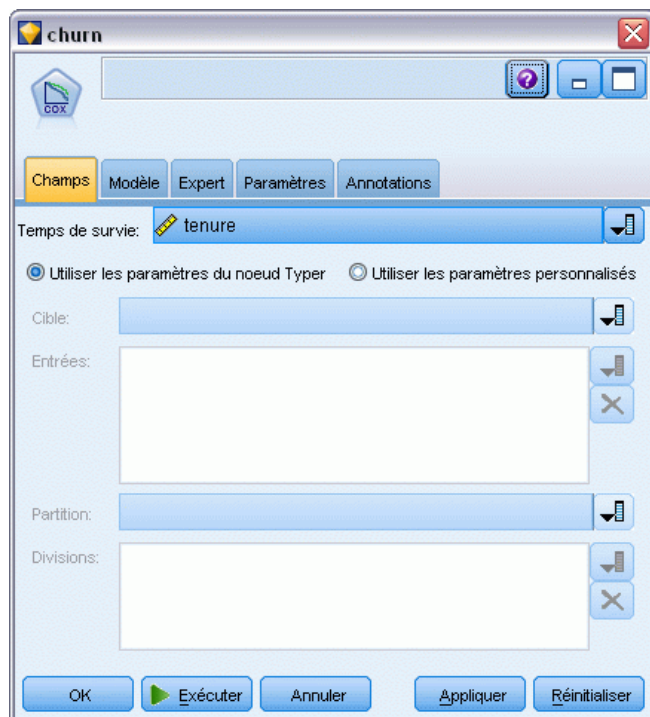
Dates & Heures. Les champs Date & Heure ne peuvent pas être utilisés pour définir directement la durée de survie. Si vous avez des champs Date & Heure, vous devez les utiliser pour créer un champ contenant des durées de survie basées sur la différence entre la date d'entrée dans l'étude et

la date d'observation. [Pour plus d'informations, reportez-vous à la section Utilisation de l'heure et de la date dans le chapitre 7 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Analyse Kaplan-Meier. La régression de Cox peut être effectuée sans champ d'entrée. Elle est semblable à l'analyse Kaplan-Meier.

Noeud de Cox - Options de champs

Figure 10-74
Boîte de dialogue du noeud de Cox - Onglet Champs



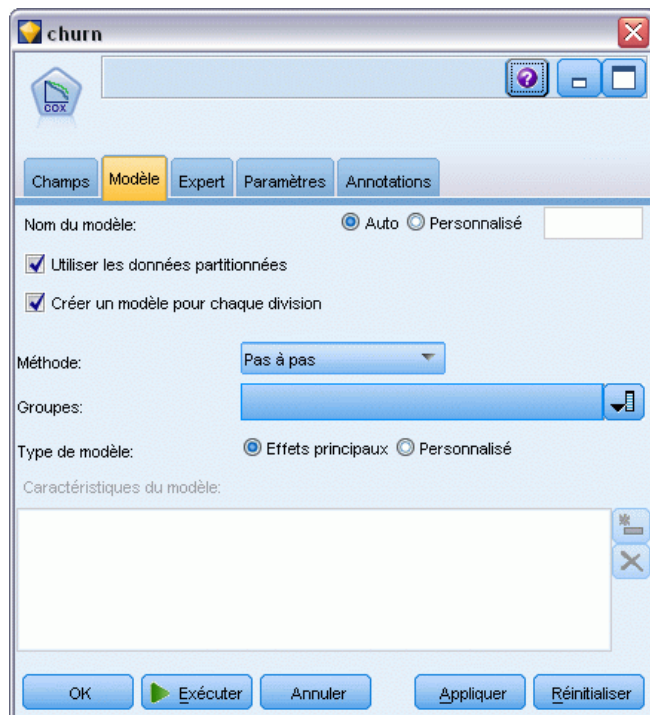
Durée de survie. Choisissez un champ numérique (qui comporte un niveau de mesure *Continu*) afin de rendre le noeud exécutable. La durée de survie indique la durée de vie de l'enregistrement prédit. Par exemple, lors de la modélisation de la durée jusqu'à l'attrition de la clientèle, ceci serait le champ qui spécifie la durée pendant laquelle une personne a été cliente de l'entreprise. La date à laquelle le client a rejoint ou quitté l'entreprise n'affecte pas le modèle, seule la durée pendant laquelle le client a bénéficié du service s'applique.

La durée de survie est une durée sans unités. Vous devez vous assurer que les champs d'entrée correspondent à la durée de survie. Par exemple, au cours d'une étude qui mesure l'attrition par mois, vous utiliseriez les ventes par mois comme entrée au lieu des ventes par année. Si vos données contiennent des dates de début et de fin au lieu d'une durée, vous devez recoder ces dates en une durée en amont du noeud de Cox.

Les autres champs de cette boîte de dialogue sont les champs habituels de IBM® SPSS® Modeler. [Pour plus d'informations, reportez-vous à la section Options de champs des noeuds de modélisation dans le chapitre 3 sur p. 39.](#)

Noeud de Cox - Options de modèle

Figure 10-75
Boîte de dialogue du noeud de Cox, 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Méthode : Les options suivantes sont disponibles pour la saisie de valeurs prédites dans le modèle :

- **Entrée.** Cette méthode par défaut intègre directement toutes les caractéristiques dans le modèle. Aucune sélection de champ n'intervient dans la création du modèle.
- **Pas à pas.** Comme son nom l'indique, la méthode de sélection de champs pas à pas génère le modèle par étapes. Le modèle initial est le modèle le plus simple : son modèle ne comporte aucune caractéristique de modèle (à l'exception de la constante). A chaque étape, les caractéristiques qui n'ont pas encore été intégrées au modèle sont évaluées et celles qui améliorent de manière significative la puissance de prévision du modèle sont alors ajoutées au modèle. De plus, les caractéristiques déjà intégrées au modèle sont réévaluées afin de déterminer si certaines d'entre elles peuvent être supprimées sans que cela affecte le

fonctionnement du modèle. Si c'est le cas, elles sont supprimées. Le processus se répète et d'autres caractéristiques sont donc ajoutées et/ou supprimées. Lorsqu'aucune caractéristique ne peut être ajoutée au modèle pour l'améliorer, ou qu'aucune caractéristique ne peut être supprimée du modèle sans risquer de le dégrader, le modèle final est généré.

- **Pas à pas descendante.** Cette méthode est le contraire de la méthode Pas à pas. Avec cette méthode, le modèle initial utilise toutes les caractéristiques en tant que valeurs prédites. A chaque étape, les caractéristiques du modèle sont évaluées et celles dont la suppression n'a aucune incidence sur le fonctionnement du modèle sont supprimées. De plus, les caractéristiques précédemment supprimées sont réévaluées afin de déterminer si la meilleure d'entre elles améliore de manière significative la puissance de prévision du modèle. Si c'est le cas, elle est rajoutée au modèle. Lorsqu'aucune caractéristique ne peut être supprimée du modèle sans risquer de le dégrader, et qu'aucune caractéristique ne peut être ajoutée au modèle pour l'améliorer, le modèle final est généré.

Remarque : les méthodes de sélection de champs automatiques (Pas à pas et Pas à pas descendante) sont des méthodes d'apprentissage extrêmement flexibles qui ont une forte tendance à surajuster les données d'apprentissage. Lorsque vous utilisez ces méthodes, il est important de vérifier la validité du modèle généré, soit à l'aide de nouvelles données, soit à l'aide d'un échantillon de test retenu créé à l'aide du noeud Partitionner. [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*.](#)

Groupes. Lorsque vous spécifiez un champ de groupes, le noeud calcule des modèles séparés pour chaque catégorie du champ. Il peut être tout champ catégoriel (Booléen ou Nominal) avec stockage de chaîne et d'entier.

Type de modèle. Deux options permettent de définir les caractéristiques de votre modèle. Les modèles **Effets principaux** incluent les champs d'entrée individuellement et ne testent pas les interactions (effets multiplicateurs) entre les champs d'entrée. Les modèles **personnalisés** comprennent uniquement les caractéristiques (effets principaux et interactions) que vous indiquez. Lorsque cette option est sélectionnée, utilisez la liste Caractéristiques du modèle pour ajouter des caractéristiques au modèle ou pour en supprimer.

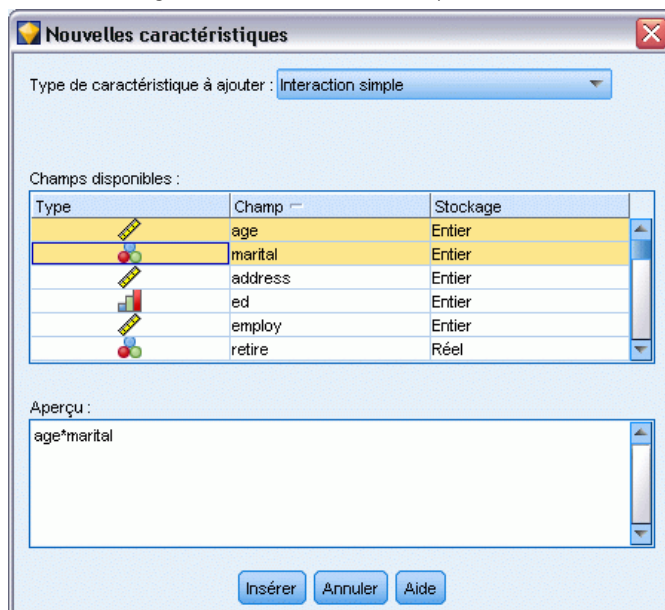
Caractéristiques du modèle. Lorsque vous créez un modèle personnalisé, vous devez indiquer explicitement ses caractéristiques. La liste répertorie toutes les caractéristiques actuelles du modèle. Les boutons situés à droite de la liste Caractéristiques du modèle permettent d'ajouter et de supprimer des caractéristiques de modèle.

- ▶ Pour ajouter des caractéristiques au modèle, cliquez sur le bouton *Ajouter de nouvelles caractéristiques au modèle*.
- ▶ Pour supprimer des caractéristiques, sélectionnez-les, puis cliquez sur le bouton *Supprimer les caractéristiques du modèle sélectionné*.

Ajout de caractéristiques à un modèle de régression de Cox

Lorsque vous demandez un modèle personnalisé, vous pouvez ajouter des caractéristiques à ce modèle en cliquant sur le bouton *Ajouter de nouvelles caractéristiques au modèle* de l'onglet Modèle. Dans la boîte de dialogue qui apparaît, vous pouvez indiquer les caractéristiques souhaitées.

Figure 10-76
Boîte de dialogue Nouvelles caractéristiques



Type de caractéristique à ajouter. Il existe plusieurs méthodes pour ajouter des caractéristiques au modèle. Ces méthodes dépendent de la sélection des champs d'entrée dans la liste Champs disponibles.

- **Interaction simple.** Insère la caractéristique représentant l'interaction de tous les champs sélectionnés.
- **Effets principaux :** Insère une caractéristique effet principal (le champ lui-même) pour chaque champ d'entrée sélectionné.
- **Toutes les interactions de second ordre.** Insère une interaction de second ordre (le produit des champs d'entrée) pour chaque paire possible de champs d'entrée sélectionnés. Par exemple, si vous avez sélectionné les champs d'entrée A , B et C dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B$, $A * C$ et $B * C$.
- **Toutes les interactions de troisième ordre.** Insère une interaction de troisième ordre (le produit des champs d'entrée) pour chaque combinaison possible de champs d'entrée sélectionnés (trois à la fois). Par exemple, si vous avez sélectionné les champs d'entrée A , B , C et D dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B * C$, $A * B * D$, $A * C * D$ et $B * C * D$.
- **Toutes les interactions de quatrième ordre.** Insère une interaction de quatrième ordre (le produit des champs d'entrée) pour chaque combinaison possible de champs d'entrée sélectionnés (quatre à la fois). Par exemple, si vous avez sélectionné les champs d'entrée A , B , C , D et E dans la liste Champs disponibles, cette méthode insère les caractéristiques $A * B * C * D$, $A * B * C * E$, $A * B * D * E$, $A * C * D * E$ et $B * C * D * E$.

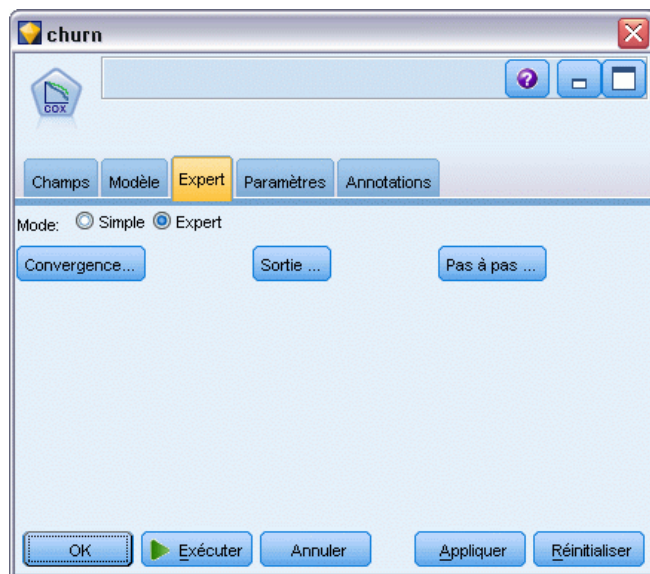
Champs disponibles. Répertoire les champs d'entrée disponibles à utiliser lors de la création des caractéristiques de modèle. Veuillez noter que cette liste peut inclure des champs qui ne sont pas des champs d'entrée valides. Vérifiez que toutes les caractéristiques de modèle incluent uniquement des champs d'entrée.

Aperçu. Affiche les caractéristiques qui seront ajoutées au modèle si vous cliquez sur Insérer, en fonction des champs sélectionnés et du type de caractéristique sélectionné ci-dessus.

Insérer. Insère les caractéristiques du modèle (en fonction des champs et du type de caractéristique sélectionnés), puis ferme la boîte de dialogue.

Noeud de Cox - Options expert

Figure 10-77
Boîte de dialogue du noeud de Cox, onglet Expert



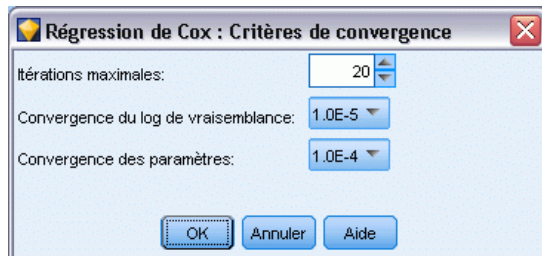
Convergence. Ces options vous permettent de contrôler les paramètres de convergence des modèles. Lorsque vous exécutez le modèle, les paramètres de convergence contrôlent le nombre d'exécutions répétées des différents paramètres pour déterminer leur degré d'adéquation. Plus les essais de paramètres sont nombreux, plus les résultats sont proches (en d'autres termes, plus les résultats convergent). [Pour plus d'informations, reportez-vous à la section Critères de convergence du noeud de Cox sur p. 377.](#)

Résultats. Ces options vous permettent de demander des statistiques et des nuages supplémentaires, y compris la courbe de survie, qui apparaîtront dans les sorties avancées du modèle généré par le noeud. [Pour plus d'informations, reportez-vous à la section Options de sortie avancée du noeud de Cox sur p. 377.](#)

Pas à pas. Ces options vous permettent de contrôler les critères d'ajout et de suppression des champs avec la méthode d'estimation Pas à pas. (Le bouton est désactivé si la méthode Entrée est sélectionnée.) [Pour plus d'informations, reportez-vous à la section Critères d'analyse pas à pas du noeud de Cox sur p. 378.](#)

Critères de convergence du noeud de Cox

Figure 10-78
Boîte de dialogue Critères de convergence du noeud Cox



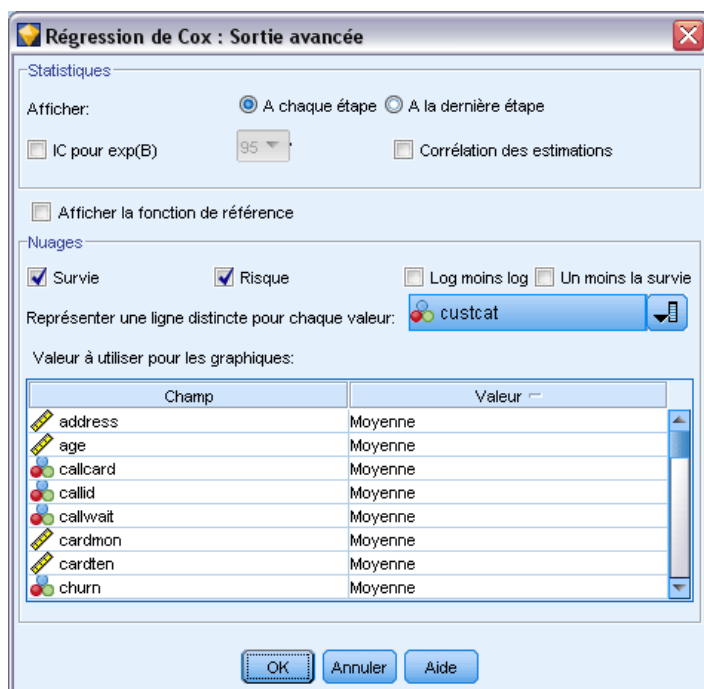
Nombre maximum d'itérations : Permet de spécifier le maximum d'itérations pour le modèle, ce qui commande la durée de la procédure pour la recherche d'une solution.

Convergence de log-vraisemblance : Les itérations sont interrompues lorsque la modification relative du log de vraisemblance est inférieure à la valeur indiquée. Le critère n'est pas utilisé si la valeur est 0.

Convergence des paramètres : Les itérations sont interrompues lorsque la modification absolue ou relative des estimations des paramètres est inférieure à la valeur indiquée. Le critère n'est pas utilisé si la valeur est 0.

Options de sortie avancée du noeud de Cox

Figure 10-79
Boîte de dialogue Sortie avancée de la régression de Cox



Statistics. Vous pouvez obtenir des statistiques sur les paramètres du modèle, y compris des intervalles de confiance pour $\exp(B)$ et la corrélation des estimations. Vous pouvez afficher les statistiques du modèle à chaque pas ou uniquement au dernier pas.

Afficher la fonction de référence : Permet d'afficher la fonction de base de hasard et la survie cumulée à la moyenne des covariables.

Diagrammes

Les diagrammes permettent d'évaluer le modèle estimé et d'interpréter les résultats. Vous ne pouvez pas représenter les fonctions Survie, Hasard, LN (-Logn), et Un moins survie.

- **Survie.** Affiche la fonction de survie cumulée d'après une échelle linéaire.
- **Hasard.** Affiche la fonction de hasard cumulée d'après une échelle linéaire.
- **LN (-Logn).** Affiche l'estimation cumulative de survie après que la transformation $\ln(-\ln)$ lui ait été appliquée.
- **Un moins survie.** Trace un moins la fonction de survie sur une échelle linéaire.

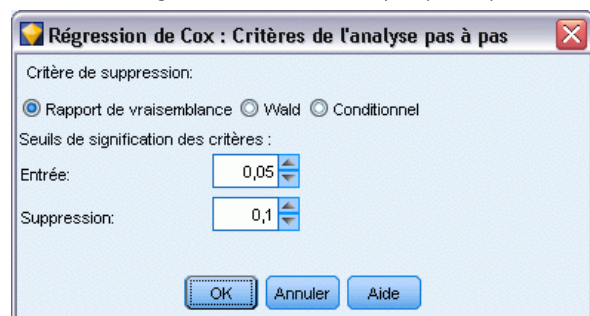
Affiche une ligne séparée pour chaque valeur. Cette option est disponible uniquement pour les champs catégoriels.

Valeur à utiliser pour les nuages. Parce que ces fonctions dépendent des valeurs des variables prédites, vous devez utiliser des valeurs de variables prédites constantes pour afficher ces fonctions et leur durée. L'option par défaut est d'utiliser la moyenne de chaque valeur prédite comme valeur constante, mais vous pouvez saisir vos propres valeurs de nuage à l'aide de la grille. Le codage d'indicateur est utilisé pour les entrées catégorielles. Par conséquent, il y a un coefficient de régression pour chaque catégorie (sauf la dernière). Ainsi, une entrée catégorielle a une valeur moyenne pour chaque contraste d'indicateur égal à la proportion des cas dans la catégorie correspondante au contraste d'indicateur.

Critères d'analyse pas à pas du noeud de Cox

Figure 10-80

Boîte de dialogue Critères de l'analyse pas à pas de la régression de Cox

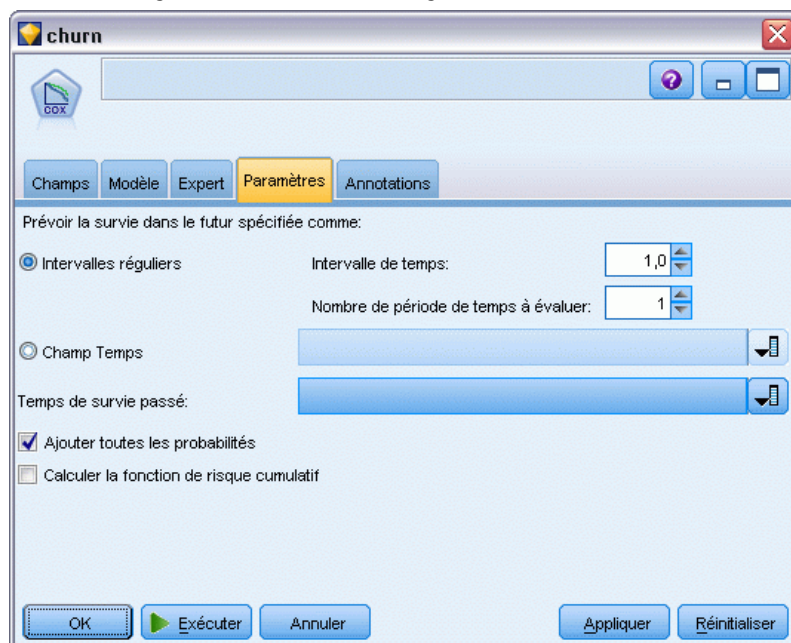


Critère de suppression. Sélectionnez Rapport de vraisemblance pour obtenir un modèle plus fiable. Pour réduire la durée de création de ce modèle, vous pouvez essayer de sélectionner Wald. L'option supplémentaire Conditionnel fournit un test de suppression basé sur la probabilité de la statistique du rapport de vraisemblance, en fonction des estimations des paramètres conditionnels.

Seuils de signification des critères. Cette option vous permet de spécifier des critères de sélection basés sur la probabilité statistique (la valeur p) associée à chaque champ. Les champs sont ajoutés au modèle uniquement si la valeur p associée est inférieure à la valeur spécifiée dans le champ Entrée ; ils sont supprimés uniquement si la valeur p est supérieure à la valeur spécifiée dans le champ Suppression. La valeur du champ Entrée doit être inférieure à la valeur du champ Suppression.

Options des paramètres du noeud de Cox

Figure 10-81
Boîte de dialogue du noeud de Cox, onglet Paramètres



Prédire la survie à des moments ultérieurs. Spécifie un ou plusieurs moments ultérieurs. La survie, c'est-à-dire si chaque cas est susceptible de survivre pour au moins cette durée (à partir de maintenant) sans que l'évènement terminal ne se produise, est prédite pour chaque enregistrement à chaque valeur temporelle, une prédiction par valeur temporelle. Veuillez noter que la survie est la valeur "fausse" du champ cible.

- **Intervalles réguliers.** Les valeurs temporelles de survie sont générées à partir de l'intervalle de temps spécifié et par le nombre de périodes à déterminer. Par exemple, si 3 périodes sont demandées avec un intervalle de 2 entre chaque période, la survie sera prédite pour les moments ultérieurs 2, 4, 6. Chaque enregistrement est évalué aux mêmes périodes temporelles.
- **Champs Temps.** Les durées de survie sont fournies pour chaque enregistrement dans chaque champ temporel choisi (un champ de prédiction est généré). Par conséquent, chaque enregistrement peut être évalué à différents moments.

Durée de survie antérieure. Définir la durée de survie de l'enregistrement jusqu'à maintenant—par exemple, la durée d'affectation d'un client existant comme champ. L'évaluation de la probabilité de survie à un moment ultérieur sera dépendant de la durée de survie antérieure.

Remarque : Les valeurs des durées de survie ultérieures et antérieures doivent être comprises dans l'intervalle des durées de vie des données utilisées pour former le modèle. Les enregistrements dont les durées de survie ne sont pas dans cet intervalle sont considérés comme nuls.

Ajouter toutes les probabilités. Spécifie l'ajout ou le non-ajout des probabilités de chacune des catégories du champ de sortie à chacun des enregistrements traités par le noeud. Lorsqu'elle ne l'est pas, seule la probabilité de la catégorie prédite est ajoutée. Les probabilités sont calculées pour chaque moment ultérieur.

Calculer la fonction cumulative des risques. Spécifie si la valeur des risques cumulatifs est ajoutée à chaque enregistrement. Les risques cumulatifs sont calculés pour chaque moment ultérieur.

Nugget de modèle Cox

Les modèles de régression de Cox représentent les équations estimées par les noeuds de Cox. Ils contiennent toutes les informations rassemblées par le modèle, ainsi que des informations sur la structure et les performances du modèle.

Lorsque vous exécutez un flux contenant un modèle de régression de Cox généré, le noeud crée deux champs contenant la prévision du modèle et la probabilité associée. Les noms des nouveaux champs sont dérivés du nom du champ de sortie prédit, précédé du préfixe *\$C-* pour la catégorie prédite *\$CP-* pour la probabilité associée, suivi du chiffre de l'intervalle du moment ultérieur ou du nom du champ temporel qui définit l'intervalle temporel. Par exemple, pour un champ de sortie nommé *attrition* et deux intervalles de temps ultérieurs définis à intervalle régulier, les nouveaux champs seront appelés *\$C-attrition-1*, *\$CP-attrition-1*, *\$C-attrition-2*, et *\$CP-attrition-2*. Si les moments ultérieurs sont définis avec un champ de temps *durée d'affectation*, les nouveaux champs seront *\$C-attrition_durée d'affectation* et *\$CP-attrition_durée d'affectation*.

Si vous avez sélectionné l'option de paramétrage *Ajouter toutes les probabilités* dans le noeud de Cox, deux champs supplémentaires seront ajoutés pour chaque moment ultérieur, contenant les probabilités de survie et d'échec pour chaque enregistrement. Ces champs supplémentaires sont nommés en fonction du nom du champ de sortie, précédé du préfixe *\$CP-<valeur faux>* pour la probabilité de survie et *\$CP-<valeur vrai>* pour la probabilité que l'évènement se soit produit, suivi du chiffre de l'intervalle de temps ultérieur. Par exemple, pour un champ de sortie où la valeur "faux" est 0 et la valeur "vrai" est 1, et deux intervalles de temps ultérieurs définis à intervalles réguliers, les nouveaux champs seront nommés *\$CP-0-1*, *\$CP-1-1*, *\$CP-0-2*, et *\$CP-1-2*. Si les moments ultérieurs sont définis avec un champ de temps unique *durée d'affectation*, les nouveaux champs seront *\$CP-0-1* et *\$CP-1-1*, car il n'y a qu'un seul intervalle ultérieur.

Si vous avez sélectionné l'option de paramétrage *Calculer la fonction cumulative des risques* dans le noeud de Cox, un champ supplémentaire sera ajouté pour chaque moment ultérieur, contenant la fonction cumulative de risques pour chaque enregistrement. Ces champs supplémentaires sont nommés en fonction du nom du champ de sortie, précédé du préfixe *\$CH-*, suivi du chiffre de l'intervalle de temps ultérieur. Par exemple, pour un champ de sortie nommé *attrition* et deux intervalles de temps ultérieurs définis à intervalle régulier, les nouveaux champs seront appelés *\$CH-attrition-1* et *\$CH-attrition-2*. Si les moments ultérieurs sont définis avec un champ de temps *durée d'affectation*, le nouveau champ sera *\$CH-attrition-1*.

Paramètres de sortie de la régression de Cox

L'onglet Paramètres du nugget contient les mêmes commandes que l'onglet Paramètres du noeud du modèle. Les valeurs par défaut des commandes du nugget sont déterminées par les valeurs définies dans le noeud du modèle. [Pour plus d'informations, reportez-vous à la section Options des paramètres du noeud de Cox sur p. 379.](#)

Sorties avancées du noeud Régression de Cox

La sortie avancée de la régression de Cox vous fournit des informations détaillées sur le modèle évalué et ses performances, y compris la courbe de survie. Dans la mesure où la plupart des informations contenues dans la sortie avancée ont un caractère technique, il est nécessaire d'avoir une bonne connaissance de la régression de Cox pour pouvoir les interpréter correctement.

Figure 10-82
Nugget de modèle Cox - Onglet Avancé

The screenshot shows the 'Scoring' dialog box with the 'Options avancées' tab selected. The 'Case Processing Summary' table is as follows:

Case Processing Summary		N	Percent
Cases available in analysis	Event(a)	274	27.4%
	Censored	726	72.6%
	Total	1000	100.0%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	0	.0%
	Total	0	.0%
Total		1000	100.0%

a. Dependent Variable: Months with service

The 'Categorical Variable Codings' table for 'marital(t)' is as follows:

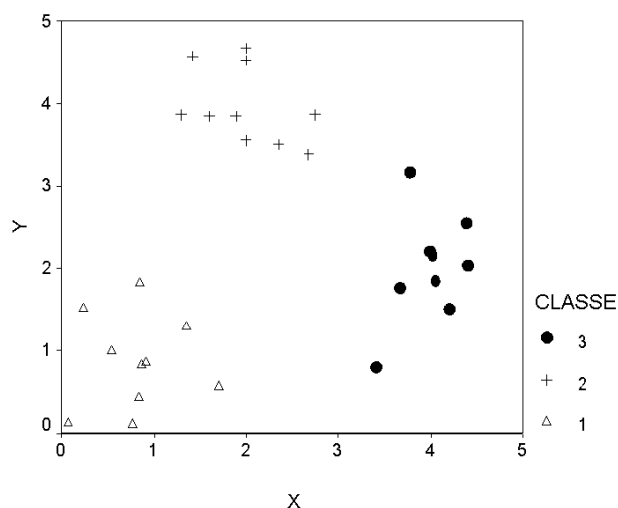
Categorical Variable Codings(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r)		Frequency	(1)(s)	(2)	(3)	(4)
marital(t)	0=Unmarried	505	1			
	1=Married	495	0			

Modèles de classification

Les modèles de classification se chargent essentiellement d'identifier des groupes d'enregistrements similaires et de répertorier les enregistrements en fonction du groupe auquel ils appartiennent. Cette opération est effectuée sans l'aide des connaissances existantes sur les groupes et leurs caractéristiques. A dire vrai, vous ne connaîtrez peut-être même pas le nombre de groupes à rechercher. Ceci permet de distinguer les modèles de classification des autres techniques d'apprentissage automatique—. En outre, il n'y a aucun champ de sortie ou cible prédéfini pour le modèle à prévoir. Ces modèles sont souvent désignés sous le nom de modèles d'**apprentissage non supervisé** du fait de l'absence de norme externe permettant d'évaluer les performances de classification du modèle. Il n'y a aucune réponse *vraie* ni *fausse* pour ces modèles. Leur valeur est déterminée par leur capacité à capturer des groupements intéressants dans les données et ils fournissent des descriptions utiles de ces mêmes groupements.

Les méthodes de classification reposent sur la mesure de la distance entre les enregistrements et les classes. Les enregistrements sont affectés à des classes de façon à réduire au maximum la distance entre les enregistrements appartenant à la même classe.

Figure 11-1
Modèle de classification simple



Trois méthodes de classification sont fournies :



Le noeud K-Means classe l'ensemble de données dans différents groupes (ou classes). La méthode définit un nombre de classes fixe, affecte à plusieurs reprises des enregistrements à des classes et ajuste les centres de classe, jusqu'à ce que le modèle ne puisse plus être amélioré. Au lieu de tenter de prédire un résultat, le modèle *k*-means utilise un processus connu sous le nom d'apprentissage non supervisé pour découvrir des tendances dans l'ensemble de champs d'entrée. [Pour plus d'informations, reportez-vous à la section Noeud K-Means sur p. 390.](#)



Le noeud TwoStep utilise une méthode de classification non supervisée en deux étapes. La première étape consiste en une exploration des données visant à compresser les données d'entrée brutes en sous-classes plus faciles à manipuler. Au cours de la seconde étape, l'utilisation d'une méthode de classification hiérarchique permet de fusionner progressivement les sous-classes en classes de plus en plus importantes. La technique TwoStep a l'avantage d'évaluer automatiquement le nombre de classes optimal pour les données d'apprentissage. Il peut prendre en charge de manière efficace des types de champ mixtes et des ensembles de données volumineux. [Pour plus d'informations, reportez-vous à la section Noeud Classification TwoStep sur p. 395.](#)



Le noeud Kohonen génère un type de réseau de neurones qui peut être utilisé pour classer les données en groupes distincts. Lorsque l'apprentissage du réseau est terminé, les enregistrements similaires doivent être regroupés dans la connexion de sortie, tandis que les enregistrements différents sont à l'opposé. Vous pouvez étudier le nombre d'observations capturées par chaque unité du nugget de modèle afin d'identifier les unités fortes. Vous pouvez ainsi vous faire une idée du nombre de classes approprié. [Pour plus d'informations, reportez-vous à la section Noeud Kohonen sur p. 383.](#)

Les modèles de classification sont souvent utilisés pour créer des classes ou segments utilisés ensuite en tant qu'entrées dans les prochaines analyses. Un exemple souvent utilisé est celui des segments de marché utilisés par les spécialistes du marketing pour partitionner leur marché global en sous-groupes homogènes. Chaque segment dispose de caractéristiques spéciales qui influent sur le succès des efforts de marketing entrepris pour ce segment. Si vous utilisez le Data mining pour optimiser votre stratégie marketing, vous pouvez généralement améliorer votre modèle de façon significative en identifiant les segments appropriés et en utilisant les informations des segments dans vos modèles prédictifs.

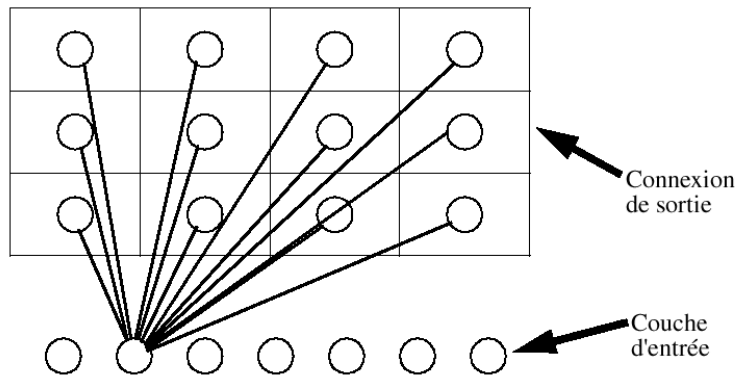
Noeud Kohonen

Les réseaux Kohonen représentent un type de réseau de neurones effectuant des opérations de classification non supervisée. Ils sont également appelés **knet** ou **cartes auto-organisatrices**. Ce type de réseau permet de classer les données en groupes distincts lorsqu'aucun groupe n'est défini au départ. Les enregistrements similaires sont rassemblés dans le même groupe ou la même classe.

Les unités de base sont les **neurones**. Ils sont organisés en deux couches : la **couche d'entrée** et la **couche de sortie** (également appelée **connexion de sortie**). Tous les neurones d'entrée sont connectés à tous les neurones de sortie. Ces connexions ont une **puissance** ou une **pondération** associée. Au cours de l'apprentissage, chaque unité entre en compétition avec les autres pour « gagner » des enregistrements.

La connexion de sortie est une grille de neurones à deux dimensions, sans connexion entre les unités. Vous trouverez ci-dessous la représentation d'une connexion 3×4 (les connexions sont généralement plus grandes).

Figure 11-2
Structure d'un réseau Kohonen



Tous les neurones entrants sont connectés à tous les neurones de la carte de résultat (non représentée par souci de clarté)

Les données d'entrée sont présentées à la couche d'entrée et les valeurs sont diffusées vers la couche de sortie. Le neurone de sortie avec la réponse la plus forte l'**emporte** et celle-ci est adoptée comme réponse pour cette entrée.

Au départ, tous les coefficients de pondération sont aléatoires. Lorsqu'une unité gagne un enregistrement, ses poids (ainsi que ceux des unités proches, regroupées sous le terme de **voisinage**) sont ajustés de façon à mieux correspondre au modèle des valeurs prédites de cet enregistrement. Tous les enregistrements d'entrée sont affichés et les coefficients de pondération sont mis à jour en conséquence. Ce processus est répété autant de fois que nécessaire jusqu'à ce que les changements soient minimes. Au fil de l'apprentissage, les poids sur les unités de la grille sont ajustés de façon à former une « carte » bidimensionnelle des classes.

Lorsque l'apprentissage du réseau est terminé, les enregistrements similaires doivent être regroupés dans la connexion de sortie, tandis que les enregistrements différents sont à l'opposé.

Contrairement à la plupart des autres méthodes d'apprentissage disponibles dans IBM® SPSS® Modeler, les réseaux Kohonen n'utilisent *pas* de champ cible. Ce type d'apprentissage qui n'utilise aucun champ cible est appelé **apprentissage non supervisé**. Au lieu de tenter de prédire un résultat, les réseaux Kohonen tentent de découvrir des tendances dans l'ensemble de champs d'entrée. En général, à la fin de l'apprentissage, un réseau Kohonen comporte un petit nombre d'unités qui résument un grand nombre d'observations (les unités **fortes**), et plusieurs autres unités qui ne correspondent à aucune observation particulière (les unités **faibles**). Les unités fortes (et parfois les unités adjacentes dans la grille) représentent des centres de classe probables.

Vous pouvez également utiliser les réseaux Kohonen via la **réduction dimensionnelle**. La caractéristique spatiale de la grille bidimensionnelle fournit un mappage entre les k valeurs prédites d'origine et deux fonctions dérivées qui préservent les relations de similarité des valeurs prédites. Dans certains cas, cela se révèle aussi avantageux qu'une analyse factorielle ou une ACP.

Remarquez que la méthode de calcul de la taille par défaut de la grille de sortie a été modifiée par rapport aux précédentes versions SPSS Modeler. La nouvelle méthode produira généralement des couches de sortie plus petites, dont l'apprentissage est plus rapide et qui produisent une

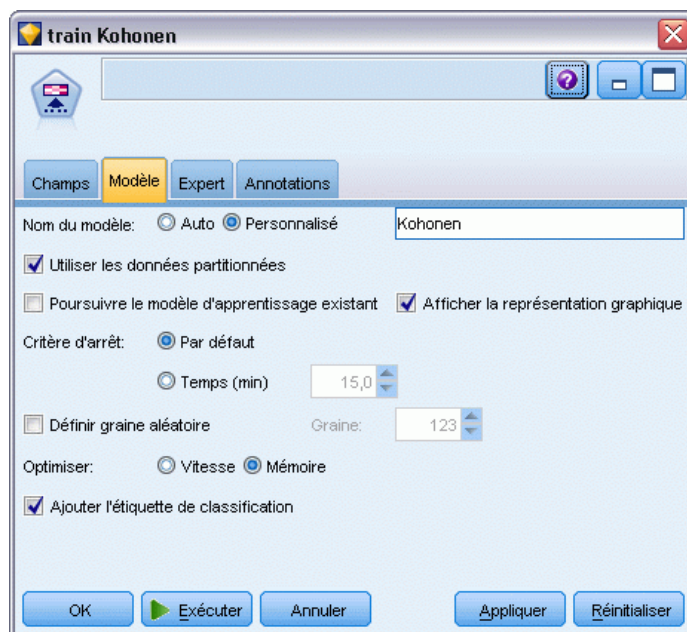
meilleure généralisation. Si vous estimez que les résultats obtenus avec la taille par défaut ne sont pas satisfaisants, essayez d'augmenter la taille de la grille de sortie dans l'onglet Expert. [Pour plus d'informations, reportez-vous à la section Noeud Kohonen - Options expert sur p. 387.](#)

Conditions requises. Pour former un réseau Kohonen, vous avez besoin d'un ou de plusieurs champs dont le rôle est configuré sur *Entrée*. Les champs dont le rôle est configuré sur *Cible*, *Les deux* ou *Aucun* sont ignorés.

Puissance. Il n'est pas nécessaire de disposer de données sur l'affectation des groupes pour générer un modèle de réseau Kohonen. Vous n'avez même pas besoin de connaître le nombre de groupes à rechercher. Les réseaux Kohonen comportent au départ un grand nombre d'unités qui, au fil de l'apprentissage, sont attirées par les classes naturelles des données. Vous pouvez étudier le nombre d'observations capturées par chaque unité du nugget de modèle afin d'identifier les unités fortes, ce qui vous donnera une idée du nombre de classes approprié.

Noeud Kohonen - Options du modèle

Figure 11-3
Options du modèle de noeud Kohonen



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.](#)

Poursuivre le modèle d'apprentissage existant. Par défaut, chaque exécution d'un noeud Kohonen engendre la création d'un réseau. Si vous désactivez cette option, l'apprentissage se poursuit sur le dernier réseau généré par le noeud.

Afficher la représentation graphique. Lorsque cette option est sélectionnée, un graphique apparaît, montrant l'évolution de la répartition, en deux dimensions, au cours de l'apprentissage. La force de chaque noeud est représentée par une couleur. Les unités qui gagnent beaucoup d'enregistrements apparaissent en rouge (unités **fortes**) tandis que les unités qui en gagnent peu, voire aucun, apparaissent en blanc (unités **faibles**). Il est possible que les commentaires ne s'affichent pas si le temps nécessaire à la construction du modèle est relativement bref. Cette fonctionnalité peut ralentir l'apprentissage. Pour accélérer le temps d'apprentissage, désélectionnez cette option.

Figure 11-4
Représentation graphique de Kohonen



Critère d'arrêt. Le critère d'arrêt de l'apprentissage par défaut est basé sur des paramètres internes. Le temps peut également constituer un critère d'arrêt. Entrez le temps (en minutes) d'apprentissage du réseau.

Définir graine aléatoire. Si aucune valeur de départ aléatoire n'est définie, la séquence de valeurs aléatoires permettant d'initialiser les pondérations du réseau est différente à chaque exécution du noeud. Le noeud crée alors des modèles différents à chaque fois qu'il est exécuté, même si les paramètres du noeud et les valeurs des données restent inchangés. Lorsque vous sélectionnez cette option, vous pouvez définir la valeur de départ aléatoire de façon à ce que le modèle généré soit entièrement reproductible. Une valeur de départ aléatoire spécifique génère toujours la même séquence de valeurs aléatoires : l'exécution du noeud produira donc toujours le même modèle.

Remarque : Lorsque vous utilisez l'option Définir graine aléatoire avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la graine aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. [Pour plus d'informations, reportez-vous à la section Noeud Trier dans le chapitre 3 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Remarque : Si vous souhaitez inclure des champs nominaux (ensemble) dans votre modèle, mais que vous rencontrez des problèmes de mémoire pour la construction du modèle ou que la construction du modèle est trop lente, réfléchissez à un autre mode de codage des grands champs d'ensemble afin de réduire le nombre de valeurs ou bien utilisez un autre champ, contenant moins de valeurs comme proxy pour un large ensemble. Par exemple, si vous rencontrez un problème avec un *champ id_produit* contenant des valeurs pour différents produits, vous pouvez le supprimer du modèle et le remplacer par un champ *catégorie_produit* moins détaillé.

Optimiser. Sélectionnez, à votre convenance, les options destinées à accroître les performances du système au cours de la création de modèles.

- Sélectionnez Vitesse pour que l'algorithme n'utilise jamais le débordement sur disque, afin d'améliorer les performances.
- Sélectionnez Mémoire pour que l'algorithme utilise le débordement sur disque en cas de besoin, ce qui a pour effet de réduire la vitesse. Par défaut, cette option est sélectionnée.

Remarque : Si vous exécutez le système en mode réparti, ce paramètre peut être ignoré par les options administrateur indiquées dans le fichier *options.cfg*. [Pour plus d'informations, reportez-vous à la section Utilisation du fichier options.cfg dans le chapitre 4 dans Guide d'administration et des performances de IBM SPSS Modeler Server 15.](#)

Ajoutez une étiquette de classification. Sélectionnée par défaut pour les nouveaux modèles, mais désélectionnée pour les modèles chargés à partir de versions précédentes de IBM® SPSS® Modeler, cette option crée un champ de score catégoriel unique du même type que celui créé par les deux noeuds K-Means et TwoStep. Ce champ de type chaîne est utilisé dans le noeud Classification automatique lors du calcul de la mesure du classement pour les différents types de modèles. [Pour plus d'informations, reportez-vous à la section Noeud Classification automatique dans le chapitre 5 sur p. 117.](#)

Noeud Kohonen - Options expert

Les options expert, qui s'adressent aux utilisateurs ayant une bonne connaissance des réseaux Kohonen, permettent d'affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 11-5
Kohonen - Options expert



Largeur et Longueur. Indiquez la taille (largeur et longueur) de la connexion de sortie bidimensionnelle en tant que nombre d'unités de sortie le long de chaque dimension.

Dégradation du taux d'apprentissage. Indiquez si vous souhaitez utiliser une dégradation linéaire ou exponentielle. Le **taux d'apprentissage** est un facteur de pondération qui décroît au fil du temps, de sorte que le réseau commence par coder les caractéristiques les plus générales des données pour progressivement se concentrer sur les caractéristiques plus spécifiques.

Phase 1 et Phase 2. L'apprentissage d'un réseau Kohonen se déroule en deux phases. La phase 1 est une phase d'estimation qui permet d'évaluer les tendances générales des données. La phase 2 est une phase d'affinement qui permet d'ajuster la carte de façon à modéliser les caractéristiques spécifiques des données. Trois paramètres sont associés à chacune de ces phases :

- **Voisinage.** Définit la taille de départ (rayon) du voisinage. Ce paramètre détermine le nombre d'unités « avoisinantes » qui seront mises à jour avec l'unité qui gagne au cours de l'apprentissage. Au cours de la phase 1, la taille de voisinage passe de la valeur *Taille Voisin 1* à la valeur (*Taille Voisin 2 + 1*). Au cours de la phase 2, la taille de voisinage passe de la valeur *Taille Voisin 2* à la valeur 1. La valeur de *Taille Voisin 1* doit être supérieure à la valeur de *Taille Voisin 2*.
- **Eta initial.** Définit la valeur de départ du taux d'apprentissage *eta*. Au cours de la phase 1, la valeur du paramètre *Eta* passe de *Phase 1 - Eta initial* à *Phase 2 - Eta initial*. Au cours de la phase 2, la valeur du paramètre *Eta* passe de *Phase 2 - Eta initial* à 0. La valeur de *Phase 1 - Eta initial* doit être supérieure à celle de *Phase 2 - Eta initial*.
- **Cycles.** Définit le nombre de cycles de chacune des phases d'apprentissage. Chacune des phases se poursuit pendant le nombre d'explorations des données spécifiées.

Nuggets de modèle Kohonen

Les nuggets de modèles Kohonen contiennent toutes les informations rassemblées par le réseau Kohonen formé, ainsi que des informations sur l'architecture du réseau.

Lorsque vous exécutez un flux contenant un nugget de modèle Kohonen, le noeud crée deux champs contenant les coordonnées X et Y de l'unité de la grille de sortie Kohonen qui a réagi le plus fortement à cet enregistrement. Les noms attribués aux nouveaux champs sont constitués du nom du modèle auquel les préfixes $\$KX-$ et $\$KY-$ sont ajoutés. Par exemple, si le nom de votre modèle est *Kohonen*, les nouveaux champs seront alors intitulés $\$KX-Kohonen$ et $\$KY-Kohonen$.

Pour mieux comprendre ce qui a été codé par le réseau Kohonen, cliquez sur l'onglet *Modèle* du navigateur du nugget de modèle. L'onglet *Afficheur (classe)* qui apparaît contient une représentation graphique des classes, champs et niveaux d'importance. [Pour plus d'informations, reportez-vous à la section Visualiseur de classes - Onglet Modèle sur p. 400.](#)

Si vous préférez visualiser les classes sous forme de grille, vous pouvez afficher les résultats du réseau Kohonen en traçant les champs $\$KX-$ et $\$KY-$ à l'aide d'un noeud *Nuage*. (Vous devez sélectionner *X-Agitation* et *Y-Agitation* dans le noeud *Nuage* pour empêcher les tracés des enregistrements de chaque unité de se superposer les uns aux autres.) Vous pouvez également superposer un champ symbolique dans le tracé afin de voir de quelle manière le réseau Kohonen a regroupé les données.

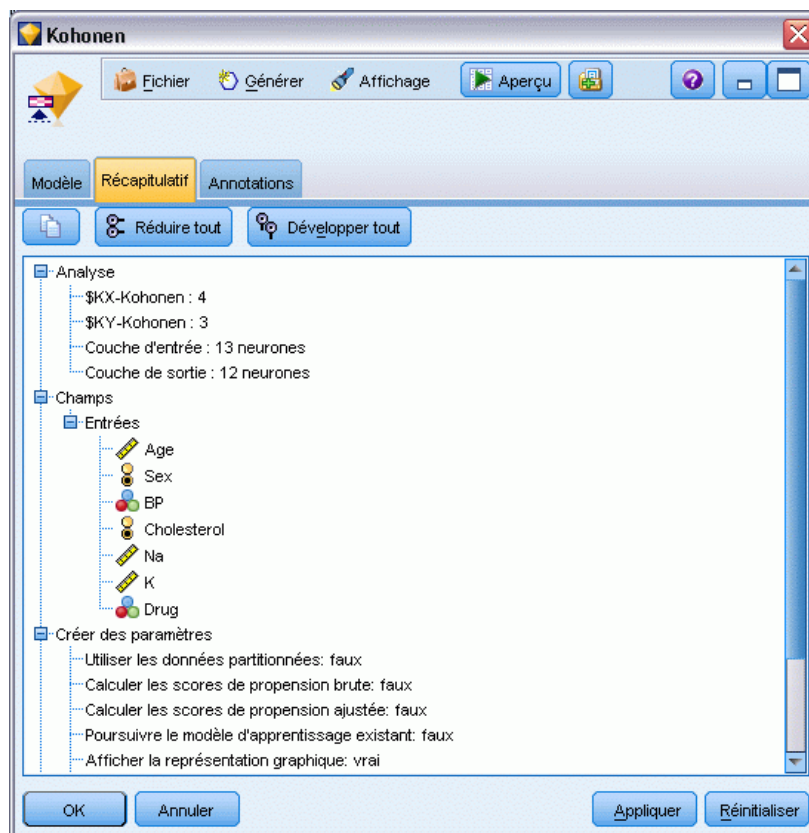
Une autre technique permettant d'évaluer le réseau Kohonen consiste à utiliser l'induction de règle afin d'identifier les caractéristiques qui distinguent les classes trouvées par le réseau [Pour plus d'informations, reportez-vous à la section Noeud C5.0 dans le chapitre 6 sur p. 180.](#)

Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle](#)

Récapitulatif du modèle Kohonen

L'onglet *Récapitulatif* d'un nugget de modèle Kohonen affiche des informations sur l'architecture ou la topologie du réseau. La longueur et la largeur de la fonction de cartographie bidimensionnelle Kohonen (la couche de sortie) sont indiquées par $\$KX-nom_du_modèle$ et $\$KY-nom_du_modèle$. Le nombre d'unités que contient chaque couche d'entrée et de sortie est indiqué.

Figure 11-6
Nugget du modèle Kohonen - Onglet Récapitulatif



Noeud K-Means

Le noeud K-Means fournit une méthode d'**analyse des classes**. Ce type de noeud permet de classer les données en groupes distincts lorsqu'aucun groupe n'est défini au départ. Contrairement à la plupart des méthodes d'apprentissage disponibles dans IBM® SPSS® Modeler, les modèles K-Means n'utilisent *pas* de champ cible. Ce type d'apprentissage qui n'utilise aucun champ cible est appelé **apprentissage non supervisé**. Le noeud K-Means n'essaie pas de générer des prévisions, mais tente de découvrir des tendances au sein des champs d'entrée. Les enregistrements sont rassemblés dans le même groupe ou la même classe.

Le noeud K-Means commence par définir un ensemble de centres de classes à partir des données. Il affecte ensuite chaque enregistrement à la classe à laquelle il s'apparente le plus, sur la base des valeurs du champ d'entrée de l'enregistrement. Une fois toutes les observations affectées, les centres de classes sont mis à jour afin de refléter le nouvel ensemble d'enregistrements affecté à chaque classe. Les enregistrements sont alors de nouveau évalués afin de déterminer si certains d'entre eux doivent être réaffectés à une autre classe. Ce processus se poursuit jusqu'à ce que le nombre maximal d'itérations soit atteint ou que le changement produit par une nouvelle itération soit inférieur à un seuil défini.

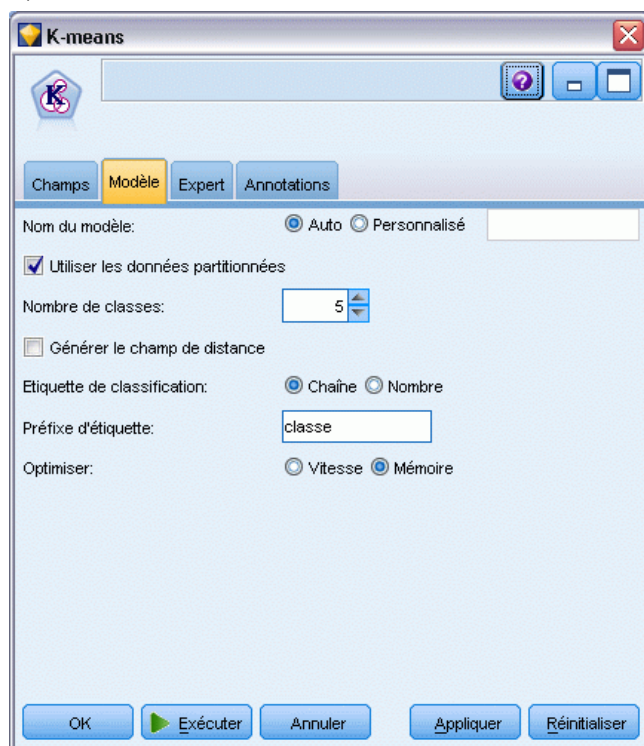
Remarque : le modèle ainsi généré dépend essentiellement de l'ordre des données d'apprentissage. Si vous réorganisez vos données et recréez le modèle, le modèle final risque d'être différent.

Conditions requises. Pour former un modèle K-Means, vous avez besoin d'un ou de plusieurs champs dont le rôle est configuré sur *Entrée*. Les champs dont le rôle est configuré sur *Sortie*, *Les deux* ou *Aucun* sont ignorés.

Puissance. Il n'est pas nécessaire de disposer de données sur les classes d'affectation pour générer un modèle K-Means. Le modèle K-Means s'avère souvent la méthode la plus rapide pour classer des ensembles de données volumineux.

Noeud K-Means - Options du modèle

Figure 11-7
Options du modèle de noeud 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.

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*.](#)

Nombre de classes spécifiées. Indiquez le nombre de classes à générer. La valeur par défaut est 5.

Générer le champ de distance. Si cette option est sélectionnée, le nugget de modèle comportera un champ contenant la distance de chaque enregistrement par rapport au centre de la classe à laquelle il a été affecté.

Étiquette de classification. Indiquez le format des valeurs du champ généré contenant les classes d'appartenance. La classe d'appartenance peut être indiquée sous la forme d'une chaîne accompagnée du préfixe d'étiquette (par exemple, "Cluster 1", "Cluster 2", etc.) ou d'un nombre.

Remarque : Si vous souhaitez inclure des champs nominaux (ensemble) dans votre modèle, mais que vous rencontrez des problèmes de mémoire pour la construction du modèle ou que la construction du modèle est trop lente, réfléchissez à un autre mode de codage des grands champs d'ensemble afin de réduire le nombre de valeurs ou bien utilisez un autre champ, contenant moins de valeurs comme proxy pour un large ensemble. Par exemple, si vous rencontrez un problème avec un *champ id_produit* contenant des valeurs pour différents produits, vous pouvez le supprimer du modèle et le remplacer par un champ *catégorie_produit* moins détaillé.

Optimiser. Sélectionnez, à votre convenance, les options destinées à accroître les performances du système au cours de la création de modèles.

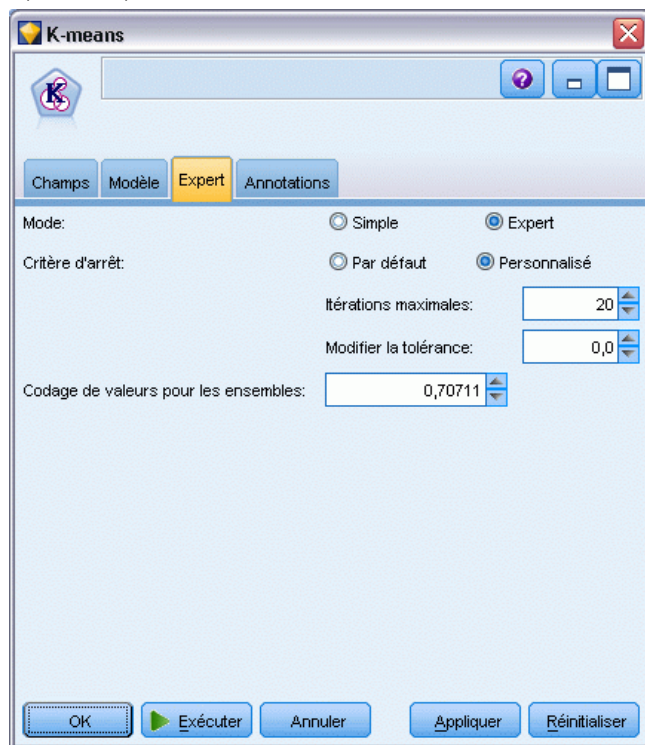
- Sélectionnez Vitesse pour que l'algorithme n'utilise jamais le débordement sur disque, afin d'améliorer les performances.
- Sélectionnez Mémoire pour que l'algorithme utilise le débordement sur disque en cas de besoin, ce qui a pour effet de réduire la vitesse. Par défaut, cette option est sélectionnée.

Remarque : Si vous exécutez le système en mode réparti, ce paramètre peut être ignoré par les options administrateur indiquées dans le fichier *options.cfg*. [Pour plus d'informations, reportez-vous à la section Utilisation du fichier options.cfg dans le chapitre 4 dans Guide d'administration et des performances de IBM SPSS Modeler Server 15.](#)

Noeud K-Means - Options expert

Les options expert, qui s'adressent aux utilisateurs ayant une connaissance approfondie des classifications *k-means*, permettent d'affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 11-8
Options expert de K-Means



Critère d'arrêt. Indiquez le critère d'arrêt à utiliser pour l'apprentissage du modèle. Le critère d'arrêt par défaut est 20 itérations ou un changement inférieur à 0.000001, le premier de ces événements se produisant étant pris en compte. Sélectionnez *Personnalisé* pour spécifier vos propres critères d'arrêt.

- **Nombre maximum d'itérations :** Cette option permet d'interrompre l'apprentissage du modèle lorsque le nombre d'itérations spécifié est atteint.
- **Modifier la tolérance.** Cette option permet d'interrompre l'apprentissage lorsque la modification des centres de classes la plus importante pour une itération est inférieure au niveau indiqué.

Codage de valeurs pour les ensembles. Indiquez une valeur comprise entre 0 et 1 qui sera utilisée pour recoder les champs d'ensemble en groupes de champs numériques. La valeur par défaut est la racine carrée de 0,5 (approximativement 0,707107), ce qui fournit la pondération appropriée pour les champs booléens recodés. Des valeurs proches de 1 ont pour effet de pondérer davantage les champs d'ensemble par rapport aux champs numériques.

Nuggets du modèle K-Means

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*.

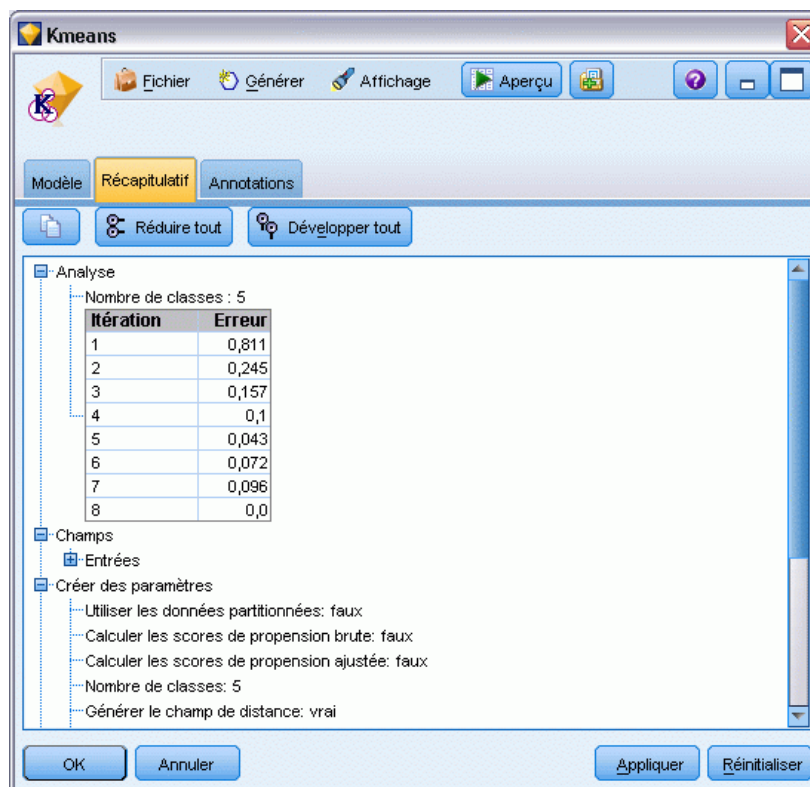
Une autre technique permettant d'évaluer le modèle K-means est disponible. Elle consiste à utiliser l'induction de règle afin d'identifier les caractéristiques qui distinguent les classes trouvées par le modèle. [Pour plus d'informations, reportez-vous à la section Noeud C5.0 dans le chapitre 6 sur p. 180.](#) Vous pouvez également cliquer sur l'onglet Modèle du navigateur du nugget de modèle pour afficher l'onglet Afficheur (classe). Vous obtenez alors une représentation graphique des classes, des champs et des niveaux d'importance. [Pour plus d'informations, reportez-vous à la section Visualiseur de classes - Onglet Modèle sur p. 400.](#)

Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle](#)

Récapitulatif du modèle K-Means

L'onglet Récapitulatif du nugget de modèle K-Means contient des informations sur les données d'apprentissage, le processus d'estimation et les classes définies par le modèle. Le nombre de classes, ainsi que l'historique d'itération, sont indiqués. Si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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.*](#)

Figure 11-9
Echantillon du nugget de modèle K-Means - Onglet Récapitulatif



Noeud Classification TwoStep

Le noeud Classification TwoStep permet une sorte d'**analyse de classes**. Ce type de noeud permet de classer les données en groupes distincts lorsqu'aucun groupe n'est défini au départ. Comme pour les noeuds Kohonen et K-Means, les modèles de classification TwoStep n'utilisent *aucun* champ cible. Le noeud Classification TwoStep n'essaie pas de générer des prévisions, mais tente de découvrir des tendances au sein des champs d'entrée. Les enregistrements sont rassemblés dans le même groupe ou la même classe.

La méthode Classification TwoStep se déroule en deux étapes. La première étape consiste en une exploration des données au cours de laquelle les données d'entrée brutes sont compressées en sous-classes plus faciles à manipuler. Au cours de la seconde étape, l'utilisation d'une méthode de classification hiérarchique permet de fusionner progressivement les sous-classes en classes de plus en plus importantes, sans qu'un nouvel examen des données soit nécessaire. Avec la classification hiérarchique, il n'est pas nécessaire de sélectionner à l'avance le nombre de classes. De nombreuses méthodes de classification non supervisée hiérarchique prennent comme classes de départ des enregistrements individuels qu'elles fusionnent ensuite pour générer des classes de plus en plus importantes. Même si de telles méthodes se révèlent souvent inefficaces face à de grands volumes de données, l'étape de pré-classification de la méthode TwoStep permet d'accélérer le processus de classification.

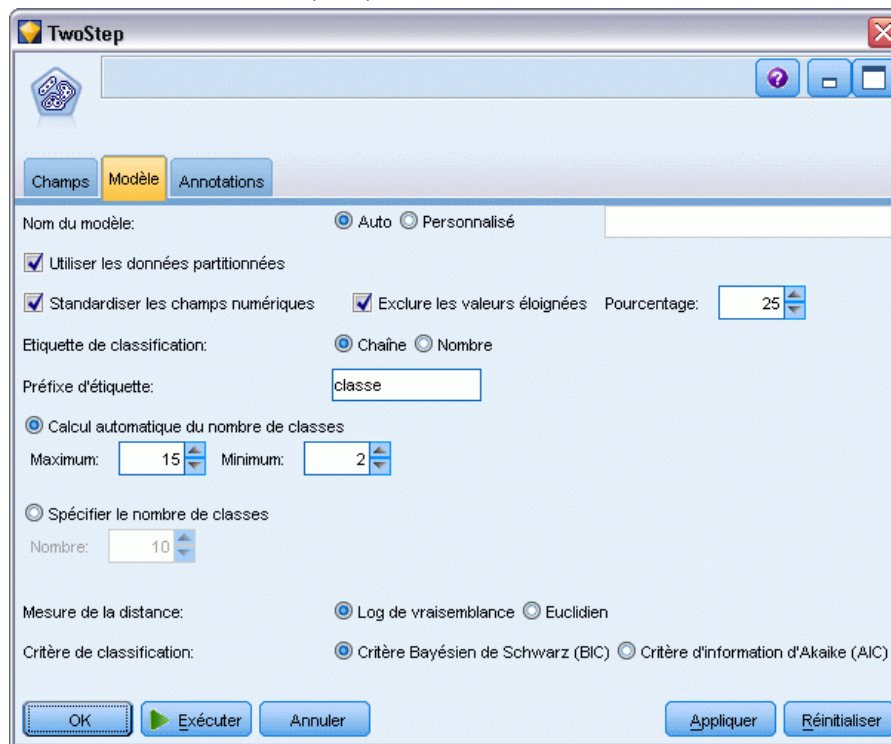
Remarque : le modèle ainsi généré dépend essentiellement de l'ordre des données d'apprentissage. Si vous réorganisez vos données et recréez le modèle, le modèle final risque d'être différent.

Conditions requises. Pour former un modèle de classification TwoStep, vous avez besoin d'un ou de plusieurs champs et dont le rôle est configuré sur *Entrée*. Les champs dont le rôle est configuré sur *Cible*, *Les deux* ou *Aucun* sont ignorés. L'algorithme de classification TwoStep ne traite pas les valeurs manquantes. Les enregistrements dont les champs d'entrée ne contiennent pas de valeur sont ignorés lors de la création du modèle.

Puissance. Le noeud de classification TwoStep peut prendre en charge de manière efficace des types de champ mixte et des volumes importants de données. Il est également capable de tester plusieurs solutions de classification pour choisir la plus performante, de sorte que vous n'avez plus besoin de spécifier au départ le nombre de classes. Le noeud Classification TwoStep peut également être configuré de façon à exclure automatiquement les **valeurs éloignées** ou extrêmement rares qui peuvent contaminer les résultats.

Noeud Classification TwoStep - Options du modèle

Figure 11-10
Noeud Classification TwoStep - 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*.](#)

Standardiser les champs numériques. Par défaut, le noeud Classification TwoStep convertit tous les champs d'entrée numériques selon une même échelle, avec une moyenne de 0 et une variance de 1. Pour conserver l'échelle d'origine des champs numériques, désélectionnez cette option. Les champs symboliques ne sont pas modifiés.

Exclure les valeurs éloignées. Si vous sélectionnez cette option, les enregistrements qui ne semblent pas adaptés à une classe substantive seront automatiquement retirés de l'analyse. Ceci permet d'empêcher que de tels cas ne viennent déformer les résultats.

La détection des valeurs éloignées intervient au cours de l'étape de préclassification. Lorsque cette option est sélectionnée, les sous-classes contenant moins d'enregistrements que les autres sous-classes sont considérées comme des valeurs éloignées potentielles et l'arbre des sous-classes recréé ne contient plus ces enregistrements. La taille en dessous de laquelle les sous-classes sont considérées comme valeurs éloignées potentielles est contrôlée par l'option Pourcentage Si certains enregistrements considérés comme des valeurs éloignées potentielles sont suffisamment similaires à l'un des nouveaux profils de sous-classe, vous pouvez les ajouter aux sous-classes recréées. Les valeurs éloignées potentielles qui ne peuvent pas être fusionnées sont considérées comme des valeurs éloignées. Elles sont ajoutées à une classe « parasite » et retirées de l'étape de classification non supervisée hiérarchique.

Lorsque vous *évaluez* des données à l'aide d'un modèle TwoStep qui utilise le traitement des valeurs éloignées, les nouvelles observations pour lesquelles la distance avec la classe substantive la plus proche est supérieure à la distance de seuil fixée (basée sur le log de vraisemblance) sont considérés comme des valeurs éloignées et sont affectés à la classe « parasite » avec le nom -1.

Etiquette de classification. Indiquez le format du champ devant contenir les classes d'appartenance. La classe d'appartenance peut être indiquée sous la forme d'une chaîne accompagnée du préfixe d'étiquette (par exemple, "Cluster 1", "Cluster 2", etc.) ou d'un nombre.

Calcul automatique du nombre de classes. Le noeud Classification TwoStep peut analyser très rapidement un grand nombre de solutions de classification afin de déterminer le nombre optimal de classes pour les données d'apprentissage. Indiquez un intervalle de solutions que le modèle devra respecter (définissez un nombre maximal et un nombre minimal de classes). Le noeud Classification TwoStep utilise un processus en deux étapes pour déterminer le nombre optimal de classes. Au cours de la première étape, une limite supérieure du nombre de classes du modèle est sélectionnée en fonction de la modification du critère de Bayes au fil de l'ajout de nouvelles classes. Au cours de la deuxième étape, la modification de la distance minimale entre les classes est calculée pour tous les modèles présentant un nombre de classes inférieur à la solution minimale du critère de Bayes. La modification la plus importante permet d'identifier le modèle de classe final.

Spécifier le nombre de classes. Si vous connaissez le nombre de classes à inclure dans votre modèle, sélectionnez cette option et entrez ce nombre.

Mesure de distance. Cette sélection détermine la façon dont la similarité entre deux classes est calculée.

- **Log-vraisemblance** : La mesure de vraisemblance place une distribution de probabilité sur les variables. Les variables continues sont considérées comme étant distribuées normalement alors que les variables qualitatives sont considérées comme étant multinomiales. Toutes les variables sont considérées comme étant indépendantes.
- **Euclidienne** : La mesure euclidienne est la distance « en ligne droite » entre deux classes. Elle peut être utilisée uniquement lorsque toutes les variables sont continues.

Critère de classification : Cette sélection détermine la façon dont l'algorithme de classification automatique détermine le nombre de classes. Vous pouvez spécifier le critère d'information bayésien (BIC) ou le critère d'information d'Akaike (AIC).

Nuggets de modèle de classification TwoStep

Les nuggets de modèles de classification TwoStep contiennent toutes les informations rassemblées par le modèle de classification non supervisée, ainsi que des informations sur les données d'apprentissage et le processus d'estimation.

Lorsque vous exécutez un flux comportant un nugget de modèle de classification TwoStep, le noeud crée un champ contenant la classe d'appartenance de cet enregistrement. Le nom du nouveau champ est formé à partir du nom du modèle auquel le préfixe *\$T-* est ajouté. Par exemple, si le nom de votre modèle est *TwoStep*, le nouveau champ sera intitulé *\$T-TwoStep*.

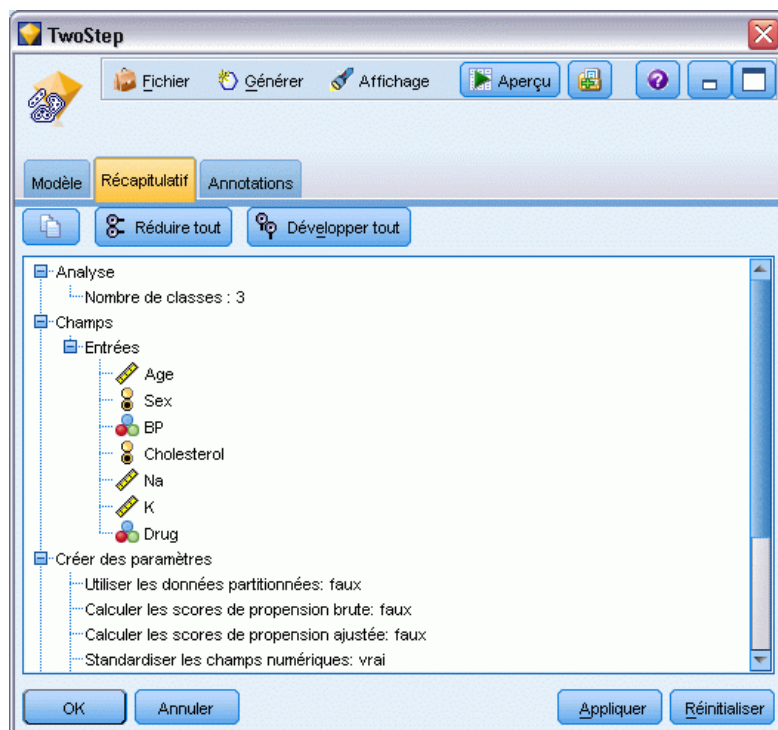
Une autre technique permettant d'évaluer le modèle TwoStep est disponible. Elle consiste à utiliser l'induction de règle afin d'identifier les caractéristiques qui distinguent les classes trouvées par le modèle. [Pour plus d'informations, reportez-vous à la section Noeud C5.0 dans le chapitre 6 sur p. 180.](#) Vous pouvez également cliquer sur l'onglet *Modèle* du navigateur du nugget de modèle pour afficher l'onglet *Afficheur (classe)*. Vous obtenez alors une représentation graphique des classes, des champs et des niveaux d'importance. [Pour plus d'informations, reportez-vous à la section Visualiseur de classes - Onglet Modèle sur p. 400.](#)

Pour obtenir des informations générales quant à l'utilisation du navigateur de modèle, reportez-vous à [Navigation dans les nuggets de modèle](#)

Récapitulatif du modèle TwoStep

L'onglet *Récapitulatif* d'un nugget de modèle de classification TwoStep affiche le nombre de classes détectées, ainsi que des informations sur les données d'apprentissage, le processus d'estimation et les paramètres de création utilisés.

Figure 11-11
Exemple : Echantillon du nugget de modèle de classification TwoStep - Onglet Récapitulatif



Pour plus d'informations, reportez-vous à la section Navigation dans les nuggets de modèle dans le chapitre 3 sur p. 55.

Le visualiseur de classes

Les modèles de classe sont généralement utilisés pour trouver des groupes (ou des classes) d'enregistrements similaires en fonction des variables examinées, où la similarité entre les membres d'un même groupe est élevée et où la similarité entre les membres de différents groupes est faible. Les résultats peuvent être utilisés pour identifier des associations qui ne seraient pas évidentes autrement. Par exemple, grâce à la classification des préférences des clients, du niveau de revenu et des habitudes d'achat, il peut être possible d'identifier les types de clients les plus susceptibles de répondre à une campagne de marketing particulière.

Il existe deux approches pour interpréter les résultats d'un affichage de classe :

- Examiner les classes afin de déterminer les caractéristiques uniques de cette classe. *Est-ce qu'une classe contient tous les emprunteurs à revenus élevés ? Est-ce que cette classe contient davantage d'enregistrements que les autres ?*
- Examiner les champs des classes afin de déterminer comment les valeurs sont distribuées parmi les classes. *Est-ce que le niveau d'éducation détermine l'appartenance à une classe ? Est-ce qu'une cote de solvabilité élevée permet de distinguer l'appartenance à une classe spécifique ?*

Grâce à l'utilisation des vues principales et des différentes vues liées dans le visualiseur de classes, vous pouvez avoir un bon aperçu qui vous aidera à répondre à ces questions.

Il est possible de générer les nuggets de modèle de classe suivants dans IBM® SPSS® Modeler :

- Nugget de modèle de réseau Kohonen
- Nugget de modèle de nuées dynamiques
- Nugget de modèle de classification TwoStep

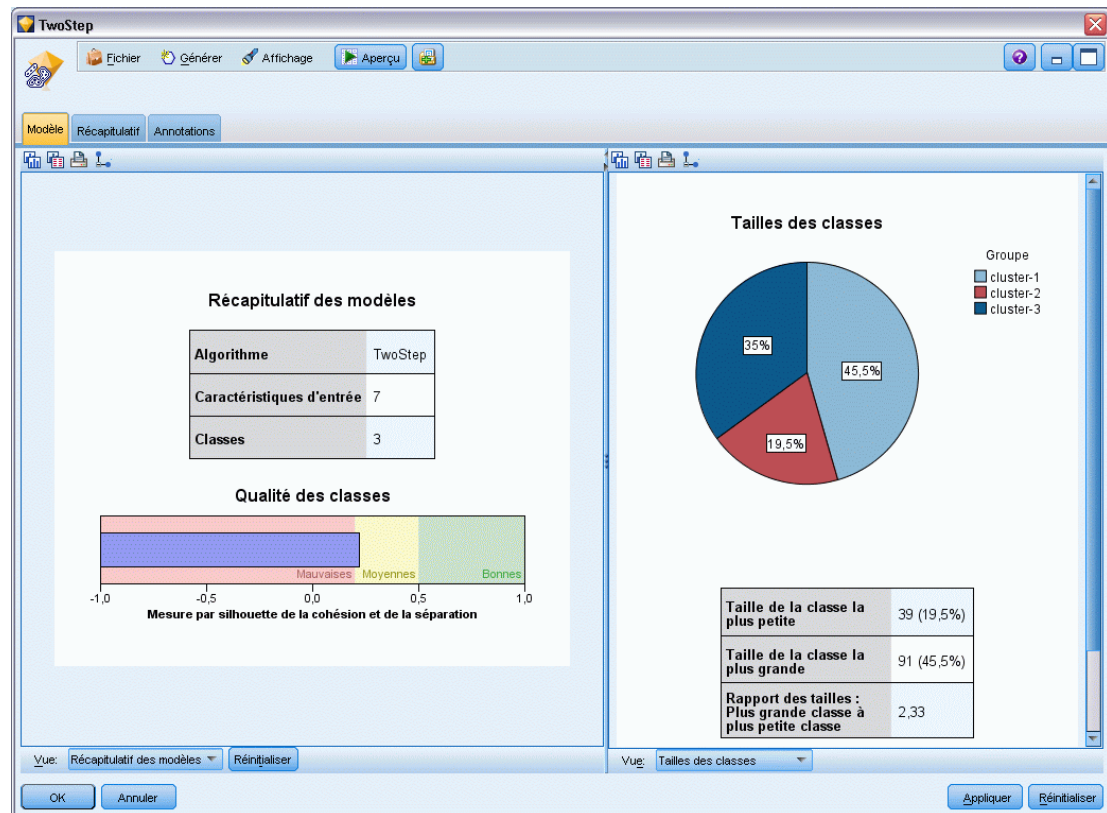
Pour consulter les informations concernant les nuggets de modèle de classe, cliquez avec le bouton droit de la souris sur le noeud de modèle et sélectionnez *Parcourir* dans le menu contextuel (ou *Modifier* pour les noeuds d'un flux). Vous pouvez aussi, si vous utilisez le noeud de modélisation de classification automatique, double-cliquer sur le nugget de classe requis dans le nugget de modèle de classification automatique. [Pour plus d'informations, reportez-vous à la section *Noeud Classification automatique* dans le chapitre 5 sur p. 117.](#)

Visualiseur de classes - Onglet Modèle

L'onglet *Modèle* des modèles de classe présente un affichage graphique des statistiques et des distributions récapitulatives des champs entre les classes, qui s'appelle **Visualiseur de classes**.

Remarque : L'onglet *Modèle* n'est pas disponible pour les modèles développés avec des versions de IBM® SPSS® Modeler antérieures à la version 13.

Figure 11-12
Visualiseur de modèles avec affichage par défaut



Le visualiseur de modèle est constitué de deux panneaux, l'affichage principal à gauche et la vue liée, ou auxiliaire, à droite. Il existe deux vues principales :

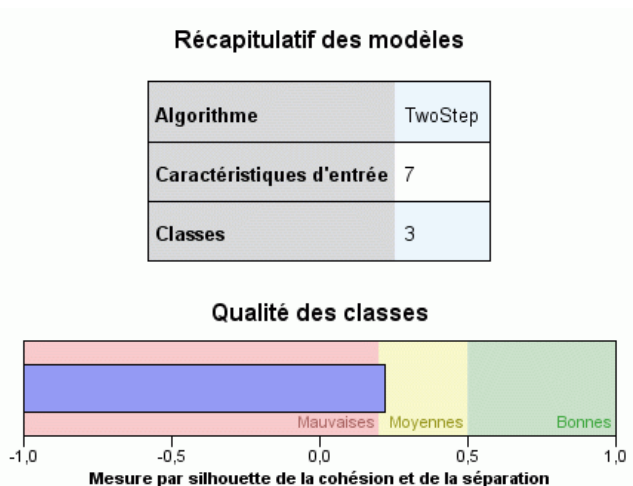
- Récapitulatif du modèle (par défaut). Pour plus d'informations, reportez-vous à la section [Vue récapitulative du modèle sur p. 402](#).
- Classes. Pour plus d'informations, reportez-vous à la section [Vue des classes sur p. 403](#).

Il existe quatre vues liées/auxiliaires :

- Importance des valeurs prédites. Pour plus d'informations, reportez-vous à la section [Vue de l'importance des valeurs prédites de classe sur p. 406](#).
- Taille des classes (par défaut). Pour plus d'informations, reportez-vous à la section [Vue de la taille des classes sur p. 407](#).
- Distribution des cellules. Pour plus d'informations, reportez-vous à la section [Vue de la distribution des cellules sur p. 408](#).
- Comparaison des classes. Pour plus d'informations, reportez-vous à la section [Vue de comparaison des classes sur p. 409](#).

Vue récapitulative du modèle

Figure 11-13
Vue récapitulative du modèle du panneau principal



La vue récapitulative du modèle affiche un instantané, ou un récapitulatif, du modèle de classe, y compris une mesure par silhouette de la cohésion et de la séparation des classes qui est ombrée pour indiquer des résultats faibles, moyens ou bons. Cet instantané vous permet de vérifier rapidement si la qualité est faible, auquel cas vous pouvez décider de revenir au noeud de modélisation afin de corriger les paramètres du modèle de classe pour obtenir un meilleur résultat.

Les résultats faibles, moyens et bons sont basés sur le travail de Kaufman et Rousseeuw (1990) concernant l'interprétation des structures de classe. Dans la vue récapitulative du modèle, d'après l'évaluation de Kaufman et Rousseeuw, un bon résultat équivaut à des données qui indiquent une preuve raisonnable ou forte de la structure de la classe, un résultat moyen signifie une preuve faible et un résultat mauvais reflète une absence de preuve significative.

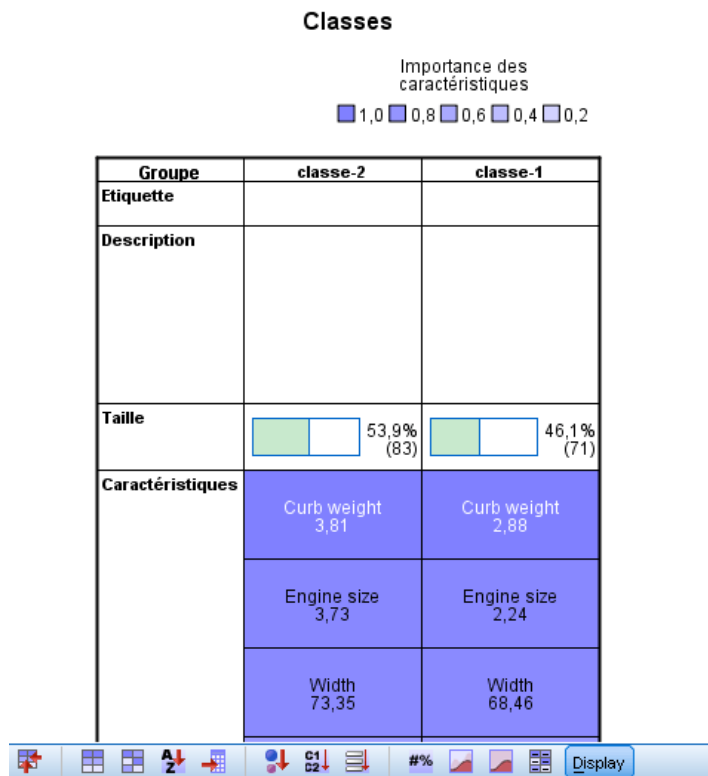
La mesure par silhouette établit la moyenne, de tous les enregistrements, $(B-A) / \max(A,B)$, où A correspond à la distance de l'enregistrement au centre de sa classe et B correspond à la distance de l'enregistrement au centre de la classe la plus proche à laquelle il n'appartient pas. Un coefficient de silhouette de 1 signifie que toutes les observations sont situées directement au centre de leurs classes. Une valeur de -1 signifie que toutes les observations sont situées au centre de classe d'autres classes. Une valeur de 0 signifie, en moyenne, que les observations sont équidistantes du centre de leur propre classe et du centre de l'autre classe la plus proche.

Le récapitulatif inclut un tableau qui contient les informations suivantes :

- **Algorithme.** L'algorithme de classification utilisé, par exemple "TwoStep".
- **Caractéristiques d'entrée.** Le nombre de champs, ou d'entrées ou de valeurs prédites.
- **Classes.** Le nombre de classes dans la solution.

Vue des classes

Figure 11-14
Vue des centres de la classe du panneau principal



La vue des classes contient une grille présentant les classes par caractéristiques, qui inclut les noms des classes, leurs tailles et les profils de chaque classe.

Les colonnes de la grille contiennent les informations suivantes :

- **Classe.** Les nombres de classes créées par l’algorithme.
- **Étiquette.** Toutes les étiquettes appliquées à chaque classe (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir une étiquette qui décrit les contenus de classe ; par exemple “Acheteurs de voitures de luxe”.
- **Description :** Toutes les descriptions du contenu de classe (celle-ci est vierge par défaut). Double-cliquez dans la cellule pour saisir une description de la classe ; par exemple « professionnels, plus de 55 ans, gagnant plus de 100 000 \$ ».
- **Taille :** La taille de chaque classe sous la forme d’un pourcentage de l’échantillon général des classes. Chaque cellule de taille de la grille affiche une barre verticale qui indique le pourcentage de taille au sein de la classe, un pourcentage de taille au format numérique et les effectifs d’observation de la classe.
- **Caractéristiques.** Les entrées ou valeurs prédites individuelles, triées par importance générale par défaut. Si des colonnes ont la même taille, elles sont affichées par ordre croissant des numéros de classe.

L'importance des caractéristiques générales est indiquée par la couleur d'ombrage de l'arrière-plan de la cellule ; la caractéristique la plus importante étant plus sombre et la moins importante n'étant pas ombrée. Un guide situé au-dessus du tableau indique l'importance correspondant à chaque couleur de cellule de caractéristique.

Lorsque vous passez la souris sur une cellule, le nom complet/l'étiquette de la caractéristique et la valeur de l'importance de la cellule s'affichent. De plus amples informations peuvent être affichées selon le type de vue et de caractéristique. Dans la vue Centres des classes, cela inclut les statistiques de la cellule et la valeur de la cellule ; par exemple : "Moyenne : 4.32". Pour les caractéristiques qualitatives, la cellule affiche le nom de la catégorie la plus fréquente (modale) et son pourcentage.

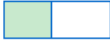


Dans la vue des classes, vous pouvez sélectionner plusieurs manières d'afficher les informations des classes :

- Transposer les classes et les caractéristiques. [Pour plus d'informations, reportez-vous à la section Transposer les classes et les caractéristiques sur p. 404.](#)
- Trier les caractéristiques. [Pour plus d'informations, reportez-vous à la section Trier les caractéristiques sur p. 405.](#)
- Trier les classes. [Pour plus d'informations, reportez-vous à la section Trier les classes sur p. 405.](#)
- Sélectionner le contenu des cellules. [Pour plus d'informations, reportez-vous à la section Contenu des cellules sur p. 405.](#)

Transposer les classes et les caractéristiques

Par défaut, les classes sont affichées en colonnes et les caractéristiques en lignes. Pour inverser cet affichage, cliquez sur le bouton Transposer les classes et les caractéristiques à gauche des boutons Trier les caractéristiques par. Par exemple, vous pouvez réaliser ceci lorsque de nombreuses classes sont affichées, afin de réduire le défilement horizontal nécessaire pour visualiser les données.

Figure 11-15
Classes transposées dans le panneau principal

Classe	Etiquette	Description	Taille	
cluster-1			 45,0% (91)	BP HIGH (41,8%)
cluster-3			 35,0% (70)	BP NORMAL (51,4%)
cluster-2			 19,0% (39)	BP HIGH (100,0%)

Trier les caractéristiques

Le bouton Trier les caractéristiques par vous permet de sélectionner la façon dont les cellules de caractéristiques sont affichées :

- **Importance générale.** Il s'agit de l'ordre de tri par défaut. Les caractéristiques sont triées en ordre décroissant de l'importance générale, et l'ordre de tri est le même pour toutes les classes. Si des caractéristiques ont des valeurs d'importance liées, les caractéristiques liées sont répertoriées par ordre croissant des noms de caractéristique.
- **Importance intra-classe.** Les caractéristiques sont triées par rapport à leur importance pour chaque classe. Si des caractéristiques ont des valeurs d'importance liées, les caractéristiques liées sont répertoriées par ordre croissant des noms de caractéristique. Lorsque cette option est sélectionnée, l'ordre de tri varie généralement d'une classe à l'autre.
- **Nom.** Les caractéristiques sont triées par nom dans l'ordre alphabétique.
- **Ordre des données.** Les caractéristiques sont triées selon leur ordre dans l'ensemble de données.

Trier les classes

Par défaut, les classes sont triées en ordre de taille décroissante. Les boutons Trier les classes par vous permettent de les trier par nom dans l'ordre alphabétique, ou, si vous avez créé des étiquettes uniques, par ordre alphabétique des étiquettes.

Les caractéristiques ayant la même étiquette sont triées selon le nom de classe. Si des classes sont triées par étiquette et que vous modifiez l'étiquette d'une classe, l'ordre de tri est automatiquement mis à jour.

Contenu des cellules

Les boutons Cellules vous permettent de modifier l'affichage du contenu des cellules pour les champs de caractéristiques et d'évaluation.

- **Centre de classes.** Par défaut, les cellules affichent les noms/étiquettes des caractéristiques et la tendance centrale pour chaque combinaison de classe/caractéristique. La moyenne est affichée pour des champs continus et le mode (de la catégorie se présentant le plus fréquemment) avec le pourcentage de catégorie des champs qualitatifs.
- **Distributions absolues.** Affiche des noms/étiquettes de caractéristiques et des distributions absolues des caractéristiques dans chaque classe. Pour les caractéristiques qualitatives, l'affichage montre des diagrammes en bâtons où sont superposées des catégories en ordre croissant de la valeur des données. Pour les caractéristiques continues, l'affichage montre un diagramme de densité lissé qui utilise les mêmes extrema et intervalles pour chaque classe. L'affichage en rouge uni montre la distribution des classes, alors qu'un affichage plus pâle représente les données générales.
- **Distributions relatives.** Affiche les noms/étiquette des caractéristiques et les distributions relatives dans les cellules. En général, les affichages sont similaires à ceux des distributions absolues, excepté les distributions relatives qui sont affichées à la place.

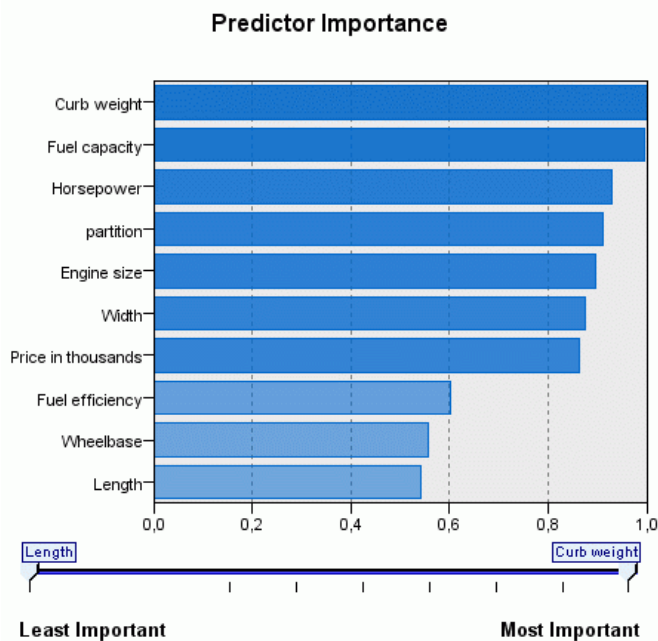
L'affichage en rouge uni montre la distribution des classes, alors qu'un affichage plus pâle représente les données générales.

- **Vue de base.** Là où il y a beaucoup de classes, il peut être difficile de distinguer tous les détails sans procéder à un défilement. Afin de réduire le défilement, sélectionnez cette vue pour modifier l'affichage en une version plus compacte du tableau.

Vue de l'importance des valeurs prédites de classe

Figure 11-16

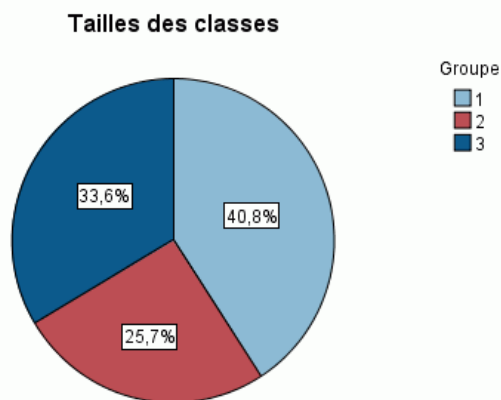
Vue de l'importance des valeurs prédites de classe dans le panneau des liens



La vue de l'importance des valeurs prédites affiche l'importance relative de chaque champ dans l'estimation du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Vue de la taille des classes

Figure 11-17
 Vue de la taille des classes dans le panneau des liens



Taille de la classe la plus petite	39 (25,7%)
Taille de la classe la plus grande	62 (40,8%)
Rapport des tailles : Plus grande classe à plus petite classe	1,59

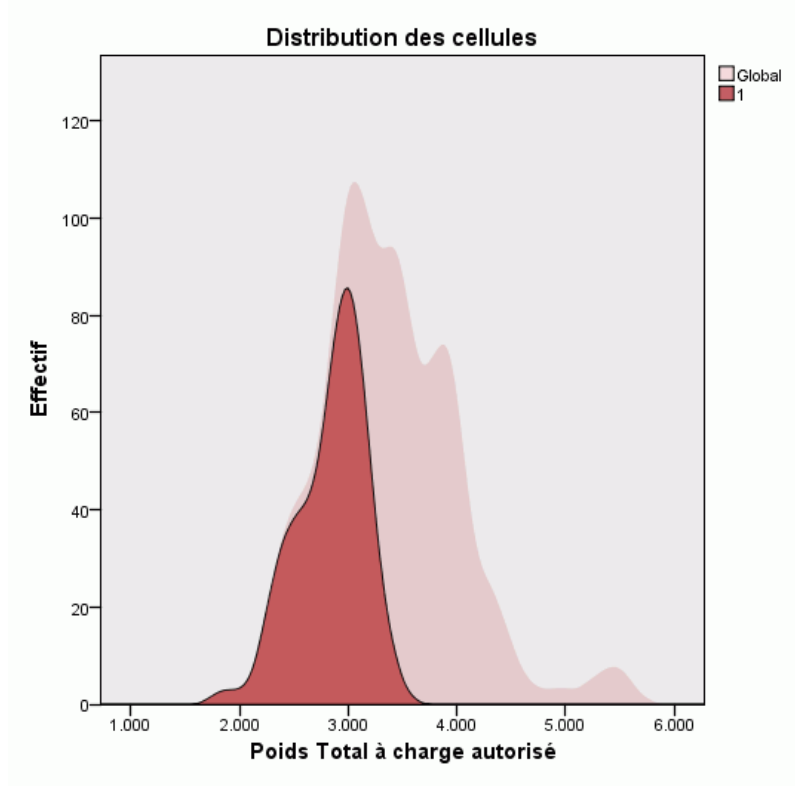
La vue de la taille des classes affiche un diagramme en secteurs qui contient chaque classe. La taille en pourcentage de chaque classe est affichée sur chaque tranche ; passez la souris sur chaque tranche pour afficher l'effectif de celle-ci.

En dessous du diagramme, un tableau répertorie les informations suivantes sur la taille :

- La taille de la classe la plus petite (en effectif et en pourcentage de l'ensemble).
- La taille de la classe la plus grande (en effectif et en pourcentage de l'ensemble).
- Le rapport de taille de la classe la plus grande par rapport à la classe la plus petite.

Vue de la distribution des cellules

Figure 11-18

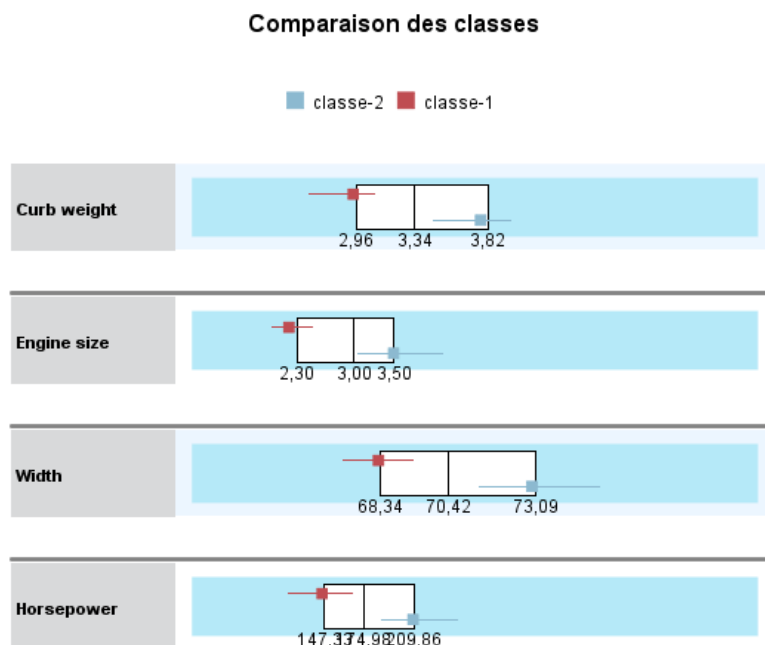
Vue de la distribution des cellules dans le panneau des liens

La vue de la distribution des cellules affiche un diagramme étendu et plus détaillé de la distribution des données pour toutes les cellules de caractéristique que vous sélectionnez dans le tableau du panneau principal des classes.

Vue de comparaison des classes

Figure 11-19

Vue de comparaison des classes dans le panneau des liens



La vue de comparaison des classes se compose d'une présentation sous forme de grille avec des caractéristiques dans les lignes et des classes sélectionnées dans les colonnes. Cette vue vous aide à mieux comprendre les facteurs qui composent les classes ; elle vous permet aussi de voir les différences entre les classes non seulement par comparaison avec les données générales mais aussi en comparant les classes les unes aux autres.

Pour sélectionner des classes à afficher, cliquez en haut de la colonne de classe sur le panneau principal des classes. Cliquez en maintenant la touche Ctrl ou Maj enfoncée pour sélectionner ou désélectionner plus d'une classe pour la comparaison.

Remarque : Vous pouvez sélectionner jusqu'à cinq classes à afficher.

Les classes sont affichées dans l'ordre où elles ont été sélectionnées, alors que l'ordre des champs est déterminé par l'option Trier les caractéristiques par. Lorsque vous sélectionnez Importance intra-classe, les champs sont toujours triés par ordre d'importance générale.

Les diagrammes d'arrière-plan affichent les distributions générales de chaque caractéristique :

- Les caractéristiques qualitatives sont affichées sous forme de tracés de points, où la taille des points indique la catégorie la plus fréquente/modale pour chaque classe (par caractéristique).
- Les caractéristiques continues seront affichées sous forme de boîtes à moustache, qui affichent les médianes générales et les intervalles interquartiles.

Des boîtes à moustache des classes sélectionnées recouvrent ces vues d'arrière-plan :

- Pour des fonctions continues, des marqueurs en points carrés et des lignes horizontales indiquent la médiane et l'intervalle interquartile de chaque classe.
- Chaque classe est représentée par une couleur différente, affichée en haut de la vue.

Navigation dans le visualiseur de classes

Le visualiseur de classes est un affichage interactif. Vous pouvez :

- sélectionner un champ ou une classe pour afficher davantage de détails ;
- comparer des classes afin de sélectionner des éléments intéressants ;
- modifier l'affichage ;
- transposer les axes.
- Générer des noeuds Calculer, Filtre et Sélectionner à l'aide du menu Générer.

Utilisation des barres d'outils

Vous contrôlez les informations affichées dans les panneaux de gauche et de droite à l'aide des options des barres d'outils. Vous pouvez modifier l'orientation de l'affichage (haut-bas, gauche-droite ou droite-gauche) à l'aide des commandes des barres d'outils. En outre, vous pouvez aussi réinitialiser le visualiseur à ses paramètres par défaut et ouvrir une boîte de dialogue pour spécifier le contenu de la vue des classes dans le panneau principal.

Figure 11-20

Barres d'outils pour contrôler les données affichées dans le visualiseur de classes



Les options Trier les caractéristiques par, Trier les classes par, Cellules et Affichage ne sont disponibles que lorsque vous sélectionnez la vue Classes du panneau principal. [Pour plus d'informations, reportez-vous à la section Vue des classes sur p. 403.](#)

	Reportez-vous à Transposer les classes et les caractéristiques sur p. 404
	Reportez-vous à Trier les caractéristiques par sur p. 405
	Reportez-vous à Trier les classes par sur p. 405
	Reportez-vous à Cellules sur p. 405

Génération de noeuds à partir de modèles de classe

Le menu Générer vous permet de créer de nouveaux noeuds basés sur le modèle de classe. Cette option est disponible dans l'onglet Modèle du modèle généré et vous permet de générer des noeuds basés sur la sélection ou l'affichage actuel (c'est-à-dire toutes les classes visibles ou sélectionnées). Par exemple, vous pouvez sélectionner une caractéristique unique puis générer un noeud Filtre pour ignorer toutes les autres caractéristiques (invisibles). Les noeuds générés

sont placés sans connexion sur le canevas. En outre, vous pouvez générer une copie du nugget de modèle dans la palette des modèles. N'oubliez pas de connecter les noeuds et de procéder à toutes les modifications souhaitées avant l'exécution.

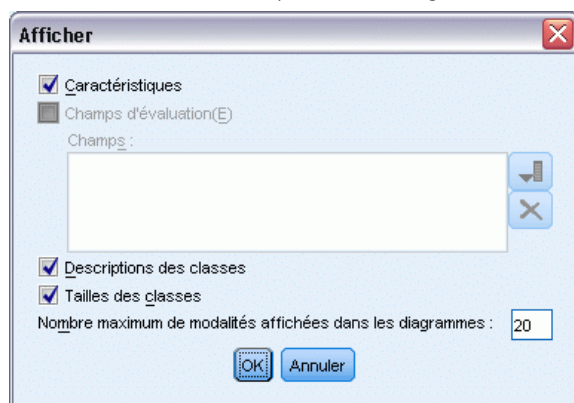
- **Générer un noeud de modélisation.** Crée un noeud de modélisation sur le canevas du flux. Ceci peut être utile, par exemple, si vous avez un flux dans lequel vous souhaitez utiliser ces paramètres de modèle mais que vous ne possédez plus le noeud de modélisation utilisé pour les générer.
- **Modèle vers palette.** Crée un nugget sur la palette des modèles. Ceci est utile dans des situations où un collègue peut vous avoir envoyé un flux contenant le modèle mais pas le modèle même.
- **Noeud Filtre.** Crée un nouveau noeud Filtre pour filtrer les champs qui ne sont pas utilisés par le modèle de classe et/ou qui ne sont pas visibles dans l'affichage actuel du visualiseur de classes. S'il existe un noeud Type en amont de ce noeud Classe, tous les champs avec le rôle *Cible* sont ignorés par le noeud Filtre généré.
- **Noeud Filtre (de la sélection).** Crée un noeud Filtre pour filtrer des champs basés sur des sélections dans le visualiseur de classes. Cliquez tout en maintenant la touche Ctrl enfoncée pour sélectionner plusieurs champs. Les champs sélectionnés dans le visualiseur de classe sont ignorés en aval, mais vous pouvez modifier ce comportement en modifiant le noeud Filtre avant l'exécution.
- **Noeud Sélectionner.** Crée un nouveau noeud Sélectionner pour sélectionner des enregistrements en fonction de leur appartenance à l'une des classes visibles dans l'affichage actuel du visualiseur de classes. Une condition de sélection est automatiquement générée.
- **Noeud Sélectionner (de la sélection).** Crée un nouveau noeud Sélectionner pour sélectionner des enregistrements basés sur l'appartenance à des classes sélectionnées dans le visualiseur de classes. Pour sélectionner plusieurs classes, cliquez tout en maintenant la touche Ctrl enfoncée.
- **Noeud Calculer.** Crée un noeud Calculer qui calcule un champ booléen qui attribue une valeur *True (vrai)* ou *False (faux)* à des enregistrements en fonction de l'appartenance à toutes les classes visibles dans le visualiseur de classes. Une condition de dérivation est automatiquement générée.
- **Noeud Calculer (de la sélection).** Crée un nouveau noeud Calculer qui calcule un champ booléen en fonction de l'appartenance à des classes sélectionnées dans le visualiseur de classes. Pour sélectionner plusieurs classes, cliquez tout en maintenant la touche Ctrl enfoncée.

Outre la génération de noeuds, vous pouvez aussi créer des diagrammes à partir du menu Générer. [Pour plus d'informations, reportez-vous à la section Génération de diagrammes à partir de modèles de classe sur p. 412.](#)

Contrôler l'affichage de la vue des classes

Pour contrôler ce qui est affiché dans la vue des classes sur le panneau principal, cliquez sur le bouton Afficher. La boîte de dialogue Afficher s'ouvre.

Figure 11-21
Visualiseur de classes - option d'affichage



Caractéristiques. Sélectionné par défaut. Pour masquer toutes les caractéristiques entrées, décochez la case.

Champs d'évaluation. Sélectionnez les champs d'évaluation à afficher (les champs qui ne sont pas utilisés pour créer le modèle de classe, mais envoyés au Visualiseur de modèles pour évaluer les classes) ; par défaut, aucun n'est affiché. *Remarque* : Cette case à cocher n'est pas disponible si aucun champ d'évaluation n'est disponible.

Descriptions des classes. Sélectionné par défaut. Pour masquer toutes les cellules de description des classes, décochez la case.

Tailles des classes. Sélectionné par défaut. Pour masquer toutes les cellules de taille des classes, décochez la case.

Nombre maximal de catégories. Spécifiez le nombre maximal de catégorie à afficher dans les diagrammes de caractéristiques qualitatives ; la valeur par défaut est de 20.

Génération de diagrammes à partir de modèles de classe

Les modèles de classes fournissent beaucoup d'informations ; cependant, elles ne sont pas toujours dans un format facilement accessible aux utilisateurs professionnels. Pour fournir les données d'une manière facilement incorporable aux rapports d'activité, aux présentations, etc., vous pouvez réaliser des diagrammes à partir des données sélectionnées. Par exemple, à partir du visualiseur de classes, vous pouvez générer un diagramme pour une classe sélectionnée, en ne créant ainsi un diagramme que pour les observations de cette classe.

Remarque : Vous ne pouvez que générer un diagramme à partir du visualiseur de classes lorsque le nugget de modèle est joint à d'autres noeuds dans un flux.

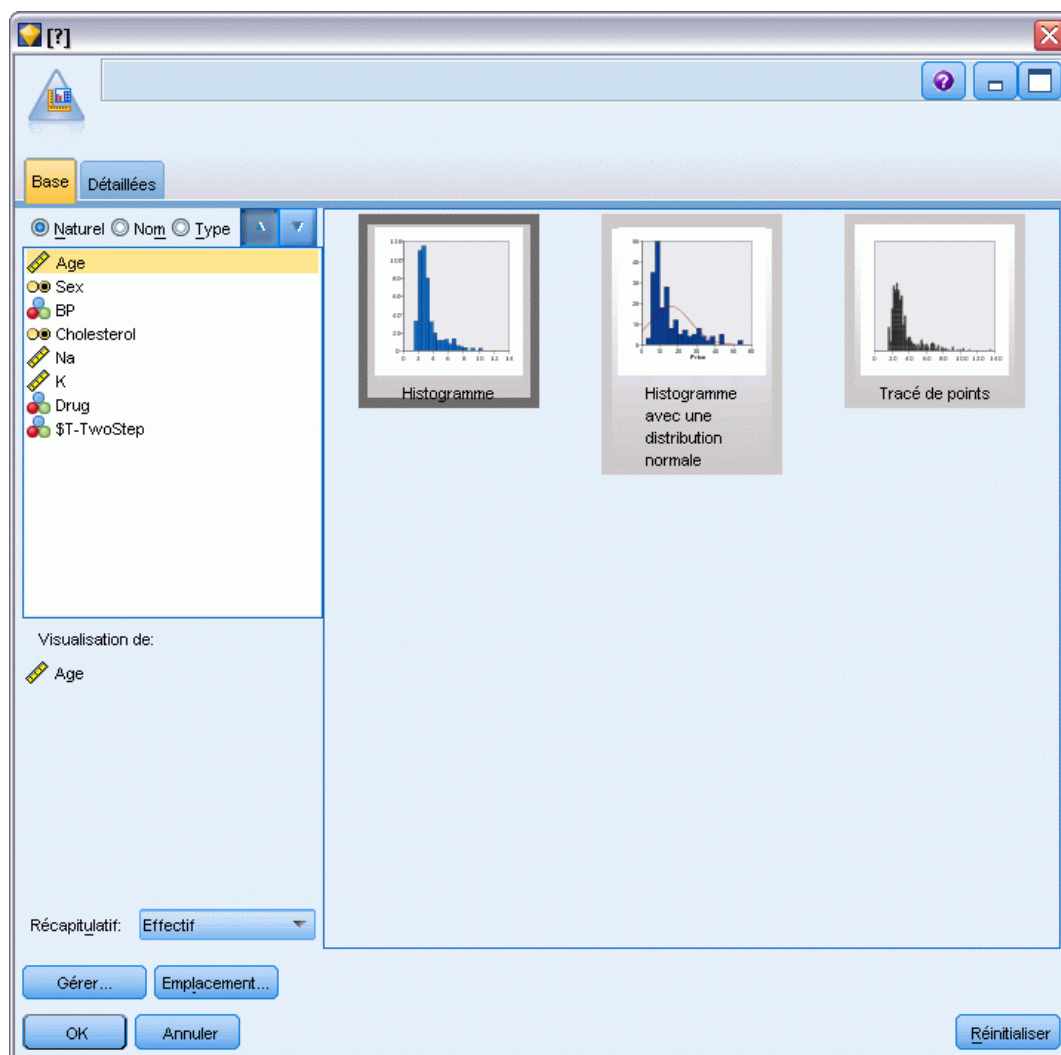
Générer un diagramme

- ▶ Ouvrez le nugget de modèles contenant le visualiseur de classes.
- ▶ Dans l'onglet Modèle, sélectionnez *Classes* à partir de la liste déroulante Affichage.

- ▶ Dans la vue principale, sélectionnez la ou les classes pour lesquelles vous souhaitez réaliser un diagramme.
- ▶ Pour le menu Générer, sélectionnez Graphique (de la sélection) ; l'onglet Graphique de base s'affiche.

Figure 11-22

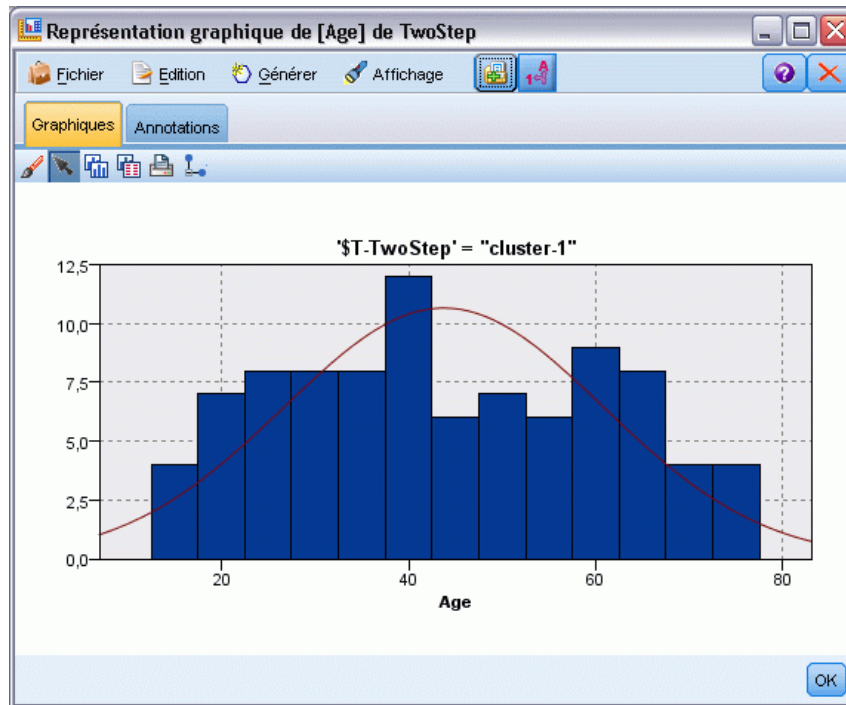
Boîte de dialogue Noeud Représentation graphique, onglet de Base



Remarque : Seuls les onglets de Base et Détaillé seront disponibles lorsque vous affichez le graphique de cette manière. [Pour plus d'informations, reportez-vous à la section Noeud Représentation Graphique dans le chapitre 5 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

- ▶ À l'aide des paramètres de l'onglet de Base ou Détaillé, spécifiez les détails à afficher sur le diagramme.
- ▶ Cliquez sur OK pour générer le diagramme.

Figure 11-23
Histogramme généré à partir de l'onglet de base de la représentation graphique



L'en-tête du diagramme identifie le type de modèle et la ou les classes que vous avez choisies d'inclure.

Règles d'association

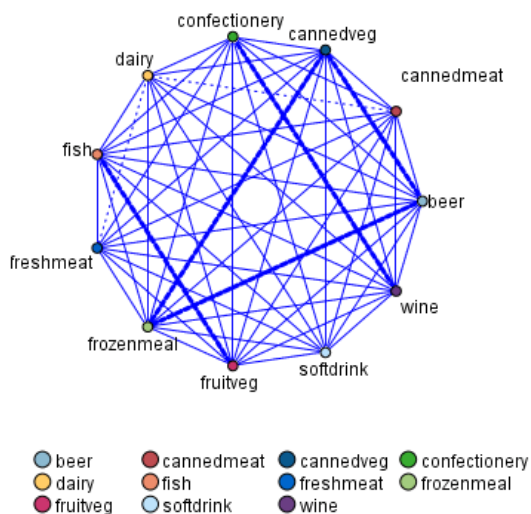
Les **règles d'association** associent une conclusion particulière (l'achat d'un produit particulier, par exemple) à un ensemble de conditions (l'achat de plusieurs autres produits, par exemple). Par exemple, la règle

bière <= conserves légumes & surgelés (173, 17,0 %, 0,84)

sous-entend que l'événement *bière* se produit lorsque les événements *conserves légumes* et *surgelés* ont lieu en même temps. La règle est fiable à 84 % et s'applique à 17 % des données ou à 173 enregistrements. Les algorithmes des règles d'association recherchent automatiquement les associations que vous pouvez trouver manuellement à l'aide de techniques de visualisation, telles que le nœud relations.

Figure 12-1

Noeud Relations présentant les associations entre les éléments du panier du marché



L'avantage des algorithmes de règles d'association par rapport aux algorithmes d'arbre décision standard (C5.0 et Arbres C&RT) est le fait qu'il puisse exister des associations entre *tous* les attributs. Un algorithme d'arbre décision peut construire une règle uniquement avec une seule conclusion. En revanche, les algorithmes d'association tentent d'en trouver plusieurs, chaque règle pouvant avoir une conclusion différente.

L'inconvénient des algorithmes d'association est qu'ils recherchent des éléments dans un espace de recherche très grand, nécessitant ainsi beaucoup plus de temps qu'un algorithme d'arbre décision. Les algorithmes utilisent une méthode de type **générer et tester** pour la recherche de règles. Tout d'abord, des règles simples sont générées, puis elles sont validées par rapport à l'ensemble de données. Les règles satisfaisantes sont conservées et toutes les règles, soumises à différentes contraintes, sont alors spécialisées. **La spécialisation** consiste à ajouter des conditions à une règle. Ces nouvelles règles sont confrontées aux données pour être validées et lors de ce

processus, les meilleures règles, ou les plus intéressantes, sont systématiquement conservées. L'utilisateur apporte généralement certaines limites au nombre possible d'antécédents pour savoir si une règle doit être conservée ou pas. Plusieurs techniques reposant sur la théorie de l'information ou des plans d'indexation efficaces sont utilisées pour réduire significativement l'espace de recherche.

A la fin du traitement, une table contenant les meilleures règles apparaît. A la différence d'un arbre décision, cet ensemble de règles ne peut pas être utilisé directement pour faire des prévisions comme pourrait l'être un modèle standard (tel qu'un arbre décision ou un réseau de neurones). Ceci est dû aux différentes conclusions possibles pour chaque règle. Un autre niveau de transformation est requis pour transformer les règles d'association en un ensemble de règles de classification. C'est pourquoi les règles d'association produites par les algorithmes d'association sont désignées sous le nom de **modèles bruts**. Même si l'utilisateur peut parcourir ces modèles, il ne peut pas les utiliser explicitement en tant que modèles de classification à moins de demander au système de générer un modèle de classe à partir du modèle brut. Cette opération est réalisée à partir du navigateur via l'option de menu Générer.

Deux algorithmes de règles d'association sont pris en charge :



Le noeud Apriori extrait des données un ensemble de règles et retient les règles contenant la plus grande quantité d'informations. Le noeud Apriori fournit cinq méthodes de sélection de règles et utilise un modèle d'indexation sophistiqué pour traiter efficacement les volumes de données importants. Pour les problèmes importants, l'apprentissage du noeud Apriori est généralement plus rapide ; il n'existe aucune limite quant au nombre de règles pouvant être conservées et il peut prendre en charge des règles faisant l'objet de 32 pré-conditions. Le noeud Apriori exige que les champs d'entrée et de sortie soient tous catégoriels, mais fournit de meilleures performances car il est optimisé de ce type de données. [Pour plus d'informations, reportez-vous à la section Noeud Apriori sur p. 417.](#)



Le noeud Séquence recherche des règles d'association dans des données dotées d'une dimension temporelle. Une séquence est une liste de jeux d'éléments ayant tendance à survenir dans un ordre prévisible. Par exemple, un client qui achète un rasoir et une lotion après-rasage achètera vraisemblablement de la crème à raser. Le noeud Séquence est basé sur l'algorithme de règles d'association CARMA, qui utilise une méthode efficace de double lecture pour rechercher des séquences. [Pour plus d'informations, reportez-vous à la section Noeud Séquence sur p. 443.](#)

Données tabulaires et données transactionnelles

Les données utilisées par les modèles de règle d'association peuvent présenter un format transactionnel ou tabulaire, comme décrit ci-dessous. Les informations fournies ici sont des descriptions d'ordre général ; les exigences particulières qui peuvent s'appliquer sont exposées dans la documentation propre à chaque type de modèle. Lors du scoring d'un modèle, les données faisant l'objet du scoring doivent refléter le format de celles utilisées pour la création du modèle. Les modèles créés à partir de données tabulaires ne peuvent servir qu'au scoring des données tabulaires, et ceux créés à partir de données transactionnelles au scoring de ce type de données.

Format transactionnel

Les données transactionnelles comportent un enregistrement distinct pour chaque transaction ou élément. Supposons, par exemple, qu'un client effectue plusieurs achats. Chacun de ces achats constitue un enregistrement distinct, comportant des éléments associés liés par un ID client. C'est ce que l'on nomme parfois format **till-roll** (déroulant).

Client	Achat
1	confiture
2	lait
3	confiture
3	pain
4	confiture
4	pain
4	lait

Les noeuds Apriori, CARMA et Séquence peuvent tous utiliser des données transactionnelles.

Données tabulaires

Les données tabulaires (également appelées données de **panier d'achat** ou données de **table des valeurs vraies**) comportent des éléments représentés par des booléens distincts, chaque champ booléen indiquant la présence ou l'absence d'un élément spécifique. Chaque enregistrement représente un ensemble complet d'éléments associés. Les champs booléens peuvent être catégoriels ou numériques (certains modèles peuvent toutefois présenter des exigences plus particulières).

Client	Confiture	Pain	Lait
1	V	F	F
2	F	F	V
3	V	V	F
4	V	V	V

Les noeuds Apriori, CARMA, et Séquence peuvent tous utiliser des données tabulaires.

Noeud Apriori

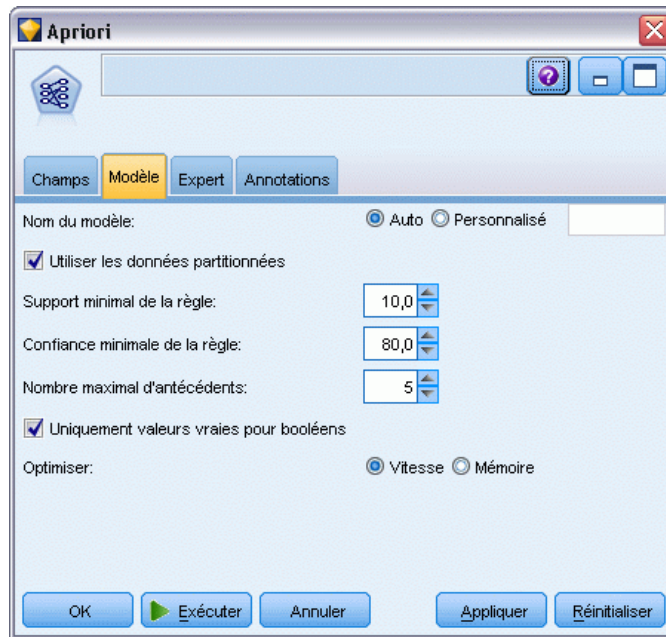
Le noeud Apriori recherche également des règles d'association dans les données. Il fournit cinq méthodes de sélection de règles et utilise un modèle d'indexation évolué pour traiter efficacement les ensembles de données importants.

Conditions requises. La création d'un ensemble de règles Apriori requiert au moins un champ *Entrée* et au moins un champ *Cible*. Les champs d'entrée et de sortie (dont le rôle *Entrée*, *Cible* ou *Les deux*) doivent être symboliques. Les champs indiquant le rôle *Aucun* sont ignorés. Les types de champ doivent être entièrement instanciés avant l'exécution du noeud. Les données peuvent être au format tabulaire ou transactionnel. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles sur p. 416.](#)

Puissance. Pour les problèmes importants, l'apprentissage du noeud Apriori est généralement plus rapide. D'autre part, il n'existe aucune limite quant au nombre de règles pouvant être conservées et le noeud Apriori peut prendre en charge des règles faisant l'objet de 32 pré-conditions. Le noeud Apriori propose cinq méthodes d'apprentissage, ce qui permet une meilleure adaptation de l'approche Data mining aux problèmes rencontrés.

Noeud Apriori - Options du modèle

Figure 12-2
Noeud Apriori - 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.

Prise en charge minimale de la règle. Vous pouvez spécifier un critère de prise en charge pour toutes les règles. L'option **Prise en charge** correspond au pourcentage d'enregistrements contenus dans les données d'apprentissage pour lesquels les antécédents (la partie « if » de la règle) sont vrais (true). (Cette définition de la prise en charge est différente de celle utilisée dans les noeuds CARMA et Séquence. [Pour plus d'informations, reportez-vous à la section Noeud Séquence - Options modèle sur p. 446.](#)) Si vous obtenez des règles qui s'appliquent à des sous-ensembles de données de très petite taille, augmentez la valeur de ce paramètre.

Remarque : La définition de la prise en charge d'Apriori repose sur le nombre d'enregistrements dotés des antécédents. Elle diffère ainsi des algorithmes CARMA et Séquence, dont la définition de prise en charge est basée sur le nombre d'enregistrements dotés de tous les éléments d'une règle (c'est-à-dire les antécédents et les conséquences). Les résultats des modèles d'association affichent à la fois la prise en charge (antécédent) et les mesures de prise en charge de règle.

Confiance minimale de la règle. Vous pouvez également spécifier un critère de confiance. La **confiance** est basée sur les enregistrements pour lesquels les antécédents de la règle sont vrais (true), et correspond au pourcentage de ces enregistrements pour lesquels les conséquences sont également vraies (true). Autrement dit, le pourcentage des prévisions basées sur la règle est correct. Les règles dont le niveau de confiance est inférieur au niveau minimal de confiance spécifié sont ignorées. Augmentez la valeur de cette option si le nombre de règles est trop important. En revanche, réduisez la valeur de ce paramètre si le nombre de règles n'est pas assez important, voire nul.

Nombre maximal d'antécédents. Vous pouvez spécifier le nombre maximal de pré-conditions s'appliquant aux règles. Cela vous permet de limiter leur complexité. Si les règles sont trop complexes ou spécifiques, réduisez la valeur de ce paramètre. Ce paramètre influe beaucoup sur la durée de l'apprentissage. Si l'apprentissage de votre ensemble de règles est trop long, réduisez la valeur de ce paramètre.

Uniquement valeurs vraies pour booléens. Si cette option est sélectionnée pour des données au format tabulaire (table des valeurs vraies), seules les valeurs vraies sont incluses dans les règles générées. Cela permet d'interpréter plus facilement les règles. Cette option ne s'applique pas aux données au format transactionnel. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles sur p. 416.](#)

Optimiser. Sélectionnez, à votre convenance, les options destinées à accroître les performances du système au cours de la création de modèles.

- Sélectionnez Vitesse pour que l'algorithme n'utilise jamais le débordement sur disque, afin d'améliorer les performances.
- Sélectionnez Mémoire pour que l'algorithme utilise le débordement sur disque en cas de besoin, ce qui a pour effet de réduire la vitesse. Par défaut, cette option est sélectionnée.
Remarque : Si vous exécutez le système en mode réparti, ce paramètre peut être ignoré par les options administrateur indiquées dans le fichier *options.cfg*. Pour plus d'informations, consultez le document intitulé *IBM® SPSS® Modeler ServerGuide de l'administrateur*.

Options expert du noeud Apriori

Les options expert décrites ci-dessous, qui s'adressent aux utilisateurs ayant une connaissance approfondie des opérations Apriori, permettent d'affiner le processus d'induction. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 12-3
Options expert Apriori



Mesure évaluation. Le noeud Apriori prend en charge cinq méthodes d'évaluation des règles potentielles.

- **Confiance règle.** La méthode par défaut évalue les règles en fonction de leur degré de confiance (ou précision). Par conséquent, l'option Limite inférieure de la mesure d'évaluation est désactivée, puisqu'elle crée un effet de redondance avec l'option Confiance minimale de la règle de l'onglet Modèles. [Pour plus d'informations, reportez-vous à la section Noeud Apriori - Options du modèle sur p. 418.](#)
- **Différence de confiance.** (Egalement appelée **différence de confiance absolue a priori**.) Cette mesure d'évaluation correspond à la différence absolue entre le degré de confiance actuel de la règle et son degré de confiance antérieur. Cette option empêche l'apparition de biais lorsque les résultats ne sont pas distribués de façon régulière. Cela permet d'éviter de conserver des règles trop « évidentes ». Par exemple, supposons que 80 % de vos clients achètent votre produit le plus apprécié. Une règle qui prédit avec 85 % de précision que votre produit sera acheté ne constitue pas une information essentielle, même si une précision de 85 % semble être un excellent résultat dans l'absolu. Paramétrez la valeur du champ Limite inférieure de la mesure d'évaluation sur la différence minimale de confiance déterminant si les règles doivent être conservées.
- **Ratio de confiance.** (également appelé **différence du quotient de confiance par rapport à 1**.) Cette mesure d'évaluation correspond au rapport entre le degré de confiance actuel et le degré de confiance antérieur de la règle (ou, si le rapport est supérieur à un, sa valeur inverse). Tout comme la différence de confiance, cette méthode prend en compte les proportions irrégulières. Elle vous permet de trouver des règles qui prédisent des événements rares. Prenons pour exemple une maladie extrêmement rare qui concerne seulement 1 % des patients. Une règle capable de prédire cette maladie avec une précision de 10 % constitue un grand progrès par rapport à l'estimation aléatoire, même si dans l'absolu une précision de 10 % ne semble pas

exceptionnelle. Dans le champ Limite inférieure de la mesure d'évaluation, indiquez la différence déterminant si les règles doivent être conservées.

- **Différence d'informations.** (Également appelée **différence d'informations a priori**.) Cette mesure est calculée sur la base de la mesure **gain d'informations**. Si la probabilité d'une conséquence particulière est considérée comme une valeur logique (un **bit**), alors le gain d'informations correspond à la proportion de l'octet qui peut être déterminée par les antécédents. La différence d'informations correspond à la différence entre le gain d'informations calculé sur la base des antécédents et le gain d'informations calculé sur la base du degré de confiance antérieur de la conséquence. L'une des principales caractéristiques de cette méthode est qu'elle prend en compte la prise en charge. Ainsi, pour un niveau de confiance donné, les règles qui seront le plus souvent retenues sont celles qui s'appliquent au plus grand nombre d'enregistrements. Dans le champ Limite inférieure de la mesure d'évaluation, indiquez la différence déterminant si les règles doivent être conservées.

Remarque : Dans la mesure où l'échelle de cette mesure est moins intuitive que les autres échelles, il peut s'avérer nécessaire de tester différentes valeurs de ce paramètre avant d'obtenir un ensemble de règles satisfaisant.

- **Chi-deux normalisé** (également appelé **mesure de Chi-deux normalisée**). Cette mesure est un index statistique d'associations entre les antécédents et les conséquences. Cette mesure est normalisée de façon à prendre des valeurs comprises entre 0 et 1. L'influence de la prise en charge sur cette mesure est encore plus importante que pour la mesure de la différence d'informations. Dans le champ Limite inférieure de la mesure d'évaluation, indiquez la différence déterminant si les règles doivent être conservées.

Remarque : Tout comme pour la différence d'informations, l'échelle de cette mesure est moins intuitive que celles des autres mesures. Il peut donc s'avérer nécessaire de tester différentes valeurs de ce paramètre avant d'obtenir un ensemble de règles satisfaisant.

Autoriser les règles ne comportant pas d'antécédents. Sélectionnez cette option pour autoriser les règles qui comprennent uniquement les conséquences (élément ou jeu d'éléments). Cela s'avère utile lorsque vous souhaitez déterminer des éléments ou des jeux d'éléments communs. Par exemple, *cannedveg* est une règle à élément unique sans antécédent qui indique que l'achat de *conserves légumes* est une occurrence commune aux données. Vous pouvez, dans certains cas, inclure des règles de ce type si vous ne souhaitez obtenir que les prévisions les plus sûres. Cette option est désactivée par défaut. Par convention, la prise en charge des antécédents des règles sans antécédent est de 100 %. Quant à la prise en charge de la règle, elle est identique au degré de confiance.

Noeud CARMA

Le noeud CARMA utilise un algorithme de découverte de règles d'association pour déterminer les règles d'association des données. Les règles d'association sont des instructions ayant la forme suivante

if antecedent(s) then consequent(s)

Par exemple, si un internaute achète une carte sans fil et un routeur sans fil haut de gamme, il est également probable qu'il achète un serveur de musique sans fil, si l'offre est disponible. Le modèle CARMA extrait un ensemble de règles des données sans que vous ayez à définir les

champs d'entrée ou les champs cible. Les règles générées peuvent ainsi être utilisées pour une plus grande variété d'applications. Par exemple, vous pouvez utiliser les règles générées par ce noeud pour rechercher une liste de produits ou de services (antécédents) dont la conséquence correspond à l'élément que vous souhaitez promouvoir à l'occasion de cette période de congés. Grâce à IBM® SPSS® Modeler, vous pouvez identifier les clients qui ont acheté les produits antécédents et élaborer une campagne marketing destinée à promouvoir le produit résultant.

Conditions requises. A l'inverse d'Apriori, le noeud CARMA ne requiert pas de champs *Entrée* ou *Cible*. Cette caractéristique est inhérente au mode de fonctionnement de l'algorithme ; cela revient à créer un modèle Apriori dont tous les champs sont paramétrés sur *Les deux*. Vous pouvez limiter les éléments qui sont uniquement répertoriés comme antécédents ou comme conséquences ; il vous suffit pour cela de filtrer le modèle une fois qu'il a été créé. Par exemple, vous pouvez utiliser le navigateur de modèle pour rechercher une liste de produits ou de services (antécédents) dont la conséquence correspond à l'élément que vous souhaitez promouvoir à l'occasion de cette période de congés.

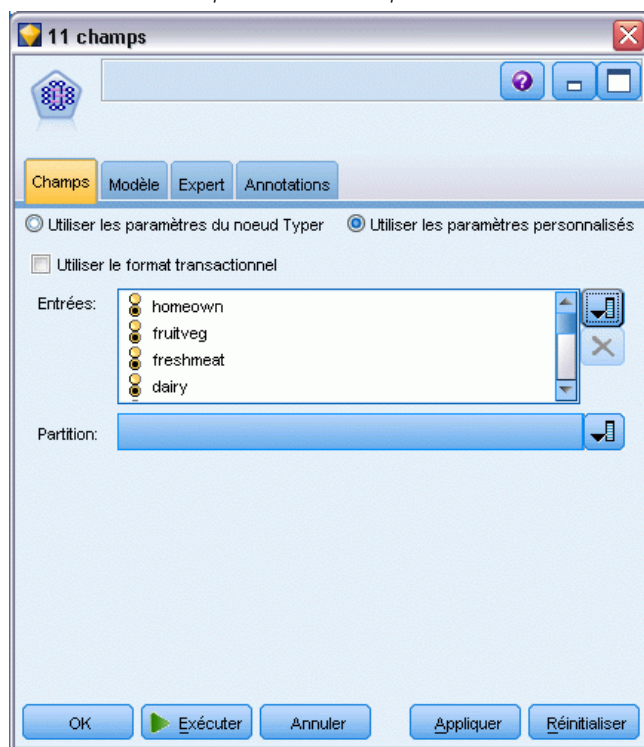
Pour créer un ensemble de règles CARMA, vous devez renseigner un champ ID, ainsi qu'un ou plusieurs champs d'analyse. Le champ ID peut être de n'importe quel rôle ou niveau de mesure. Les champs indiquant le rôle *Aucun* sont ignorés. Les types de champ doivent être entièrement instanciés avant l'exécution du noeud. De même que pour Apriori, les données peuvent être au format tabulaire ou transactionnel. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles sur p. 416.](#)

Puissance. Le noeud CARMA est basé sur l'algorithme de règles d'association CARMA. Contrairement aux noeuds Apriori, le noeud CARMA offre des paramètres de création pour la prise en charge de la règle (à la fois pour les antécédents et les conséquences), et non une prise en charge d'antécédents. CARMA autorise également les règles comportant plusieurs conséquences. A l'instar d'Apriori, les modèles générés par un noeud CARMA peuvent être insérés dans un flux de données afin de créer des prévisions. [Pour plus d'informations, reportez-vous à la section Nuggets de modèle dans le chapitre 3 sur p. 48.](#)

Noeud CARMA - Options des champs

Avant d'exécuter un noeud CARMA, vous devez renseigner les champs d'entrée dans l'onglet Champs du noeud CARMA. Alors que la plupart des noeuds de modélisation partagent les mêmes options dans l'onglet Champs, le noeud CARMA contient plusieurs options exclusives. Toutes ces options sont décrites ci-dessous.

Figure 12-4
Noeud CARMA - Options des champs



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. Une fois cette option sélectionnée, choisissez les champs ci-dessous en fonction du format (tabulaire ou transactionnel) de lecture des données.

Utiliser le format transactionnel. Cette option permet de modifier les commandes de champ situées ailleurs dans cette boîte de dialogue, en fonction du format des données (tabulaire ou transactionnel). Si vous utilisez plusieurs champs avec des données transactionnelles, le système suppose que les éléments spécifiés dans ces champs pour un enregistrement spécifique représentent les éléments trouvés dans une même transaction avec un seul horodatage. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles sur p. 416.](#)

Données tabulaires

Si Utiliser le format transactionnel n'est pas sélectionné, les champs suivants apparaissent.

- **Entrées.** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un nœud *Typé*.
- **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 nœuds *Typé* 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 nœud. (Désélectionnez cette option pour pouvoir désactiver la partition sans modifier les paramètres du champ.)

Données transactionnelles

Si vous sélectionnez *Utiliser le format transactionnel*, les champs suivants apparaissent.

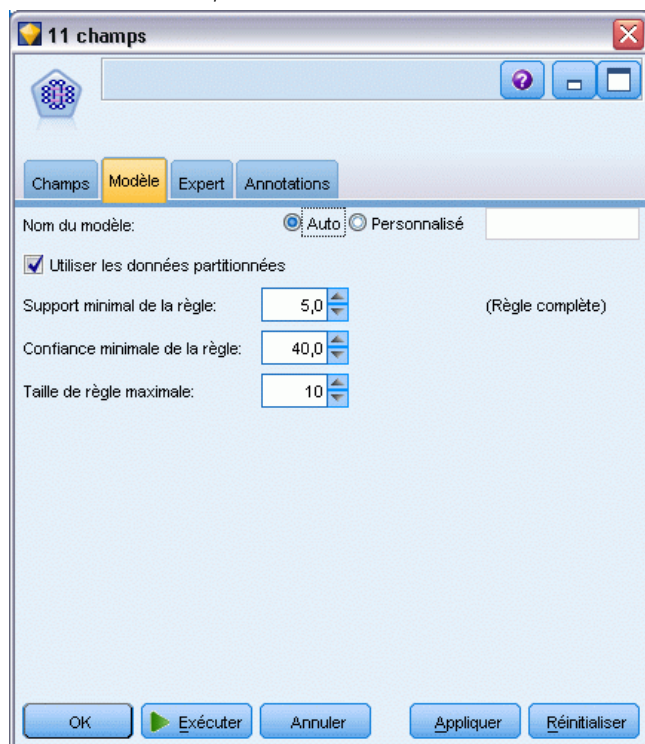
- **ID.** Pour des données transactionnelles, sélectionnez un champ ID dans la liste. Vous pouvez utiliser un champ numérique ou symbolique en tant que champ ID. Chaque valeur unique de ce champ doit indiquer une unité d'analyse spécifique. Par exemple, dans une analyse de paniers du marché, chaque ID 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).
- **Les ID sont contigus.** (Pour les nœuds *Apriori* et *CARMA* seulement) Si vos données sont pré-triées de façon à ce que tous les enregistrements avec le même ID soient regroupés dans le flux de données, sélectionnez cette option pour accélérer le traitement. Si vos données ne sont pas pré-triées (ou si vous n'en êtes pas certain), ne sélectionnez pas cette option ; le nœud triera automatiquement les données.

Remarque : Si vos données ne sont pas triées et si vous sélectionnez cette option, vous risquez d'obtenir des résultats incorrects dans votre modèle.

- **Contenu.** Permet d'ajouter des informations aux champs d'analyse du modèle. Ces champs contiennent les éléments d'intérêt concernant la modélisation des associations. Vous pouvez définir plusieurs champs booléens (si les données sont au format tabulaire) ou un champ nominal unique (si les données sont au format transactionnel).

Noeud CARMA - Options du modèle

Figure 12-5
Noeud CARMA - 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.

Prise en charge minimale de la règle (%). Vous pouvez spécifier un critère de prise en charge. L'option de **prise en charge de la règle** correspond à la proportion d'ID dans les données d'apprentissage qui contiennent la règle, (Cette définition de la prise en charge est différente de la prise en charge des antécédents, utilisée dans les noeuds Apriori.) Si vous souhaitez extraire des règles plus standard, augmentez la valeur de ce paramètre.

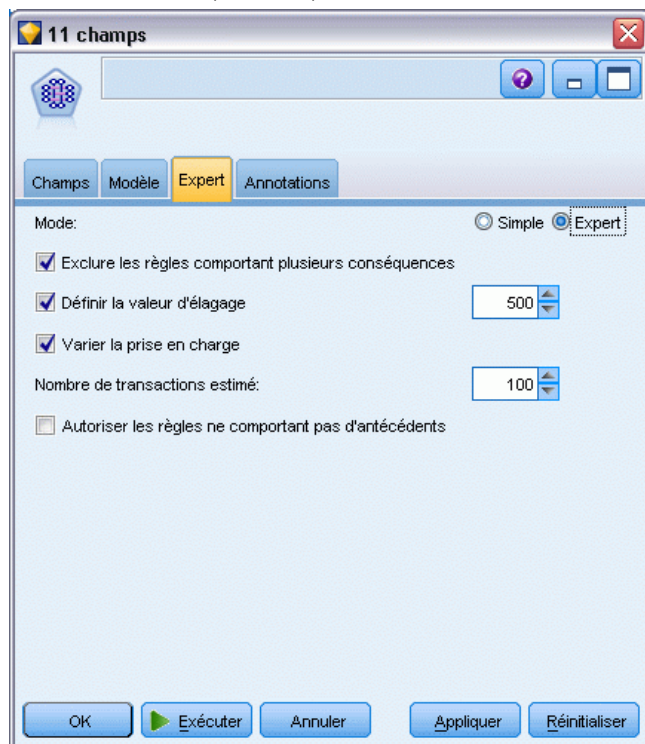
Confiance minimale de la règle (%). Vous pouvez spécifier un critère de confiance pour toutes les règles. La **confiance** correspond au pourcentage des ID pour lesquels une prévision correcte est effectuée, par rapport à l'ensemble des ID pour lesquels la règle effectue une prévision. Elle est calculée comme suit : nombre d'ID pour lesquels la règle entière est trouvée, divisé par le nombre d'ID pour lesquels les antécédents sont trouvés, sur la base des données d'apprentissage. Les règles dont le niveau de confiance est inférieur au niveau minimal de confiance spécifié sont ignorées. Augmentez la valeur de cette option si le nombre de règles est trop important ou si les règles obtenues sont sans intérêt. En revanche, réduisez la valeur si le nombre de règles n'est pas assez important.

Taille de règle maximale. Vous pouvez définir le nombre maximal de *jeux d'éléments* (et non d'*éléments*) que doit contenir une règle. Si les règles qui vous intéressent sont relativement courtes, diminuez la valeur de ce paramètre afin d'accélérer la génération de l'ensemble de règles.

Noeud CARMA - Options expert

Les options expert suivantes, qui s'adressent aux utilisateurs ayant une bonne connaissance du fonctionnement du noeud CARMA, permettent d'affiner le processus de création de modèle. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 12-6
Noeud CARMA - Options expert



Exclure les règles comportant plusieurs conséquences. Sélectionnez cette option pour exclure les conséquences « doubles », c'est-à-dire les conséquences qui contiennent deux éléments. Par exemple, la règle bread & cheese & fish -> wine&fruit contient une conséquence double, wine&fruit. Par défaut, ces règles sont incluses.

Définir une valeur d'élagage. L'algorithme CARMA supprime (**élague**) périodiquement de sa liste de jeux d'éléments potentiels les jeux peu fréquents afin d'économiser de la mémoire. Sélectionnez cette option pour régler la fréquence d'élagage ; le nombre que vous indiquez ici détermine la fréquence d'élagage. Entrez une valeur faible pour diminuer les besoins en mémoire de l'algorithme (le temps d'apprentissage requis risque d'être augmenté). Entrez une valeur élevée pour accélérer l'apprentissage (les besoins en mémoire de l'algorithme risquent d'être augmentés). La valeur par défaut est 500.

Varier la prise en charge. Sélectionnez cette option pour accroître l'efficacité grâce à l'exclusion des jeux d'éléments peu fréquents, qui semblent fréquents lorsqu'ils sont inclus de manière irrégulière. Pour cette opération, choisissez d'abord un niveau de prise en charge plus élevé, puis diminuez-le progressivement jusqu'à atteindre le niveau indiqué dans l'onglet Modèle. Entrez la

valeur du champ Nombre de transactions estimé afin d'indiquer la rapidité à laquelle le niveau de prise en charge doit être diminué.

Autoriser les règles ne comportant pas d'antécédents. Sélectionnez cette option pour autoriser les règles qui comprennent uniquement les conséquences (élément ou jeu d'éléments). Cela s'avère utile lorsque vous souhaitez déterminer des éléments ou des jeux d'éléments communs. Par exemple, *cannedveg* est une règle à élément unique sans antécédent qui indique que l'achat de *conserves légumes* est une occurrence commune aux données. Vous pouvez, dans certains cas, inclure des règles de ce type si vous ne souhaitez obtenir que les prévisions les plus sûres. Cette option est désactivée par défaut.

Nuggets du modèle de règle d'association

Les nuggets du modèle de règles d'association représentent les règles découvertes par l'un des noeuds de modélisation de règles d'association suivants :

- A priori
- CARMA

Les nuggets de modèles contiennent des informations sur les règles extraites des données lors de la génération de modèles.

Affichage des résultats

Vous pouvez parcourir les règles générées par les modèles d'association (Apriori et CARMA) et les modèles Séquence via l'onglet Modèle de la boîte de dialogue. Parcourir un nugget de modèle permet de prendre connaissance d'informations sur les règles, et fournit des options de filtrage et de tri des résultats avant de générer de nouveaux noeuds ou de déterminer le score du modèle.

Détermination du score du modèle

Vous pouvez ajouter des nuggets de modèles affinés (Apriori, CARMA et Séquence) à un flux et les utiliser pour déterminer des scores. [Pour plus d'informations, reportez-vous à la section Utilisation de nuggets de modèle dans les flux dans le chapitre 3 sur p. 70.](#) Les nuggets de modèles utilisés pour le scoring comportent un onglet supplémentaire, Paramètres, dans leur boîte de dialogue respective. [Pour plus d'informations, reportez-vous à la section Paramètres des nuggets de modèles de règle d'association sur p. 434.](#)

Un nugget de modèle non affiné ne peut pas être utilisé pour le scoring. En revanche, vous pouvez générer un ensemble de règles et l'utiliser pour déterminer des scores. [Pour plus d'informations, reportez-vous à la section Génération d'un ensemble de règles à partir d'un nugget de modèle d'association sur p. 437.](#)

Détails du nugget de modèle de règle d'association

Dans l'onglet Modèle d'un nugget de modèle de règle d'association, vous pouvez voir un tableau contenant les règles extraites par l'algorithme. Chaque ligne du tableau représente une règle. La première colonne représente les conséquences (la partie « then » de la règle) tandis que la

colonne suivante représente les antécédents (la partie « if » de la règle). Les colonnes suivantes contiennent des informations sur la règle, telle que la confiance, la prise en charge et Lift.

Figure 12-7

Onglet des modèles des nuggets de modèles de règle d'association

Conséquence	Antécédent	% de prise en charge	% de confiance
frozenmeal	beer cannedveg	16,7	87,425
cannedveg	beer frozenmeal	17,0	85,882
beer	frozenmeal cannedveg	17,3	84,393
frozenmeal	beer	29,3	58,02
cannedveg	frozenmeal	30,2	57,285
frozenmeal	cannedveg	30,3	57,096
cannedveg	beer	29,3	56,997
beer	frozenmeal	30,2	56,291
beer	cannedveg	30,3	55,116
wine	confectionery	27,6	52,174
confectionery	wine	28,7	50,174

Les règles d'association apparaissent souvent au format suivant :

Conséquence	Antécédent
Drug = drugY	Sex = F BP = HIGH

L'exemple de règle est interprété comme suit : *si Sex = "F" et BP = "HIGH", alors Médicament sera probablement médicamentY*. En d'autres termes, *dans les enregistrements où Sex = "F" et BP = "HIGH", Médicament est certainement médicamentY*. A l'aide de la barre d'outils de la boîte de dialogue, vous pouvez choisir d'afficher d'autres informations, telles que la confiance, la prise en charge et les instances.

Menu Trier. Le bouton de menu Trier de la barre d'outils contrôle le tri des règles. Pour modifier l'ordre de tri (ascendant ou descendant), utilisez le bouton d'ordre du tri (flèche vers le haut ou vers le bas).

Figure 12-8

Options de tri de la barre d'outils

Trier par : % de confiance

Vous pouvez trier les règles par :

- Prise en charge
- Confiance
- Prise en charge de la règle
- Conséquence
- Lift
- Deployabilité

Menu Afficher/Masquer. Le menu Afficher/Masquer (bouton de la barre d'outils indiquant les critères) contrôle les options d'affichage des règles.

Figure 12-9
Bouton Afficher/Masquer



Les options d'affichage suivantes sont disponibles :

- **ID de règle** affiche l'ID de règle attribué lors de la création du modèle. Un ID de règle permet d'identifier les règles actuellement appliquées à une prévision donnée. Les ID de règle permettent également de fusionner des informations de règle supplémentaires ultérieurement, telles que la capacité de déploiement, les informations sur le produit ou les antécédents.
- L'option **Instances** affiche des informations sur le nombre d'ID uniques auxquels s'applique la règle, c'est-à-dire les enregistrements dont les antécédents sont vrais (true). Par exemple, pour la règle `bread -> cheese`, le nombre d'enregistrements incluant l'antécédent `pain` dans les données d'apprentissage est appelé **instances**.
- L'option **Prise en charge** affiche la prise en charge des antécédents—, c'est-à-dire la proportion des ID pour laquelle les antécédents sont true (vrais), sur la base des données d'apprentissage. Par exemple, si 50 % des données d'apprentissage comprennent l'achat de pain, la règle `bread -> cheese` offre une prise en charge de l'antécédent de 50 %. *Remarque :* La prise en charge définie ici est identique aux instances, mais indiquée sous forme de pourcentage.
- **Confiance** affiche le rapport de la prise en charge de la règle sur la prise en charge de l'antécédent. Ce rapport indique la proportion d'ID comportant les antécédents indiqués pour lesquels les conséquences sont également vraies (true). Par exemple, si 50 % des données d'apprentissage contiennent du pain (indiquant la prise en charge de l'antécédent), mais que seulement 20 % contiennent du pain et du fromage (indiquant la prise en charge de la règle), la confiance de la règle `bread -> cheese` est $\text{Rule Support} / \text{Antecedent Support}$ ou, dans ce cas, 40 %.
- L'option **Prise en charge de la règle** affiche la proportion d'ID pour lesquels l'intégralité de la règle, les antécédents et les conséquences sont vrais (true). Par exemple, si 20 % des données d'apprentissage comprennent pain et fromage, le support de la règle `bread -> cheese` est de 20 %.
- **Lift** affiche le rapport de confiance de la règle sur la probabilité a priori d'obtenir la conséquence. Par exemple, si 10 % de la population totale achètent du pain, une règle prévoyant l'achat de pain par la population avec une confiance de 20 % comporte une augmentation de $20/10 = 2$. Si une autre règle indique que la population achète du pain

avec une confiance de 11 %, l'augmentation de cette règle est de près de 1, ce qui signifie que la présence d'antécédents n'affecte pas réellement la probabilité que la conséquence soit obtenue. En général, les règles dont l'augmentation n'est pas 1 sont plus intéressantes que celles dont l'augmentation est proche de 1.

- **Capacité de déploiement** est la mesure du pourcentage des données d'apprentissage remplissant les conditions de l'antécédent, mais pas de la conséquence. En ce qui concerne l'achat du produit, il s'agit du pourcentage de la base de clientèle totale possédant (ou ayant acheté) les antécédents, mais pas encore la conséquence. La statistique de capacité de déploiement est définie comme $((\text{Antecedent Support in \# of Records} - \text{Rule Support in \# of Records}) / \text{Number of Records}) * 100$, où *Prise en charge de l'antécédent* signifie le nombre d'enregistrements pour lesquels les antécédents sont vrais et *Support de règle* signifie le nombre d'enregistrements pour lesquels les antécédents et les conséquences sont vrais.

Bouton Filtrer. Le bouton Filtrer (icône d'entonnoir) du menu développe la partie inférieure de la boîte de dialogue, comportant un panneau où figurent les filtres de règle actifs. Les filtres servent à réduire le nombre de règles affichées dans l'onglet Modèles.

Figure 12-10
Bouton Filtrer



Pour créer un filtre, cliquez sur l'icône de filtre située à droite du panneau développé. La boîte de dialogue qui apparaît alors permet d'indiquer les contraintes d'affichage des règles. Notez que le bouton Filtrer est fréquemment utilisé avec le menu Générer pour filtrer les règles, puis créer un modèle comportant ce sous-ensemble de règles. Pour plus d'informations, reportez-vous à [Définition de filtres pour les règles](#) ci-dessous.

Bouton Rechercher la règle. Le bouton Rechercher la règle (icône des jumelles) permet de rechercher l'ID de règle indiqué dans les règles affichées. La boîte d'affichage adjacente indique le nombre de règles actuellement affichées sur le nombre disponible. Les ID de règle sont attribués par le modèle dans l'ordre dans lequel ils sont découverts et sont ajoutés aux données lors de la détermination des scores.

Figure 12-11
Bouton Rechercher la règle



Pour réorganiser les ID de règle :

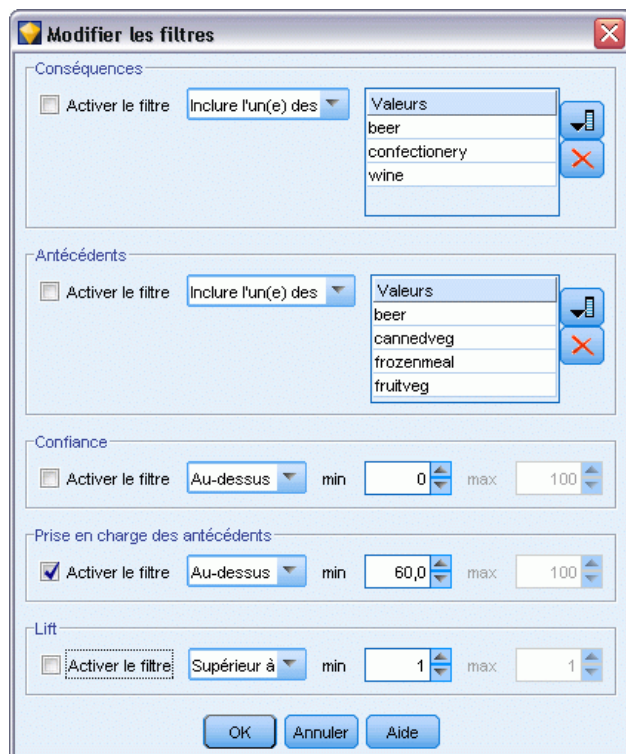
- ▶ Dans IBM® SPSS® Modeler, vous pouvez organiser les ID de règle en triant d'abord le tableau d'affichage des règles selon la mesure voulue, telle que la confiance ou Lift.
- ▶ Puis, à l'aide des options du menu Générer, créez un modèle filtré.
- ▶ Dans la boîte de dialogue Modèle filtré, sélectionnez Renuméroter les règles de manière consécutive en commençant par et indiquez un numéro de début.

Pour plus d'informations, reportez-vous à la section [Génération d'un modèle filtré](#) sur p. 438.

Définition de filtres pour les règles

Par défaut, les algorithmes de règle, tels qu'Apriori, CARMA et Séquence peuvent générer un grand nombre de règles, généralement encombrant. Pour plus de clarté lors de vos recherches ou pour simplifier la détermination des scores des règles, pensez à filtrer les règles pour que les conséquences et les antécédents intéressants soient affichés de manière plus visible. Les options de filtrage de l'onglet Modèle d'un navigateur de règles permettent d'ouvrir une boîte de dialogue dans laquelle vous indiquez les qualifications de filtre.

Figure 12-12
Boîte de dialogue de filtrage du navigateur de règles



Conséquences. Sélectionnez Activer le filtre pour activer les options de filtrage des règles en fonction de l'inclusion ou de l'exclusion des conséquences indiquées. Sélectionnez Inclure l'un(e) des pour créer un filtre dans lequel les règles contiennent au moins l'une des conséquences indiquées. Vous pouvez également sélectionner Exclut pour créer un filtre excluant les conséquences indiquées. Vous pouvez sélectionner les conséquences à l'aide de l'icône de sélection située à droite de la zone de liste. Une boîte de dialogue répertoriant toutes les conséquences présentes dans les règles générées apparaît.

Remarque : Les conséquences peuvent contenir plusieurs éléments. Les filtres vérifient uniquement qu'une conséquence contient l'un des éléments indiqués.

Antécédents. Sélectionnez Activer le filtre pour activer les options de filtrage des règles en fonction de l'inclusion ou de l'exclusion des antécédents indiqués. Vous pouvez sélectionner les éléments à l'aide de l'icône de sélection située à droite de la zone de liste. Une boîte de dialogue répertoriant tous les antécédents présents dans les règles générées apparaît.

- Sélectionnez Inclure tou(te)s les pour définir le filtre comme étant inclusif, dans lequel tous les antécédents indiqués doivent être inclus dans une règle.
- Sélectionnez Inclure l'un(e) des pour créer un filtre dans lequel les règles contiennent au moins l'un des antécédents indiqués.
- Sélectionnez Exclut pour créer un filtre excluant les règles contenant un antécédent précis.

Confiance. Sélectionnez Activer le filtre pour activer les options de filtrage des règles en fonction du niveau de confiance d'une règle. Vous pouvez utiliser les commandes Min et Max pour indiquer un intervalle de confiance. Lorsque vous parcourez les modèles générés, la confiance est présentée sous forme de pourcentage. Lorsque vous déterminez le score de sorties, la confiance est exprimée sous la forme d'un chiffre compris entre 0 et 1.

Prise en charge des antécédents. Sélectionnez Activer le filtre pour activer les options de filtrage des règles en fonction du niveau de prise en charge des antécédents d'une règle. La prise en charge des antécédents indique la proportion des données d'apprentissage contenant les mêmes antécédents que la règle en cours, ce qui la rend identique à un indice de popularité. Vous pouvez utiliser les commandes Min et Max pour indiquer un intervalle utilisé pour filtrer les règles selon le niveau de prise en charge.

Lift. Sélectionnez Activer le filtre pour activer les options de filtrage des règles en fonction de la mesure du Lift (augmentation) d'une règle. *Remarque* : Le filtrage du Lift est uniquement disponible pour les modèles d'association créés après la version 8.5 ou pour les modèles antérieurs contenant une mesure d'augmentation. Les modèles Séquence ne comportent pas cette option.

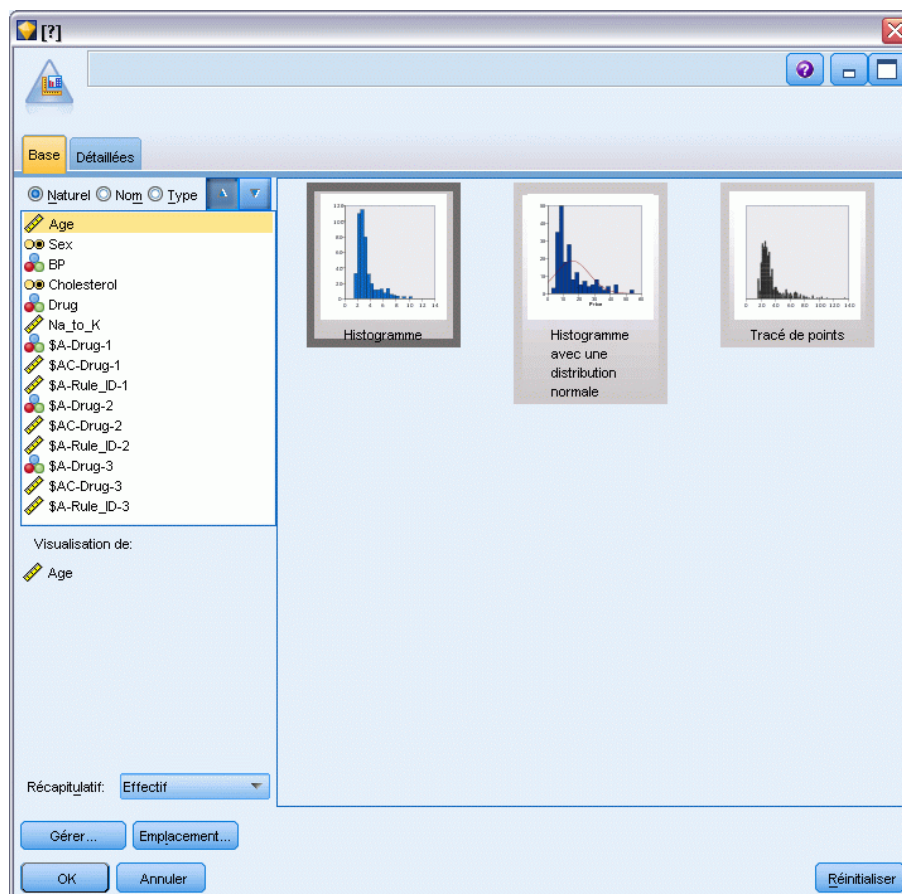
Cliquez sur OK pour appliquer tous les filtres activés dans cette boîte de dialogue.

Génération de graphiques pour les règles

Les noeuds d'association offrent de nombreuses informations, mais ils ne sont pas toujours dans un format très pratique pour les utilisateurs. Pour fournir les données d'une manière facilement incorporable aux rapports d'activité, aux présentations, etc., vous pouvez réaliser des diagrammes à partir des données sélectionnées. Dans l'onglet Modèle, vous pouvez générer un graphique pour la règle sélectionnée et créer ainsi un graphique pour les observations de cette règle uniquement.

- ▶ Dans l'onglet Modèle, sélectionnez la règle qui vous intéresse.
- ▶ Dans le menu Générer, sélectionnez Graphique (à partir de la sélection). L'onglet Base de Représentation graphique apparaît.

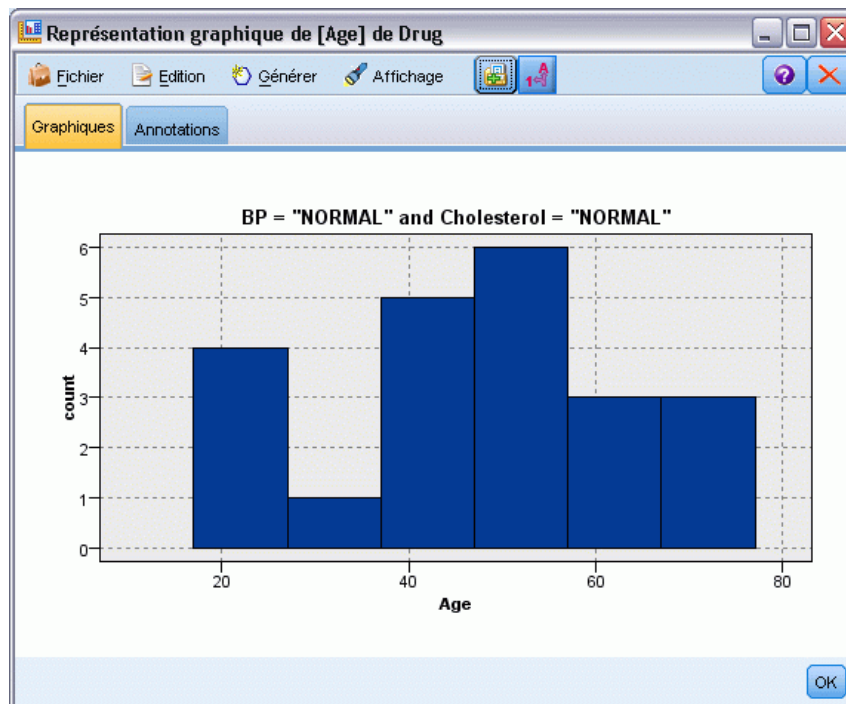
Figure 12-13
Boîte de dialogue Noeud Représentation graphique, onglet de Base



Remarque : Seuls les onglets de Base et Détaillé seront disponibles lorsque vous affichez le graphique de cette manière. [Pour plus d'informations, reportez-vous à la section Noeud Représentation Graphique dans le chapitre 5 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

- ▶ À l'aide des paramètres de l'onglet de Base ou Détaillé, spécifiez les détails à afficher sur le diagramme.
- ▶ Cliquez sur OK pour générer le diagramme.

Figure 12-14
Boîte de dialogue Noeud Représentation graphique, onglet de Base



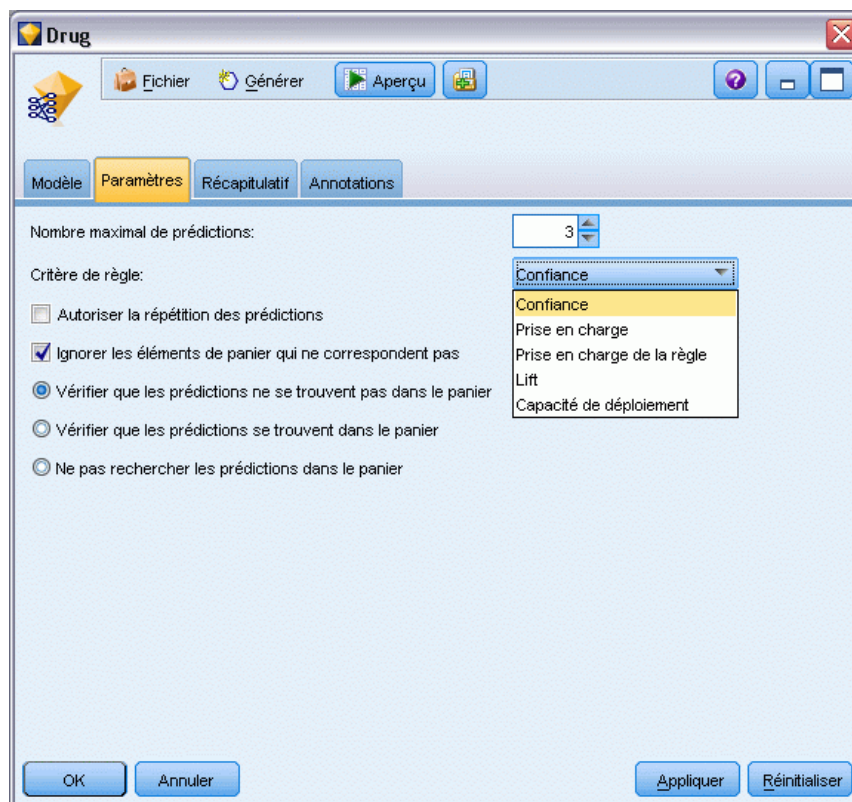
L'en-tête de graphique identifie la règle et les détails de l'antécédent à inclure.

Paramètres des nuggets de modèles de règle d'association

Cet onglet Paramètres est utilisé pour indiquer les options de détermination des scores des modèles d'association (Apriori et CARMA). Il est disponible uniquement lorsque le nugget de modèle a été ajouté à un flux à des fins de scoring.

Remarque : La boîte de dialogue permettant d'accéder à un modèle non affiné ne comprend pas l'onglet Paramètres, puisque les scores de ce modèle ne peuvent pas être déterminés. Pour déterminer les scores du modèle "brut", vous devez d'abord générer un ensemble de règles. [Pour plus d'informations, reportez-vous à la section Génération d'un ensemble de règles à partir d'un nugget de modèle d'association sur p. 437.](#)

Figure 12-15
Onglet Paramètres des nuggets de modèles de règle d'association



Nombre maximal de prédictions. Indiquez le nombre maximal de prévisions incluses pour chaque ensemble d'éléments de panier. Cette option est utilisée avec l'option Critères de règle ci-après pour produire les prévisions "supérieures" où *supérieures* indique le plus haut niveau de confiance, de prise en charge, de Lift, etc., indiqué ci-après.

Critère de règle. Sélectionnez la mesure utilisée pour déterminer la puissance des règles. Les règles sont triées selon la force des critères sélectionnés ici afin de renvoyer les prévisions de niveau supérieur d'un jeu d'éléments. Les critères disponibles sont les suivants :

- Confiance
- Prise en charge
- Prise en charge de la règle (Prise en charge * Confiance)
- Lift
- Deployabilité

Autoriser la répétition des prédictions. Sélectionnez cette option pour inclure plusieurs règles ayant la même conséquence lors de la détermination des scores. Par exemple, lorsque cette option est sélectionnée, les règles suivantes peuvent être évaluées :

bread & cheese -> wine
cheese & fruit -> wine

Désactivez cette option pour exclure la répétition des prévisions lors de la détermination des scores.

Remarque : Les règles avec plusieurs conséquences (bread & cheese & fruit -> wine & pate) sont considérées comme des répétitions des prévisions uniquement si toutes les conséquences (wine & pate) ont été prévues auparavant.

Ignorer les éléments de panier qui ne correspondent pas. Sélectionnez cette option pour ignorer la présence d'éléments supplémentaires dans le jeu d'éléments. Par exemple, lorsque cette option est sélectionnée pour un panier qui contient [tent & sleeping bag & kettle], la règle tent & sleeping bag -> gas_stove s'applique malgré l'élément supplémentaire (kettle) présent dans le panier.

Dans certaines circonstances, les éléments supplémentaires doivent être exclus. Par exemple, une personne achetant une tente, un sac de couchage et une bouilloire possède probablement déjà un réchaud à gaz, comme l'indique la présence de la bouilloire. En d'autres termes, un réchaud à gaz n'est probablement pas une bonne prévision. Dans ce cas, désélectionnez l'option Ignorer les éléments de panier qui ne correspondent pas pour vous assurer que les antécédents de règle correspondent exactement au contenu d'un panier. Par défaut, les éléments ne correspondant pas sont ignorés.

Vérifier que les prédictions ne se trouvent pas dans le panier. Sélectionnez cette option pour vous assurer que les conséquences ne se trouvent pas également dans le panier. Par exemple, si la détermination du score a pour objet la recommandation d'un meuble, une personne dont le panier contient déjà une table de salle à manger n'en achètera probablement pas d'autre. Dans ce cas, sélectionnez cette option. Cependant, si les produits concernés sont périssables ou jetables (comme le fromage, le lait pour nourrisson ou les mouchoirs), les règles dans lesquelles la conséquence figure déjà dans le panier peuvent être valables. Dans ce dernier cas, l'option la plus adaptée est l'option Ne pas rechercher les prédictions dans le panier suivante.

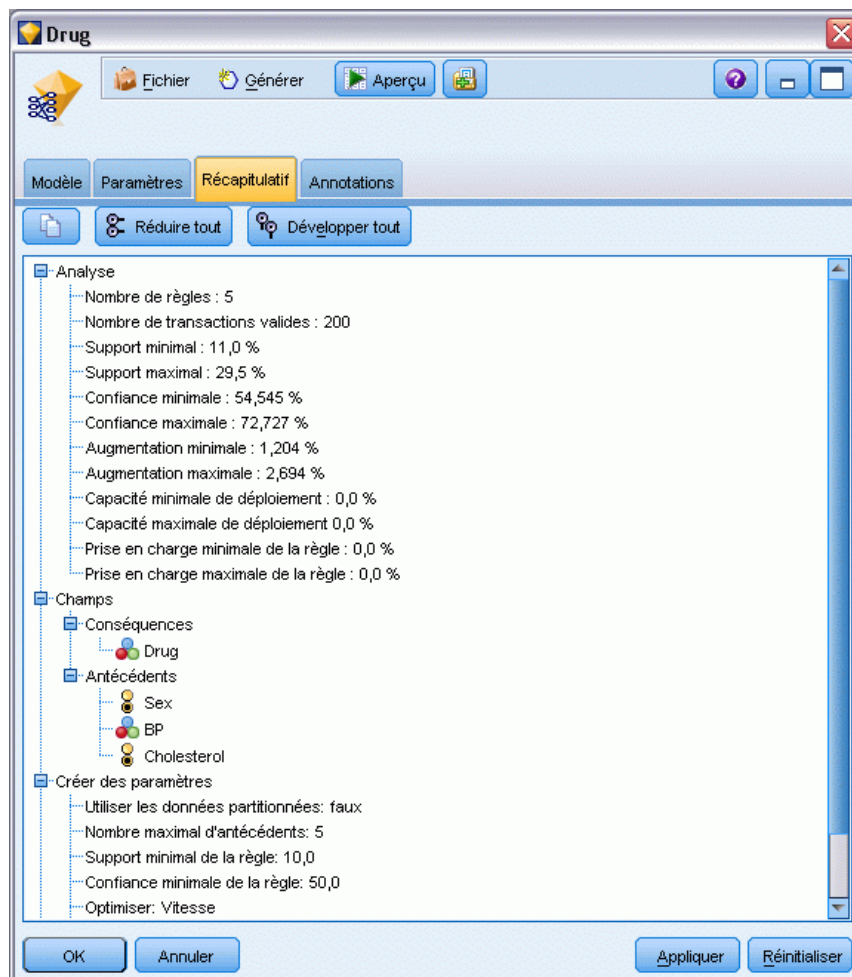
Vérifier que les prédictions se trouvent dans le panier. Sélectionnez cette option pour vous assurer que les conséquences se trouvent également dans le panier. Cette approche est utile lorsque vous essayez de vous faire une idée des transactions ou des clients existants. Par exemple, vous souhaitez peut-être identifier les règles ayant la plus grande augmentation (lift), puis explorer les clients qui respectent ces règles.

Ne pas rechercher les prédictions dans le panier. Sélectionnez cette option pour inclure toutes les règles lors de la détermination des scores, indépendamment de la présence ou de l'absence de conséquences dans le panier.

Récapitulatif du nugget du modèle de règle d'association

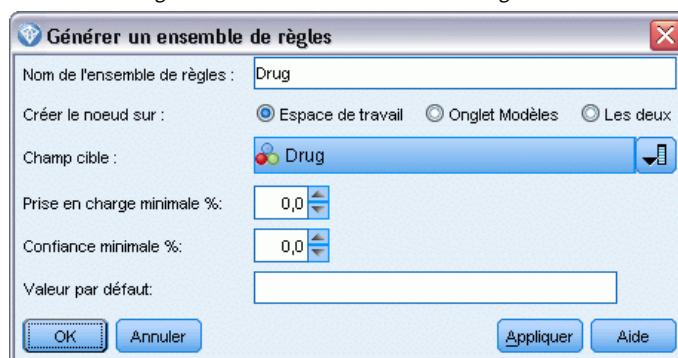
L'onglet Récapitulatif d'un nugget de modèle Règle d'association affiche le nombre de règles trouvées, ainsi que les valeurs minimale et maximale de prise en charge, de Lift, de confiance et de capacité de déploiement des règles.

Figure 12-16
Onglet Récapitulatif - Nuggets de modèles de règle d'association



Génération d'un ensemble de règles à partir d'un nugget de modèle d'association

Figure 12-17
Boîte de dialogue Générer un ensemble de règles



Vous pouvez utiliser les nuggets de modèles d'association, tels que CARMA et Apriori, pour déterminer directement le score des données ou vous pouvez d'abord générer un sous-ensemble de règles appelé **ensemble de règles**. Les ensembles de règles sont particulièrement utiles lorsque vous utilisez un modèle brut, qui ne peut pas être utilisé directement pour le scoring. [Pour plus d'informations, reportez-vous à la section Modèles bruts dans le chapitre 3 sur p. 77.](#)

Pour générer un ensemble de règles, sélectionnez Ensemble de règles dans le menu Générer du navigateur de nuggets de modèles. Vous pouvez utiliser les options suivantes pour convertir des règles en un ensemble de règles :

Nom de l'ensemble de règles. Permet d'attribuer un nom au nouveau noeud Ensemble de règles généré.

Créer le noeud sur. Détermine l'emplacement du nouveau noeud Ensemble de règles généré. Sélectionnez Canevas, Palette GM ou Les deux.

Champ cible. Détermine quel champ de sortie sera utilisé pour le noeud Ensemble de règles généré. Sélectionnez un champ de sortie dans la liste.

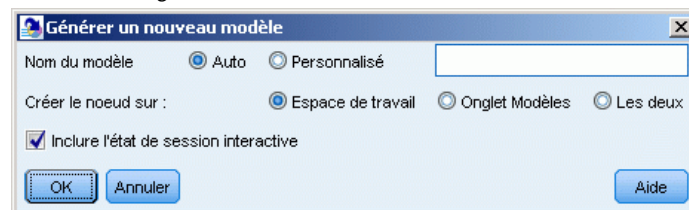
Prise en charge minimale. Indique la prise en charge minimale que doivent avoir les règles pour apparaître dans l'ensemble de règles généré. Les règles dont la prise en charge est inférieure à la valeur spécifiée ne seront pas incluses dans le nouvel ensemble de règles.

Confiance minimale. Indique la confiance minimale que doivent avoir les règles pour apparaître dans l'ensemble de règles généré. Les règles avec une confiance inférieure à la valeur spécifiée ne seront pas incluses dans le nouvel ensemble de règles.

Valeur par défaut. Permet de spécifier une valeur par défaut pour le champ cible attribué aux enregistrements évalués auxquels aucune règle ne s'applique.

Génération d'un modèle filtré

Figure 12-18
Boîte de dialogue Générer un nouveau modèle



Pour générer un modèle filtré à partir d'un modèle d'association, tel qu'un noeud Apriori, CARMA ou Ensemble de règles de séquence, sélectionnez Modèle filtré dans le menu Générer du navigateur de nuggets de modèles. Le modèle de sous-ensemble qui est créé comporte uniquement les règles actuellement affichées dans le navigateur. *Remarque* : Vous ne pouvez pas générer de modèles filtrés pour les modèles non affinés.

Vous pouvez indiquer les options suivantes pour filtrer les règles :

Nom du nouveau modèle. Permet d'attribuer un nom au nouveau noeud Modèle filtré.

Créer le noeud sur. Détermine l'emplacement du nouveau noeud Modèle filtré. Sélectionnez Canevas, Palette GM ou Les deux.

Numérotation des règles. Indiquez comment les ID de règle seront numérotés dans le sous-ensemble des règles incluses dans le modèle filtré.

- **Conserver les numéros d'ID de règle d'origine.** Sélectionnez cette option pour conserver la numérotation d'origine des règles. Par défaut, les règles se voient attribuer un ID correspondant à leur ordre de découverte par l'algorithme. Cet ordre peut varier en fonction de l'algorithme employé.
- **Renommer les règles de manière consécutive en commençant par.** Sélectionnez cette option pour attribuer de nouveaux ID de règle aux règles filtrées. Les nouveaux ID sont attribués selon l'ordre de tri affiché dans le tableau du navigateur de règles de l'onglet Modèle et commencent par le chiffre indiqué ici. Vous pouvez indiquer le premier chiffre des ID à l'aide des flèches situées à droite.

Scoring des règles d'association

Les scores produits lors de l'exécution de nouvelles données via un nugget de modèle de règle d'association sont renvoyés dans des champs distincts. Trois nouveaux champs sont ajoutés pour chaque prévision : *P* représente la prévision, *C*, la confiance et *I*, l'ID de règle. L'organisation de ces champs de sortie dépend du format des données d'entrée (transactionnel ou tabulaire). Reportez-vous à Données tabulaires et données transactionnelles sur p. 416 pour obtenir une présentation de ces formats.

Par exemple, supposons que vous déterminiez les scores des données du panier à l'aide d'un modèle qui génère des prévisions sur la base des trois règles suivantes :

Rule_15 bread&wine -> meat (confidence 54%)
 Rule_22 cheese -> fruit (confidence 43%)
 Rule_5 bread&cheese -> frozveg (confidence 24%)

Données tabulaires. Pour les données tabulaires, les trois prévisions (3 est la valeur pas défaut) sont renvoyées dans un enregistrement unique.

Table 12-1
 Scores au format tabulaire

ID	Pain	Vin	Fromage	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	viande	0.54	15	fruit	0.43	22	légumes surgelés	.24	5

Données transactionnelles. En ce qui concerne les données transactionnelles, un enregistrement distinct est généré pour chaque prévision. Les prévisions sont toujours ajoutées dans des colonnes distinctes, mais les scores sont renvoyés au fur et à mesure de leur calcul. Vous obtenez ainsi des enregistrements comportant des prévisions incomplètes, comme l'illustre l'exemple de sortie ci-dessous. La deuxième et la troisième prévisions (P2 et P3) sont vides dans le premier enregistrement, ainsi que dans les confiances et les ID de règle associés. Lorsque les scores sont renvoyés, l'enregistrement final comporte alors les trois prévisions.

Table 12-2
Scores au format transactionnel

ID	Élément	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	pain	viande	0.54	14	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$	\$null\$
Fred	fromage	viande	0.54	14	fruit	0.43	22	\$null\$	\$null\$	\$null\$
Fred	vin	viande	0.54	14	fruit	0.43	22	légumes surgelés	0.24	5

Pour inclure uniquement les prévisions complètes à des fins de création de rapport ou de déploiement, utilisez un noeud Sélectionner pour sélectionner les enregistrements complets.

Remarque : Pour plus de clarté, les noms de champ utilisés dans ces exemples sont abrégés. Lors de l'utilisation réelle, les champs de résultats des modèles d'association portent les noms suivants :

Nouveau champ	Exemple de nom de champ
Prévision	<i>\$A-TRANSACTION_NUMBER-I</i>
Confiance (ou autre critère)	<i>\$AC-TRANSACTION_NUMBER-I</i>
ID de règle	<i>\$A-Rule_ID-I</i>

Règles à plusieurs conséquences

L'algorithme CARMA autorise les règles à plusieurs conséquences— par exemple :

bread -> wine&cheese

Lorsque vous déterminez des scores de règles “doubles” de ce type, les prévisions sont renvoyées au format affiché dans le tableau suivant :

Table 12-3
Résultats des scores comprenant une prévision à plusieurs conséquences

ID	Pain	Vin	Fromage	P1	C1	I1	P2	C2	I2	P3	C3	I3
Fred	1	1	1	viande&lég	0.54	16	fruit	0.43	22	légumes surgelés	.24	5

Dans certains cas, vous pourrez avoir besoin de partager ces scores avant le déploiement. Pour diviser une prévision à plusieurs conséquences, vous devez analyser le champ via les fonctions de chaîne CLEM. [Pour plus d'informations, reportez-vous à la section Fonctions sur chaînes dans le chapitre 8 dans *Guide de l'utilisateur de IBM SPSS Modeler 15*.](#)

Déploiement des modèles d'association

Lors du scoring des modèles d'association, les prévisions et les confiances apparaissent dans des colonnes distinctes (où *P* représente la prévision, *C* la confiance et *I* l'ID de règle). Ce type d'affichage s'applique quel que soit le format des données d'entrée (tabulaire ou transactionnel). [Pour plus d'informations, reportez-vous à la section Scoring des règles d'association sur p. 439.](#)

Figure 12-19
Scores tabulaires, avec les prévisions affichées sous forme de colonnes

	ID	A	B	C	P1	C1	I1	P2	C2	I2	P3	C3	I3
1	Tom	1	1	1	D	9	1	E	5	23	F	3	9
2	Bob	0	1	1	F	3	9	E	2	15	D	1	4

Lorsque vous préparez des scores en vue d'une opération de déploiement, il se peut que votre application requiert la transposition des données de sortie en un format présentant les prévisions sous forme de lignes plutôt que sous forme de colonnes (une prévision par ligne ; ce format est également appelé format « till-roll » (déroulant)).

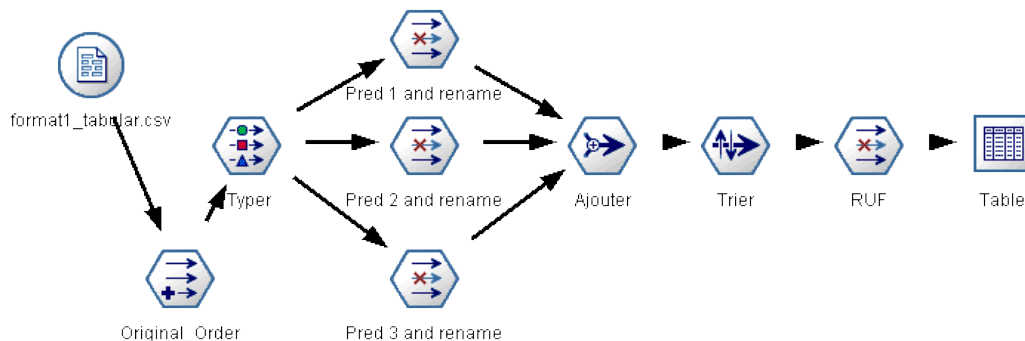
Figure 12-20
Scores transposés, avec les prévisions affichées sous forme de lignes

	ID	A	B	C	Pred	Crit	Rule_ID
1	Tom	1	1	1	D	9	1
2	Tom	1	1	1	E	5	23
3	Tom	1	1	1	F	3	9
4	Bob	0	1	1	F	3	9
5	Bob	0	1	1	.	\$null\$	\$null\$
6	Bob	0	1	1	.	\$null\$	\$null\$

Transposition des scores tabulaires

Vous pouvez transposer des scores tabulaires de colonnes en lignes en réalisant une série d'étapes dans IBM® SPSS® Modeler, comme indiqué dans l'énoncé qui suit.

Figure 12-21
Exemple de flux utilisé pour transposer les données tabulaires au format till-roll

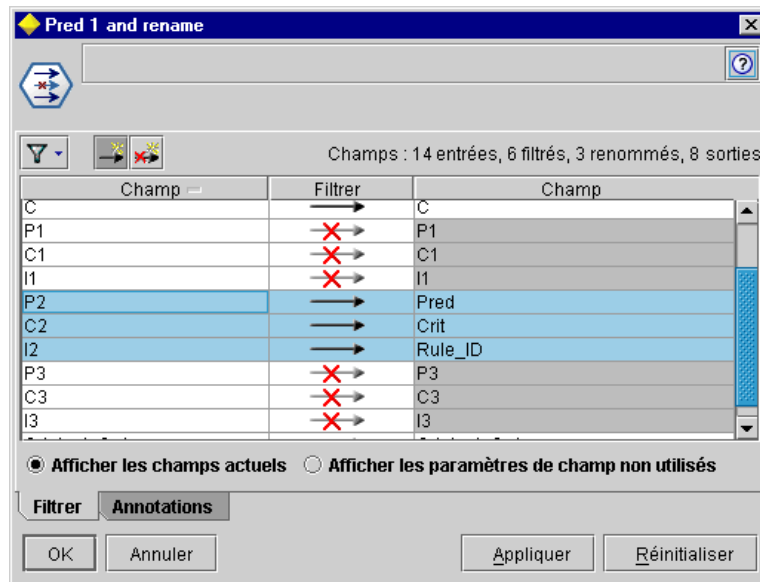


- ▶ Utilisez la fonction @INDEX d'un nœud Calculer pour vérifier l'ordre actuel des prévisions et enregistrer cet indicateur dans un nouveau champ, tel que *Ordre d'origine*.
- ▶ Ajoutez un nœud Typer pour vous assurer que tous les champs ont été instanciés.

- Utilisez un noeud Filtrer pour renommer les champs de prévision, de confiance et d'ID par défaut (*P1, C1, I1*) en champs communs, tels que *Prévisions, Critères* et *ID de règle*, qui seront utilisés ultérieurement pour ajouter des enregistrements. Vous aurez besoin d'un noeud Filtrer pour chaque prévision générée.

Figure 12-22

Filtrage des champs des prévisions 1 et 3 lors de l'attribution de nouveaux noms aux champs de la prévision 2.



- Utilisez un noeud Ajouter pour ajouter des valeurs aux valeurs de *prévision, de confiance* et d'*ID de règle* partagées.
- Reliez un noeud Trier pour trier les enregistrements dans l'ordre croissant pour le champ *Ordre d'origine* et décroissant pour *Critères*, qui est le champ utilisé pour trier les prévisions selon des critères tels que la confiance, Lift et la prise en charge.
- Utilisez un autre noeud Filtrer pour filtrer le champ *Ordre d'origine* à partir de la sortie.

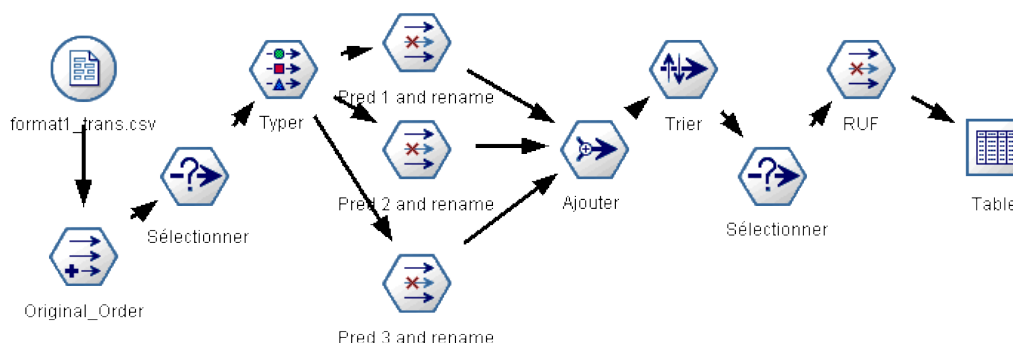
A ce stade, les données sont prêtes à être déployées.

Transposition de scores transactionnels

Le procédé est identique à celui de transposition des scores transactionnels. Par exemple, le flux ci-dessous transpose les scores dans un format présentant une prévision unique sur chaque ligne, comme le requiert le déploiement.

Figure 12-23

Exemple de flux utilisé pour transposer les données transactionnelles au format till-roll



Si vous ajoutez deux noeuds Sélectionner, le procédé est identique à celui décrit précédemment pour les données tabulaires.

- Le premier noeud Sélectionner est utilisé pour comparer les ID de règle entre les enregistrements adjacents et inclure uniquement des enregistrements uniques ou non définis. Ce noeud Sélectionner utilise l'expression CLEM pour sélectionner les enregistrements : $ID \neq @OFFSET(ID,-1)$ or $@OFFSET(ID,-1) = \text{undef}$.
- Le deuxième noeud Sélectionner est utilisé pour supprimer les règles superflues ou celles pour lesquelles Rule_ID a une valeur nulle. Ce noeud Sélectionner utilise l'expression CLEM suivante pour supprimer les enregistrements : $\text{not}(@NULL(\text{Rule_ID}))$.

Pour plus d'informations sur la transposition des scores en vue du déploiement, contactez l'assistance technique.

Noeud Séquence

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.

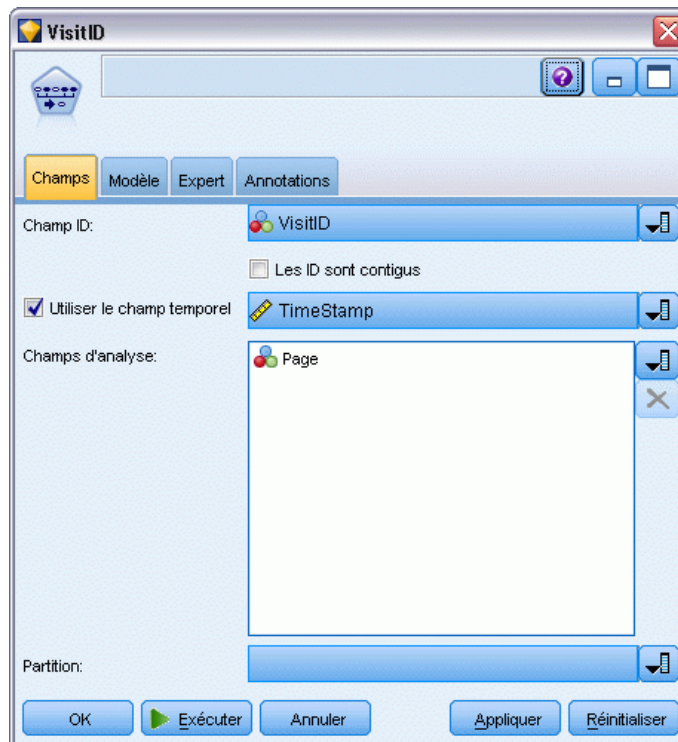
Conditions requises. Pour créer un ensemble de règles Séquence, vous devez renseigner un champ ID obligatoire, un champ de temps facultatif et au moins un champ d'analyse. Remarque : ces paramètres doivent être définis dans l'onglet Champs du noeud de modélisation ; ils ne peuvent pas être lus à partir d'un noeud Typier en amont. Le champ ID peut être de n'importe quel rôle ou niveau de mesure. Si vous indiquez un champ Temps, il peut avoir n'importe quel rôle mais son stockage doit afficher un nombre, une date, une heure ou un horodatage. Si vous ne définissez pas de champ Temps, le noeud Séquence utilise un horodatage implicite et se sert des numéros de ligne comme valeurs temporelles. Les champs d'analyse peuvent être de n'importe quel niveau de

mesure ou rôle, mais tous les champs d'analyse doivent être du même type. S'ils sont numériques, ils doivent contenir des intervalles d'entiers (et non des intervalles de réels).

Puissance. Le noeud Séquence est basé sur l'algorithme de règles d'association CARMA, qui utilise une méthode efficace de double lecture pour rechercher des séquences. En outre, le noeud de modèle généré créé par un noeud Séquence peut être inséré dans un flux de données pour créer des prévisions. Le noeud de modèle généré peut également créer des super noeuds pour détecter et compter des séquences spécifiques, et effectuer des prévisions basées sur des séquences spécifiques.

Noeud Séquence - Options de champs

Figure 12-24
Noeud Séquence - Options de champs



Avant d'exécuter un noeud Séquence, vous devez renseigner le champ ID et le champ d'analyse dans l'onglet Champs du noeud Séquence. Si vous souhaitez utiliser le champ Temps, sélectionnez l'option correspondante dans cet onglet.

Champ 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).

- **Les ID sont contigus.** Si vos données sont prétriées de façon à ce que tous les enregistrements avec le même ID sont regroupés dans le flux de données, sélectionnez cette option pour accélérer le traitement. Si vos données ne sont pas pré-triées (ou si vous n'en êtes pas certain), ne sélectionnez pas cette option ; le noeud Séquence triera automatiquement les données.

Remarque : Si vos données ne sont pas triées et que vous sélectionnez cette option, vous risquez d'obtenir des résultats incorrects dans votre modèle Séquence.

Champ Temps. Pour utiliser un champ dans les données afin d'indiquer l'heure des événements, sélectionnez Utiliser le champ temporel et choisissez le champ à utiliser. Le champ Temps doit afficher un nombre, une date, une heure ou un horodatage. Si aucun champ de temps n'est spécifié, le système suppose que les enregistrements proviennent de la source de données selon un ordre séquentiel et les numéros des enregistrements sont alors utilisés comme valeurs temporelles (le premier enregistrement arrive au moment "1"; le deuxième au moment "2"; etc.).

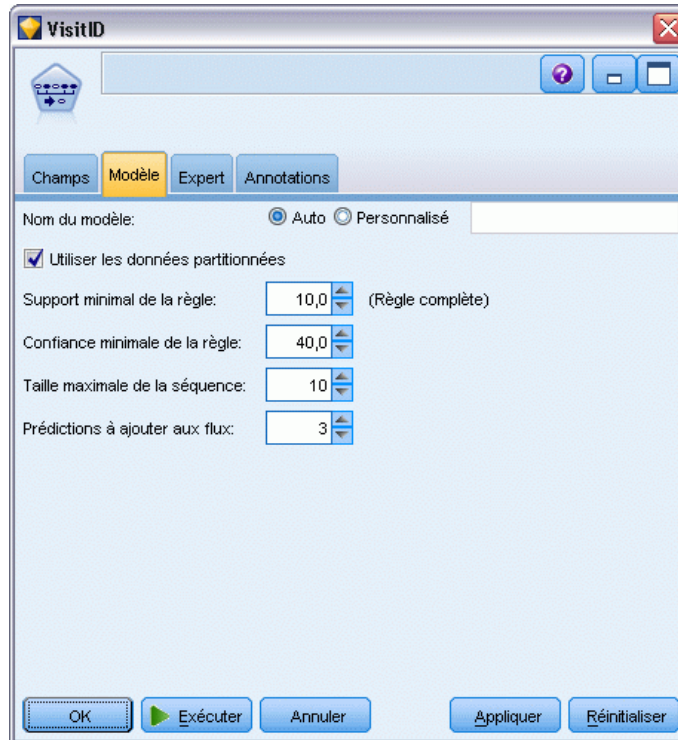
Champs d'analyse. Permet d'ajouter des informations aux champs d'analyse du modèle. Ces champs contiennent les événements d'intérêt concernant la modélisation des séquences.

Le noeud Séquence peut gérer des données au format tabulaire ou transactionnel. Si vous utilisez plusieurs champs avec des données transactionnelles, le système suppose que les éléments spécifiés dans ces champs pour un enregistrement spécifique représentent les éléments trouvés dans une même transaction avec un seul horodatage. [Pour plus d'informations, reportez-vous à la section Données tabulaires et données transactionnelles sur p. 416.](#)

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 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.)

Noeud Séquence - Options modèle

Figure 12-25
Noeud Séquence - Options de 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 (%). Vous pouvez spécifier un critère de prise en charge. L'option de **prise en charge de la règle** correspond à la proportion d'ID des données d'apprentissage qui contiennent la séquence. Si vous souhaitez extraire des séquences plus standard, augmentez la valeur de ce paramètre.

Confiance minimale de la règle (%). Vous pouvez indiquer le degré de confiance minimale pour toutes les séquences. La **confiance** correspond au pourcentage des ID pour lesquels une prévision correcte est effectuée, par rapport à l'ensemble des ID pour lesquels la règle effectue une prévision. Elle est calculée de la manière suivante : nombre d'ID pour lesquels la séquence entière est trouvée, divisé par le nombre d'ID pour lesquels les antécédents sont trouvés, sur la base des données d'apprentissage. Les séquences dont le niveau de confiance est inférieur au niveau spécifié sont ignorées. Si vous obtenez des séquences inintéressantes ou des séquences trop nombreuses, augmentez la valeur de ce paramètre. En revanche, réduisez la valeur si le nombre de séquences n'est pas assez important.

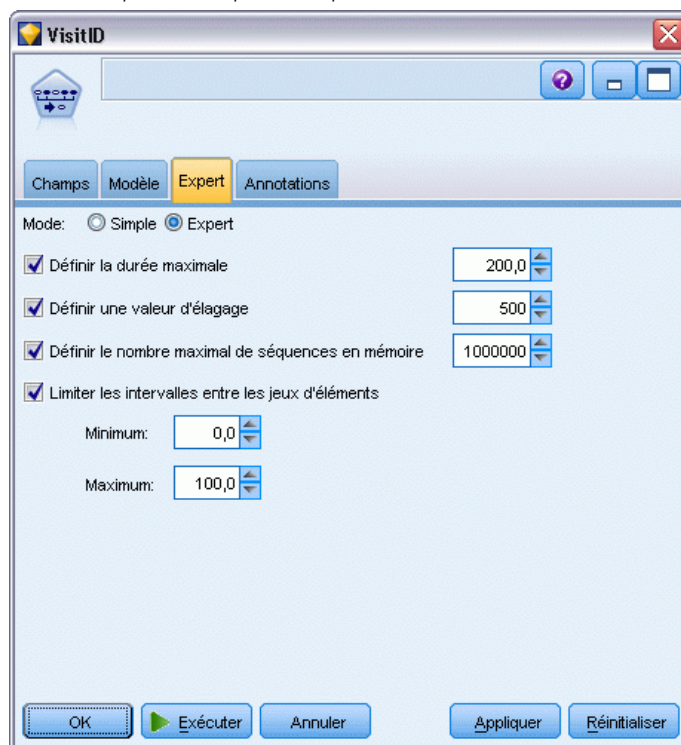
Taille maximale de la séquence. Vous pouvez définir le nombre maximal de *jeux d'éléments* (et non d'*éléments*) que doit contenir une séquence. Si les séquences qui vous intéressent sont relativement courtes, diminuez la valeur de ce paramètre afin d'accélérer la génération de l'ensemble de séquences.

Prédictions à ajouter aux flux. Indiquez le nombre de prévisions à ajouter au flux par le noeud de modèle généré. [Pour plus d'informations, reportez-vous à la section Nuggets de modèles de séquences sur p. 449.](#)

Noeud Séquence - Options expert

Les options expert suivantes, qui s'adressent aux utilisateurs ayant une bonne connaissance du fonctionnement du noeud Séquence, permettent d'affiner le processus de création de modèle. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 12-26
Noeud Séquence - Options expert



Définir la durée maximale. Lorsque cette option est sélectionnée, les séquences extraites sont celles dont la durée (intervalle de temps entre le premier et dernier jeu d'éléments) est inférieure ou égale à la valeur indiquée. Si aucun champ Temps n'a été spécifié, la durée est alors exprimée en utilisant les numéros de ligne (enregistrements) des données brutes. Si le champ temporel utilisé est un champ d'heure, de date ou d'horodatage, la durée est exprimée en secondes. Pour les champs numériques, la durée est exprimée dans la même unité que le champ.

Définir une valeur d'élagage. L'algorithme CARMA utilisé dans le noeud Séquence supprime (**élague**) périodiquement de sa liste de jeux d'éléments potentiels les jeux peu fréquents afin d'économiser de la mémoire. Sélectionnez cette option pour régler la fréquence de l'élagage. Le nombre spécifié détermine la fréquence de l'élagage. Entrez une valeur faible pour diminuer les besoins en mémoire de l'algorithme (le temps d'apprentissage requis risque d'être augmenté). Entrez une valeur élevée pour accélérer l'apprentissage (les besoins en mémoire de l'algorithme risquent d'être augmentés).

Définir le nombre maximal de séquences en mémoire. Si cette option est sélectionnée lors de la construction de modèle, l'algorithme CARMA limite son stockage en mémoire des séquences potentielles au nombre de séquences défini. Sélectionnez cette option si IBM® SPSS® Modeler utilise trop de mémoire pendant la construction de modèles de séquences. Le nombre maximal de séquences indiqué ici correspond au nombre de séquences potentielles enregistrées en interne lors de la construction du modèle. Ce nombre doit être supérieur au nombre de séquences que vous souhaitez voir apparaître dans le modèle final.

Limiter les intervalles entre les jeux d'éléments. Cette option vous permet de spécifier des limites pour les intervalles de temps qui séparent les jeux d'éléments. Lorsque cette option est sélectionnée, les jeux d'éléments dont les intervalles de temps sont inférieurs à la valeur du paramètre Intervalle minimal entre les éléments ou supérieurs à la valeur du paramètre Intervalle maximal entre les éléments sont ignorés. Utilisez cette option pour éviter d'extraire des séquences dont les intervalles de temps sont trop importants ou, au contraire, des séquences qui surviennent dans un intervalle de temps très court.

Remarque : Si le champ temporel utilisé est un champ d'heure, de date ou d'horodatage, l'intervalle est exprimé en secondes. Pour les champs numériques, l'intervalle de temps est exprimé dans la même unité que le champ Temps.

Prenons cette liste de transactions comme exemple :

ID	Heure	Contenu
1001	1	pommes
1001	2	pain
1001	5	fromage
1001	6	vinaigrette

Si vous créez un modèle à partir de ces données en définissant le paramètre Intervalle minimal entre les éléments sur 2, vous obtiendrez les séquences suivantes :

apples -> cheese

apples -> dressing

bread -> cheese

bread -> dressing

Vous n'obtiendrez pas de séquences du type *apples* -> *bread* dans la mesure où l'intervalle entre *apples* et *bread* est inférieur à l'intervalle minimal spécifié. De même, si les données sont les suivantes :

ID	Heure	Contenu
1001	1	pommes
1001	2	pain
1001	5	fromage
1001	20	vinaigrette

et que le paramètre Intervalle maximal entre les éléments est défini sur 10, vous n'obtiendrez aucune séquence comportant l'élément *dressing*, dans la mesure où l'intervalle entre *cheese* et *dressing* est trop important pour que l'on puisse considérer que ces deux éléments appartiennent à la même séquence.

Nuggets de modèles de séquences

Les nuggets de modèles de séquence représentent les séquences trouvées pour un champ de sortie particulier détecté par le noeud Séquence ; ils peuvent être ajoutés à des flux pour générer des prévisions.

Lorsque vous exécutez un flux contenant un noeud de séquence, ce dernier crée pour chaque prévision générée par le modèle de séquence deux nouveaux champs, l'un contenant la prévision et l'autre la valeur de confiance associée. Par défaut, trois paires de champs contenant les trois premières prévisions (et les valeurs de confiance associées) sont ajoutées. Vous pouvez modifier le nombre de prévisions générées en définissant les options de modèle du noeud Séquence lors de la création du modèle ; vous pouvez également le faire après l'ajout du nugget de modèle à un flux dans l'onglet Paramètres. [Pour plus d'informations, reportez-vous à la section Paramètres du nugget de modèle de séquence sur p. 453.](#)

Les noms des nouveaux champs sont constitués à partir du nom du modèle. Les noms des champs sont *\$\$sequence-n* pour le champ de prévision (*n* indiquant la *n*ème prévision) et *\$\$SC-sequence-n* pour le champ de confiance. Dans un flux comportant plusieurs noeuds Règles de séquence connectés en série, un nombre séquentiel sera ajouté au préfixe du nom des nouveaux champs afin de pouvoir les distinguer. Le premier noeud Ensemble de séquences du flux utilise les noms d'origine, le deuxième noeud les noms commençant par *\$\$1-* et *\$\$SC1-*, le troisième noeud les noms commençant par *\$\$2-* et *\$\$SC2-*, et ainsi de suite. Les prévisions sont affichées en fonction de leur degré de confiance, par ordre décroissant. Ainsi le champ *\$\$sequence-1* contient la prévision dotée du degré de confiance le plus élevé, le champ *\$\$sequence-2* contient la prévision dotée du second degré de confiance le plus élevé et ainsi de suite. Pour les enregistrements où le nombre de prévisions disponibles est inférieur au nombre de prévisions demandées, les prévisions restantes contiennent la valeur *\$null\$*. Par exemple, si seulement deux prévisions peuvent être établies pour un enregistrement donné, la valeur de *\$\$sequence-3* et *\$\$SC-sequence-3* sera *\$null\$*.

Pour chaque enregistrement, les règles du modèle sont comparées à l'ensemble de transactions traité jusque-là pour l'ID en cours, y compris l'enregistrement actuel et tous les enregistrements précédents dotés du même ID et dont l'horodatage est antérieur. Les *k* règles présentant les valeurs de confiance les plus élevées et s'appliquant à cet ensemble de transactions sont utilisées pour générer les *k* prévisions de l'enregistrement (*k* étant le nombre de prévisions défini dans

l'onglet Paramètres une fois le modèle ajouté au flux). (Si plusieurs règles prévoient le même résultat pour l'ensemble de transactions, c'est celle dont le degré de confiance est le plus élevé qui est utilisée.) [Pour plus d'informations, reportez-vous à la section Paramètres du nugget de modèle de séquence sur p. 453.](#)

Tout comme avec d'autres types de modèle de règle d'association, le format des données doit correspondre au format utilisé lors de la création du modèle de séquence. Par exemple, vous ne pouvez utiliser les modèles créés à l'aide de données tabulaires que pour déterminer les scores des données tabulaires. [Pour plus d'informations, reportez-vous à la section Scoring des règles d'association sur p. 439.](#)

Remarque : lorsque vous évaluez des données en utilisant dans un flux un noeud Ensemble de séquences généré, tout paramètre de tolérance et d'intervalle que vous avez défini lors de la création du modèle est ignoré.

Prévisions des règles de séquence

Le noeud gère les enregistrements en fonction de l'heure (ou de l'ordre si aucun champ d'horodatage n'a été défini lors de la création du modèle). Les enregistrements doivent être triés par les champs d'ID et d'horodatage (s'ils existent). Toutefois, les prévisions ne sont pas liées à l'horodatage de l'enregistrement auquel elles sont ajoutées. Elles font simplement référence aux éléments les plus susceptibles de se produire *à un certain moment dans le futur* étant donné l'historique des transactions, de l'ID actuel jusqu'à l'enregistrement actuel.

Notez que les prévisions pour chaque enregistrement ne dépendent pas nécessairement des transactions de l'enregistrement en question. Si les transactions de l'enregistrement en cours ne déclenchent pas de règle spécifique, les règles sont sélectionnées en fonction des transactions antérieures relatives à l'ID actuel. Autrement dit, si l'enregistrement en cours n'ajoute aucune information prévisionnelle utile à la séquence, la prévision de la dernière transaction utile de cet ID est reportée sur cet enregistrement.

Par exemple, supposons que vous disposez d'un modèle de séquence avec pour seule règle

Jam -> Bread (0.66)

et que vous lui transmettez les enregistrements suivants :

ID	Achat	Prévision
001	jam	bread
001	milk	bread

Comme vous vous y attendiez, le premier enregistrement génère la prévision *pain*. Le second enregistrement contient également la prévision relative à *pain*, car aucune règle ne propose *confiture* suivi de *lait*. Par conséquent, la transaction *lait* n'apportant aucune information utile, la règle Jam -> Bread reste en vigueur.

Génération de noeuds

Le menu Générer vous permet de créer des super noeuds sur la base du modèle de séquence.

- **Super noeud Règle.** Crée un super noeud pouvant détecter et comptabiliser les occurrences de séquences dans des données évaluées. Lorsqu'aucune règle n'est sélectionnée, cette option est désactivée. [Pour plus d'informations, reportez-vous à la section Génération d'un super noeud Règle à partir d'un nugget de modèle de séquence sur p. 455.](#)
- **Modèle vers palette.** Renvoie le modèle à la palette de modèles. Ceci est utile dans des situations où un collègue peut vous avoir envoyé un flux contenant le modèle mais pas le modèle même.

Détails du nugget de modèle de séquence

L'onglet Modèle d'un nugget de modèle de séquence affiche les règles extraites par l'algorithme. Chaque ligne du tableau représente une règle, l'antécédent (partie « if » de la règle) se trouvant dans la première colonne et la conséquence (partie « then » de la règle) dans la seconde.

Figure 12-27
Onglet Modèle - Nugget de Séquence

Antécédent	Conséquence	% de prise en charge	% de confiance
login.asp?	personal.asp	21,591	100,0
login.asp?	login.asp?	21,591	100,0
login.asp?	personal.asp	21,591	100,0
login.asp?	personal.asp	21,591	100,0
splash.htm	main.htm	70,455	83,871
splash.htm	login.asp	59,091	51,923
main.htm	login.asp	70,455	43,548
splash.htm	login.asp	76,136	41,791
main.htm	main.htm	76,136	37,313
splash.htm	main.htm	59,091	36,538
main.htm	main.htm		

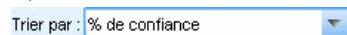
Les règles sont affichées dans le format suivant :

Antécédent	Conséquence
beer and cannedveg	beer
fish	fish
fish	fish

La première règle exemple est interprétée de la façon suivante : *pour les ID comportant « bière » et « conserves légumes » dans la même transaction, il est probable qu'il existe une autre occurrence de « bière »*. La deuxième règle exemple est interprétée de la façon suivante : *pour les ID comportant « poisson » dans une transaction et « poisson » dans une autre, il est probable qu'il existe une autre occurrence de « poisson »*. Dans la première règle, les achats *bière* et *conserves légumes* sont simultanés ; dans la seconde règle, l'achat *poisson* fait l'objet de deux transactions distinctes.

Menu Trier. Le bouton de menu Trier de la barre d'outils contrôle le tri des règles. Pour modifier l'ordre de tri (ascendant ou descendant), utilisez le bouton d'ordre du tri (flèche vers le haut ou vers le bas).

Figure 12-28
Options de tri de la barre d'outils



Vous pouvez trier les règles par :

- % de support
- % de confiance
- % de prise en charge de la règle
- Conséquence
- Premier antécédent
- Dernier antécédent
- Nombre d'éléments (antécédents)

Par exemple, le tableau suivant est trié dans l'ordre décroissant en fonction du nombre d'articles : Les règles qui présentent plusieurs articles dans le jeu antécédent précèdent celles qui en présentent moins.

Antécédent	Conséquence
beer and cannedveg and frozenmeal	frozenmeal
beer and cannedveg	beer
fish fish	fish
softdrink	softdrink

Menu Afficher/masquer les critères. Le bouton du menu Afficher/Masquer les critères (icône de la grille) contrôle les options d'affichage des règles. Les options d'affichage suivantes sont disponibles :

- L'option **Instances** affiche des informations sur le nombre d'ID uniques pour lesquels la *séquence complète*—(à la fois les antécédents et les conséquences)—se produit. (Une différence est à noter avec les modèles d'association : pour ces derniers, le nombre d'instances se rapporte au nombre d'ID pour lesquels *seuls* les antécédents s'appliquent.) Par exemple, pour la règle *bread -> cheese*, le nombre d'ID incluant les antécédents *pain* et *fromage* dans les données d'apprentissage est appelé **instances**.

- L'option **Prise en charge** affiche la proportion d'ID, dans les données d'apprentissage, pour lesquels les antécédents sont vrais (true). Par exemple, si 50 % des données d'apprentissage incluent l'antécédent *pain*, le support de la règle *bread -> cheese* est alors de 50 %. (Contrairement aux modèles d'association, la prise en charge n'est *pas* basée sur le nombre d'instances, comme mentionné précédemment.)
- L'option **Confiance** affiche le pourcentage des ID pour lesquels une prévision correcte est effectuée, par rapport à l'ensemble des ID pour lesquels la règle réalise une prévision. Elle est calculée de la manière suivante : nombre d'ID pour lesquels la séquence entière est trouvée, divisé par le nombre d'ID pour lesquels les antécédents sont trouvés, sur la base des données d'apprentissage. Par exemple, si 50 % des données d'apprentissage contiennent *cannedveg* (indiquant la prise en charge de l'antécédent), mais que seulement 20 % contiennent *cannedveg* et *frozenmeal*, alors la confiance de la règle *cannedveg -> frozenmeal* est $\text{Rule Support} / \text{Antecedent Support}$ ou, dans ce cas, 40 %.
- L'option **Prise en charge de la règle** des modèles Séquence repose sur les instances et affiche la proportion des enregistrements d'apprentissage pour lesquels l'intégralité de la règle, les antécédents et les conséquences sont vrais (true). Par exemple, si 20 % des données d'apprentissage comprennent *pain* et *fromage*, le support de la règle *bread -> cheese* est de 20 %.

Remarquez que les proportions sont basées sur des transactions valides (celles pour lesquelles au moins un article ou une valeur vraie a été observé) plutôt que sur toutes les transactions. Les transactions incorrectes (celles ne comportant aucun article ou aucune valeur vraie) sont ignorées dans ces calculs.

Bouton Filtrer. Le bouton Filtrer (icône d'entonnoir) du menu développe la partie inférieure de la boîte de dialogue, comportant un panneau où figurent les filtres de règle actifs. Les filtres servent à réduire le nombre de règles affichées dans l'onglet Modèles.

Figure 12-29
Bouton Filtrer

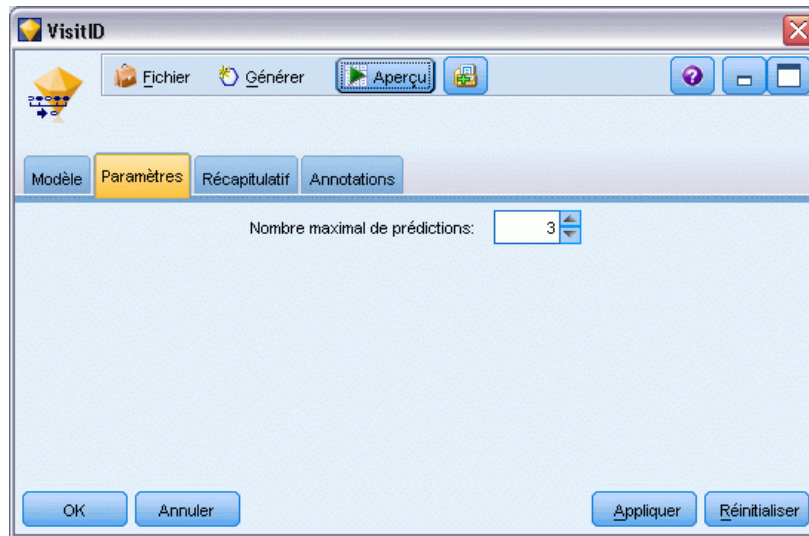


Pour créer un filtre, cliquez sur l'icône de filtre située à droite du panneau développé. La boîte de dialogue qui apparaît alors permet d'indiquer les contraintes d'affichage des règles. Notez que le bouton Filtrer est fréquemment utilisé avec le menu Générer pour filtrer les règles, puis créer un modèle comportant ce sous-ensemble de règles. Pour plus d'informations, reportez-vous à [Définition de filtres pour les règles](#) ci-dessous.

Paramètres du nugget de modèle de séquence

L'onglet Paramètres d'un nugget de modèle de séquence affiche les options de scoring du modèle. Il est disponible uniquement lorsque le modèle a été ajouté à l'espace de travail de flux à des fins de scoring.

Figure 12-30
Onglet Paramètres - Nugget de Séquence

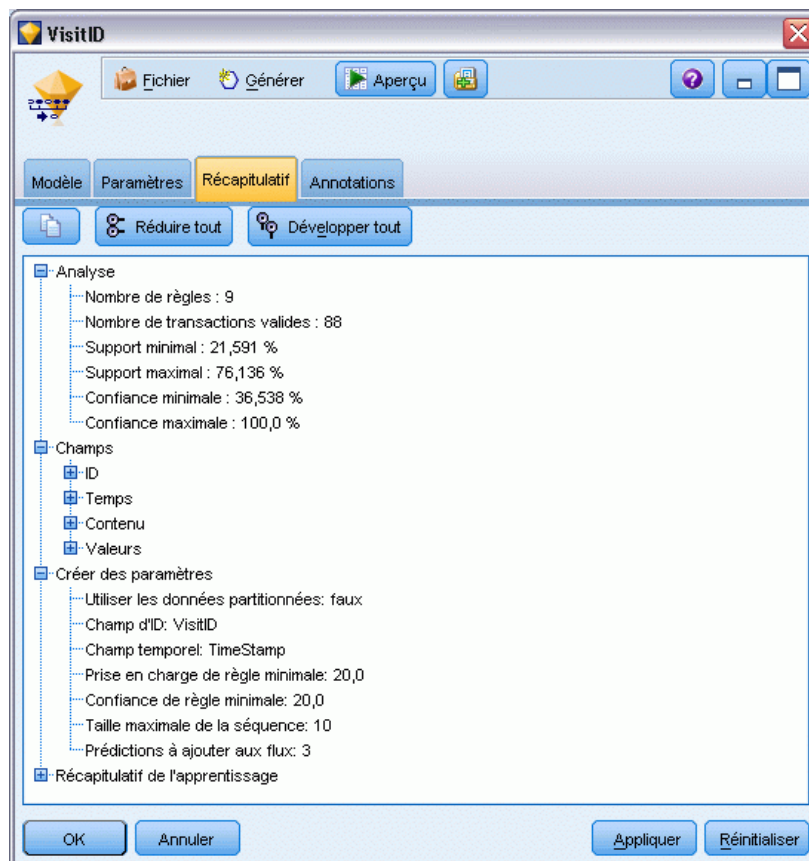


Nombre maximal de prédictions. Indiquez le nombre maximal de prévisions incluses pour chaque ensemble d'éléments de panier. Les règles présentant les valeurs de confiance les plus élevées et s'appliquant à cet ensemble de transactions sont utilisées pour générer les prévisions de l'enregistrement jusqu'à la limite indiquée.

Récapitulatif du nugget de modèle de séquence

L'onglet Récapitulatif d'un nugget de modèle Règle de séquence affiche le nombre de règles trouvées, ainsi que les valeurs minimale et maximale de prise en charge et de confiance des règles. Si vous avez exécuté un noeud Analyse relié à ce noeud de modélisation, 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*.](#)

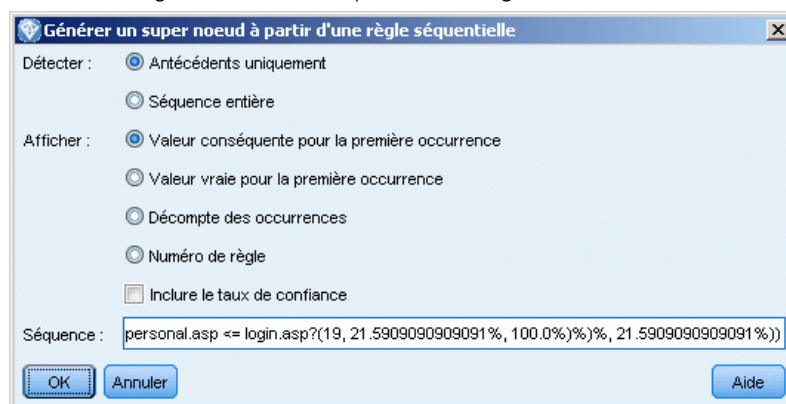
Figure 12-31
Onglet Récapitulatif - Nugget de Séquence



Pour plus d'informations, reportez-vous à la section Navigation dans les nuggets de modèle dans le chapitre 3 sur p. 55.

Génération d'un super noeud Règle à partir d'un nugget de modèle de séquence

Figure 12-32
Boîte de dialogue Générer un super noeud Règle



Pour générer un super noeud Règle sur la base d'une règle de séquence :

- ▶ Dans l'onglet Modèle du nugget de modèle de la règle de séquence, cliquez sur une ligne du tableau pour sélectionner la règle souhaitée.
- ▶ Dans les menus du navigateur de règles, choisissez :
Générer > Super noeud Règle

Important : Pour utiliser le super noeud généré, vous devez trier les données en fonction du champ ID (et du champ Temps le cas échéant) avant de les transmettre au super noeud. Le super noeud ne détecte pas correctement les séquences si les données ne sont pas triées.

Vous pouvez définir les options suivantes pour la génération d'un super noeud Règle :

Détecter. Spécifie la façon dont les correspondances sont définies pour les données transmises au super noeud.

- **Antécédents uniquement.** Le super noeud identifie une correspondance chaque fois qu'il trouve des antécédents pour la règle sélectionnée dans le bon ordre dans un ensemble d'enregistrements ayant le même ID, qu'une conséquence soit trouvée ou non. Cette opération ne prend pas en compte les paramètres de tolérance d'horodatage ou de contrainte d'intervalle entre les éléments du noeud de modélisation Séquence d'origine. Lorsque le dernier jeu d'éléments antécédent est détecté dans le flux (tous les autres antécédents ont été trouvés dans l'ordre qui convient), tous les enregistrements suivants comportant l'ID en cours contiennent le récapitulatif sélectionné ci-dessous.
- **Séquence entière.** Le super noeud identifie une correspondance chaque fois qu'il trouve les antécédents et les conséquences de la règle sélectionnée, dans le bon ordre, dans un ensemble d'enregistrements ayant le même ID. Cette opération ne prend pas en compte les paramètres de tolérance d'horodatage ou de contrainte d'intervalle entre les éléments du noeud de modélisation Séquence d'origine. Lorsque la conséquence est détectée dans le flux (et que tous les antécédents ont également été trouvés dans l'ordre qui convient), l'enregistrement en cours, ainsi que tous les enregistrements suivants ayant l'ID actuel contiendront le récapitulatif sélectionné ci-dessous.

Afficher : Contrôle la façon dont les récapitulatifs de correspondance sont ajoutés aux données dans la sortie du super noeud Règle.

- **Valeur conséquente pour la première occurrence.** La valeur ajoutée aux données correspond à la valeur conséquente prévue en fonction de la première occurrence de la correspondance. Les valeurs sont ajoutées en tant que nouveau champ appelé *rule_n_consequent*, *n* correspondant au numéro de la règle (basé sur l'ordre de création des super noeuds Règle dans le flux).
- **Valeur vraie pour la première occurrence.** La valeur ajoutée aux données est vraie s'il existe au moins une correspondance pour l'ID et fausse s'il n'en existe pas. Les valeurs sont ajoutées en tant que nouveau champ appelé *rule_n_flag*
- **Décompte des occurrences.** La valeur ajoutée aux données est le nombre de correspondances pour l'ID. Les valeurs sont ajoutées dans un nouveau champ appelé *rule_n_count*.
- **Numéro de règle.** La valeur ajoutée est le numéro de la règle sélectionnée. Les **numéros de règle** sont affectés en fonction de l'ordre dans lequel le super noeud a été ajouté au flux. Par exemple, le premier super noeud Règle est considéré comme étant *règle 1*, le deuxième super noeud Règle *règle 2*, etc. Cette option est particulièrement utile lorsque vous devez

inclure plusieurs super noeuds Règle dans votre flux. Les valeurs sont ajoutées en tant que nouveau champ appelé *rule_n_number*

- **Inclure le taux de confiance.** Lorsqu'elle est sélectionnée, cette option permet d'ajouter le degré de confiance de la règle au flux de données, ainsi que le récapitulatif sélectionné. Les valeurs sont ajoutées en tant que nouveau champ appelé *rule_n_confidence*.

Modèles de séries temporelles

Prévoir, à quoi ça sert ?

Prévoir signifie prédire les valeurs d'une ou de plusieurs séries dans le temps. Par exemple, vous pouvez prévoir la demande pour une gamme de produits ou de services de façon à allouer les ressources relatives à la fabrication ou à la distribution. Etant donné que la mise en oeuvre des décisions de planification est longue, les prévisions constituent un outil essentiel dans de nombreux processus de planification.

Les méthodes de modélisation des séries temporelles supposent que l'histoire se répète. Même s'il ne s'agit pas de répétitions exactement identiques, la similarité est suffisante pour permettre de prendre de meilleures décisions d'après l'étude du passé. Par exemple, pour prévoir les ventes de l'année prochaine, il est très probable que vous observiez d'abord les ventes de cette année, puis que vous remontiez le temps pour cerner les tendances ou schémas éventuels qui se sont développés au cours des dernières années. Mais il peut s'avérer difficile d'évaluer les schémas. Si vos ventes augmentent pendant plusieurs semaines d'affilée, par exemple, doit-on parler d'un cycle saisonnier ou du début d'une tendance à long terme ?

Grâce aux techniques de modélisation statistique, vous pouvez analyser les schémas des données passées et pratiquer une extrapolation afin de déterminer un intervalle au sein duquel les valeurs futures des séries sont susceptibles de se situer. Vous obtenez alors des prévisions plus précises sur lesquelles vous pouvez baser vos décisions.

Séries temporelles.

Une **série temporelle** est un ensemble ordonné de mesures prises à intervalles réguliers, par exemple, le cours journalier des valeurs boursières ou les données de ventes hebdomadaires. L'objet des mesures peut correspondre à un élément qui représente un intérêt pour vous et il est généralement possible de classer chaque série dans l'une des catégories suivantes :

- **Dépendante.** Série que vous souhaitez prévoir.
- **Valeur prédite.** Série qui peut contribuer à expliquer la cible (par exemple, utilisation d'un budget publicitaire pour prévoir les ventes). Il n'est possible d'utiliser les valeurs prédites qu'avec les modèles ARIMA.
- **Événement.** Série spéciale de valeurs prédites utilisée pour représenter les incidents récurrents prévisibles (par exemple, les promotions).
- **Intervention.** Série spéciale de valeurs prédites utilisée pour représenter les incidents passés à caractère exceptionnel (par exemple, une panne de courant ou une grève des employés).

Les intervalles peuvent représenter toute unité de temps, mais doivent être les mêmes pour toutes les mesures. De plus, tout intervalle pour lequel il n'existe pas de mesure doit être défini comme valeur manquante. Par conséquent, le nombre d'intervalles comportant des mesures (y compris ceux présentant des valeurs manquantes) définit la durée de l'étendue historique des données.

Caractéristiques des séries temporelles

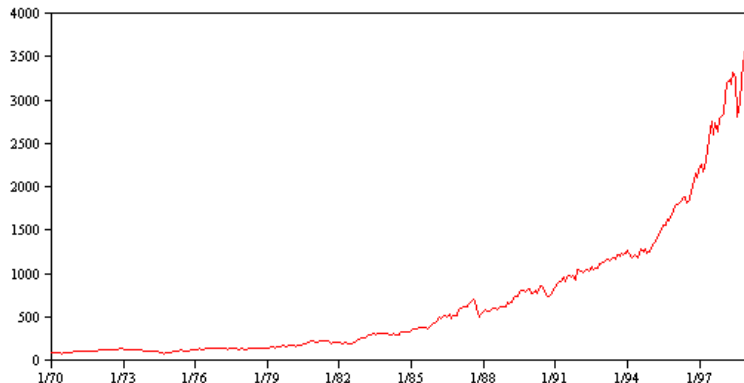
L'étude du comportement passé d'une série permet d'identifier des schémas et de réaliser de meilleures prévisions. Lorsqu'elles sont représentées graphiquement, de nombreuses séries temporelles affichent une ou plusieurs des caractéristiques suivantes :

- Tendances
- Cycles saisonniers et non saisonniers
- Impulsions et étapes
- Valeurs éloignées

Tendances

Une **tendance** est un changement ascendant ou descendant progressif du niveau de la série, ou encore la tendance des valeurs de la série à augmenter ou à diminuer dans le temps.

Figure 13-1
Tendance



Les tendances sont soit **locales**, soit **globales**, mais une série unique peut faire état des deux types. Du point de vue historique, les représentations graphiques des séries de l'indice boursier montrent une tendance globale ascendante. Les tendances locales descendantes apparaissent en période de récession et les tendances locales ascendantes, en période de croissance.

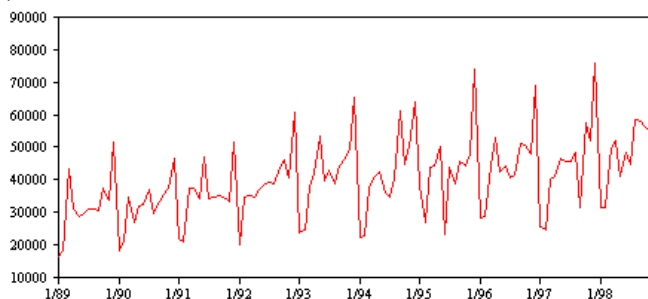
Les tendances peuvent également être **linéaires** ou **non linéaires**. Les tendances linéaires sont des incréments additifs positifs ou négatifs qui s'appliquent au niveau de la série, comparables à l'effet des intérêts simples sur le capital. Les tendances non linéaires sont souvent multiplicatives, avec des incréments proportionnels aux valeurs des séries précédentes.

Les modèles de lissage exponentiel et les modèles ARIMA sont adaptés aux tendances linéaires globales et prévoient ces dernières de façon correcte. Lors de la création de modèles ARIMA, les séries qui présentent des tendances sont généralement différenciées afin de supprimer l'effet de la tendance.

Cycles saisonniers

Un **cycle saisonnier** est un schéma répétitif et prévisible qui caractérise les valeurs d'une série.

Figure 13-2
Cycle saisonnier



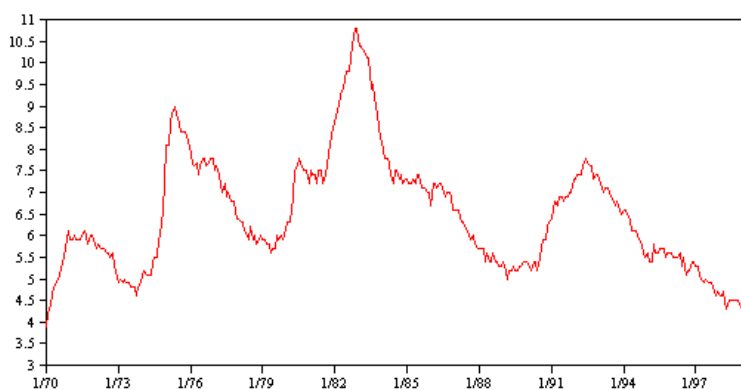
Les cycles saisonniers sont liés à l'intervalle des séries. Par exemple, des données mensuelles constituent généralement des cycles sur des trimestres et des années. Une série mensuelle peut faire état d'un cycle trimestriel significatif avec un niveau bas au premier trimestre ou encore d'un cycle annuel avec un pic chaque mois de décembre. Les séries faisant état d'un cycle saisonnier sont considérées comme présentant des **effets saisonniers**.

Les schémas saisonniers sont utiles pour obtenir de bons ajustements et de bonnes prévisions ; il existe des modèles de lissage exponentiel et ARIMA qui capturent les effets saisonniers.

Cycles non saisonniers

Un **cycle non saisonnier** est un schéma répétitif, éventuellement imprévisible, qui caractérise les valeurs d'une série.

Figure 13-3
Cycle non saisonnier



Certaines séries, comme le taux de chômage, présentent clairement un comportement cyclique ; toutefois, la périodicité du cycle varie dans le temps, ce qui rend difficile la prévision d'un pic ou d'un niveau bas. D'autres séries peuvent présenter des cycles prévisibles, mais ne correspondent pas exactement au calendrier grégorien ou intègrent des cycles plus longs qu'une année. Par exemple, les marées suivent le calendrier lunaire, les déplacements à l'étranger et le commerce international liés aux Jeux olympiques augmentent tous les quatre ans et il existe un grand nombre de fêtes religieuses dont les dates grégoriennes changent d'une année à l'autre.

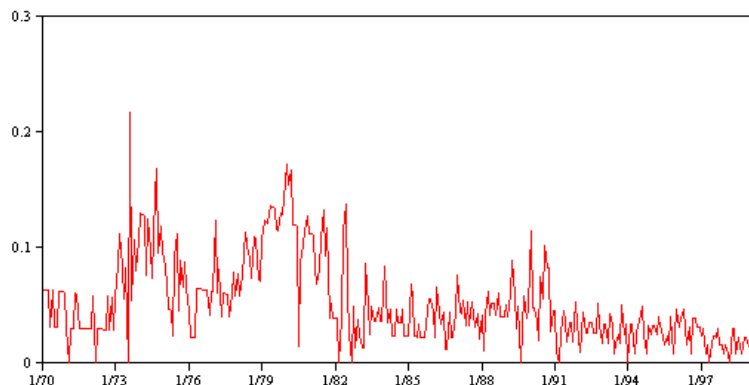
Les schémas cycliques non saisonniers sont difficiles à modéliser et augmentent généralement l'incertitude des prévisions. La Bourse, par exemple, fournit de nombreux exemples de séries qui ont résisté aux efforts des prévisionnistes. Toutefois, les schémas non saisonniers doivent être justifiés lorsqu'ils existent. Dans de nombreux cas, vous pouvez toujours identifier un modèle qui s'ajuste aux données historiques de façon relativement correcte ; vous avez ainsi de grandes chances de minimiser l'incertitude des prévisions.

Impulsions et étapes

De nombreuses séries font état de variations de niveau soudaines. Elles sont généralement de deux types :

- Un changement soudain et *temporaire*, ou **impulsion**, dans le niveau de la série ;
- Un changement soudain et *permanent*, ou **étape**, dans le niveau de la série.

Figure 13-4
Série avec impulsion



Si vous observez des étapes ou des impulsions, il est important de trouver une explication plausible. Les modèles de séries temporelles sont conçus pour représenter les changements progressifs, et non soudains. Par conséquent, ils tendent à sous-estimer les impulsions et à être altérés par les étapes, ce qui entraîne de médiocres ajustements de modèle et des prévisions incertaines. (Il se peut que certains exemples d'effets saisonniers montrent des changements soudains dans le niveau, mais celui-ci est constant d'une période saisonnière à l'autre.)

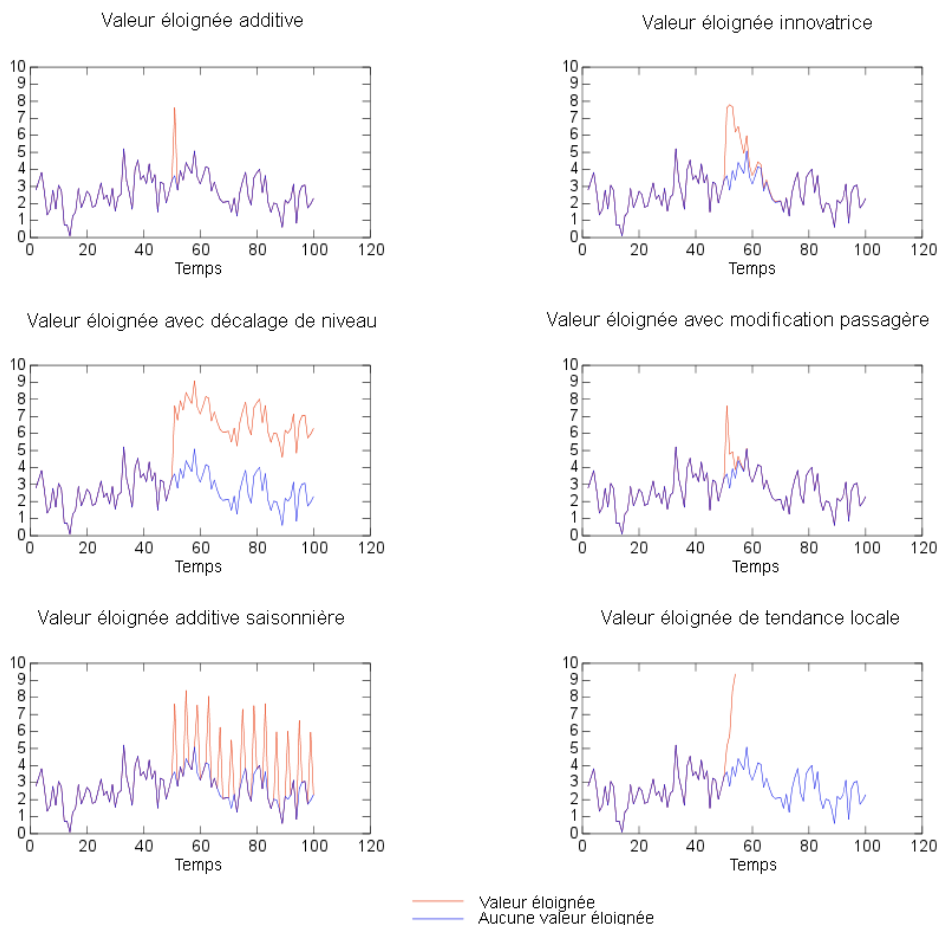
S'il est possible d'expliquer une perturbation, vous pouvez procéder à une modélisation à l'aide d'une **intervention** ou d'un **événement**. Par exemple, en août 1973, un embargo imposé par l'OPEP (Organisation des pays exportateurs de pétrole) sur le pétrole a généré une variation radicale du taux d'inflation, qui est revenu à un niveau normal dans les mois suivants. En spécifiant une **intervention ponctuelle** pour le mois de l'embargo, vous pouvez améliorer l'ajustement du modèle et, par conséquent, améliorer indirectement les prévisions. Par exemple, un commerce de détail peut découvrir que les ventes étaient largement supérieures le jour où les articles faisaient l'objet d'une réduction de 50 %. En spécifiant la promotion de - 50 % comme **événement** récurrent, vous pouvez améliorer l'ajustement du modèle et estimer l'effet de la répétition de la promotion sur des dates futures.

Valeurs éloignées

Les changements dans le niveau d'une série temporelle ne pouvant pas être expliqués sont nommés **valeurs éloignées**. Ces observations sont incohérentes par rapport au reste de la série et peuvent considérablement influencer l'analyse et, par conséquent, affecter la capacité de prévision du modèle de série temporelle.

La figure suivante présente plusieurs types de valeurs éloignées figurant couramment dans les séries temporelles. Les lignes bleues représentent une série sans valeurs éloignées. Les lignes rouges suggèrent un schéma pouvant être présent si les séries contiennent des valeurs éloignées. Ces valeurs éloignées sont toutes classées comme étant **déterministes** car elles n'affectent que le niveau moyen des séries.

Figure 13-5
Valeurs éloignées types



- **Valeur éloignée additive.** Une valeur éloignée additive apparaît comme une valeur étonnamment élevée ou faible rencontrée pour une observation unique. Les observations suivantes ne sont pas affectées par une valeur éloignée additive. Les valeurs éloignées additives sont généralement nommées **correctifs de valeur éloignée additive**.

- **Valeur éloignée innovatrice.** Une valeur éloignée innovatrice est caractérisée par un impact initial dont les effets subsistent sur des observations ultérieures. L'influence des valeurs éloignées peut augmenter avec le temps.
- **Valeur éloignée avec décalage de niveau.** Pour un décalage de niveau, toutes les observations apparaissant après la valeur éloignée passent à un nouveau niveau. Par rapport aux valeurs éloignées additives, une valeur éloignée avec décalage de niveau affecte de nombreuses observations et a un effet permanent.
- **Valeur éloignée avec modification passagère.** Les valeurs éloignées avec modification passagère sont semblables aux valeurs éloignées avec décalage de niveau, mais l'effet de la valeur éloignée diminue de manière exponentielle sur les observations ultérieures. La série finit par retourner à son niveau normal.
- **Valeur éloignée additive saisonnière.** Une valeur éloignée additive saisonnière apparaît comme une valeur étonnamment élevée ou faible rencontrée de manière répétée à intervalles réguliers.
- **Valeur éloignée de tendance locale** Une valeur éloignée de tendance locale génère une déviation générale dans la série causée par un schéma dans les valeurs éloignées après le début de la valeur éloignée initiale.

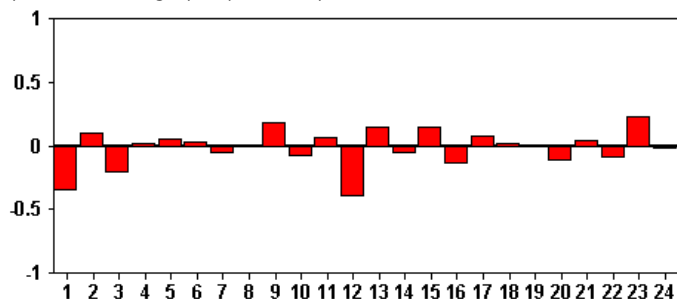
La détection des valeurs éloignées dans les séries temporelles implique de déterminer l'emplacement, le type et la grandeur des valeurs éloignées présentes. Tsay (1988) a proposé une procédure itérative pour détecter les changements de niveau moyen afin d'identifier les valeurs éloignées déterministes. Ce processus implique de comparer un modèle de série temporelle qui suppose qu'aucune valeur éloignée n'est présente dans un autre modèle qui incorpore des valeurs éloignées. Les différences entre les modèles génèrent des estimations de l'effet du traitement d'un point en tant que valeur éloignée.

Fonctions d'autocorrélation et d'autocorrélation partielle

L'autocorrélation et l'autocorrélation partielle sont des mesures de l'association entre des valeurs de séries actuelles et passées ; elles indiquent les valeurs de séries passées les plus utiles à la prévision de valeurs futures. Avec ces données, vous pouvez déterminer l'ordre des processus d'un modèle ARIMA. De façon plus spécifique :

- **Fonction d'autocorrélation (ACF).** Au décalage k , il s'agit de la corrélation entre les valeurs de séries séparées par k intervalles.
- **Fonction d'autocorrélation partielle (PACF).** Au décalage k , il s'agit de la corrélation entre les valeurs de séries séparées par k intervalles, compte tenu des valeurs des intervalles intermédiaires.

Figure 13-6
Représentation graphique ACF pour une série



L'axe x de la représentation graphique ACF indique le décalage auquel l'autocorrélation est calculée ; l'axe y indique la valeur de la corrélation (entre -1 et 1). Par exemple, une pointe au décalage 1 dans une représentation graphique ACF indique une forte corrélation entre chaque valeur de série et la valeur précédente ; une pointe au décalage 2 indique une forte corrélation entre chaque valeur et la valeur apparaissant deux points auparavant, etc.

- Une corrélation positive indique que des valeurs actuelles élevées correspondent à des valeurs élevées au niveau du décalage spécifié ; une corrélation négative indique que des valeurs actuelles élevées correspondent à des valeurs faibles au niveau du décalage spécifié.
- La valeur absolue d'une corrélation est une mesure de la force de l'association, des valeurs absolues élevées indiquant des relations plus fortes.

Transformations de série

Les transformations sont souvent utiles pour stabiliser une série avant d'estimer des modèles. Ceci est particulièrement important pour les modèles ARIMA, qui requièrent que les séries soient **stationnaires** avant l'estimation des modèles. Une série est stationnaire si le niveau global (moyenne) et l'écart moyen à partir du niveau (variance) sont constants dans l'ensemble de la série.

Bien que les séries les plus intéressantes ne soient pas stationnaires, le modèle ARIMA est efficace tant que la série peut être rendue stationnaire par l'application de transformations, comme le log naturel, la différenciation ou la différenciation saisonnière.

Transformations de stabilisation de variance. Il est souvent possible de stabiliser les séries dans lesquelles la variance évolue dans le temps à l'aide d'une transformation par le logarithme naturel ou par la racine carrée. Ces transformations sont également appelées transformations fonctionnelles.

- **Log népérien :** Le logarithme naturel est appliqué aux valeurs de la série.
- **Racine carrée :** La fonction racine carrée est appliquée aux valeurs de la série.

Il n'est pas possible d'utiliser les transformations par le logarithme naturel et par la racine carrée pour les séries comportant des valeurs négatives.

Transformations de stabilisation de niveau. Une baisse lente des valeurs de la fonction d'autocorrélation indique que chaque valeur de la série est fortement corrélée à la valeur précédente. En analysant la variation des valeurs de la série, vous obtenez un niveau stable.

- **Différenciation simple.** Les différences entre chaque valeur et la valeur précédente de la série sont calculées, à l'exception, évidemment, de la valeur la plus ancienne. En d'autres termes, la série différenciée inclura une valeur de moins que la série d'origine.
- **Différenciation saisonnière.** Différenciation identique à la différenciation simple, excepté que le calcul porte sur les différences entre chaque valeur et la valeur saisonnière précédente.

Lorsque vous utilisez la différenciation simple ou la différenciation saisonnière simultanément avec la transformation par le logarithme naturel ou par la racine carrée, la transformation de type Stabilisation de variance est toujours appliquée en premier lieu. Lorsque vous utilisez à la fois la différenciation simple et la différenciation saisonnière, les valeurs de série obtenues sont les mêmes, quelle que soit la différenciation appliquée en premier lieu.

Série de valeurs prédites

Les séries de valeurs prédites contiennent des données connexes qui peuvent contribuer à expliquer le comportement des séries sur lesquelles portera la prévision. Une société de vente sur le Web ou par correspondance peut prévoir ses ventes par rapport au nombre de catalogues envoyés, de lignes téléphoniques ouvertes ou d'accès à sa page Web.

N'importe quelle série peut être utilisée comme valeur prédite, dès l'instant que sa durée correspond à la durée que voulez prédire et que ses données sont complètes, sans valeurs manquantes.

Faites preuve de prudence lorsque vous ajoutez des valeurs prédites à un modèle. L'ajout d'un grand nombre de prédicteurs augmente le temps nécessaire à l'estimation des modèles. Si l'ajout de prédicteurs peut améliorer la capacité du modèle à s'ajuster aux données historiques, cela ne signifie pas pour autant qu'il effectue de meilleures prévisions. Par conséquent, vu la difficulté supplémentaire, le jeu n'en vaut peut-être pas la chandelle. Idéalement, l'objectif devrait être d'identifier le modèle le plus simple pour effectuer de bonnes prévisions.

En général, il est recommandé d'utiliser un nombre de valeurs prédites inférieur à la taille de l'échantillon divisée par 15 (au maximum, une valeur prédite pour 15 observations).

Valeurs prédites contenant des données manquantes. Les valeurs prédites contenant des données incomplètes ou manquantes ne peuvent pas être utilisées pour la prévision. Cela s'applique à la fois aux données historiques et aux valeurs futures. Dans certains cas, il est possible de contourner cet obstacle en définissant l'amplitude d'estimation du modèle de façon à exclure les données les plus anciennes lors de l'estimation des modèles.

Noeud de modélisation Séries temporelles

Le noeud Séries temporelles estime les modèles de lissage exponentiel, d'ARIMA (Autoregressive Integrated Moving Average) univariable et d'ARIMA multivariable (ou fonction de transfert) pour les séries temporelles, et il génère des prévisions basées sur les séries temporelles.

Le **lissage exponentiel** est une méthode qui vise à prévoir des valeurs futures à partir de valeurs pondérées d'observations de séries antérieures. De part sa nature, le lissage exponentiel n'est pas basé sur une compréhension théorique des données. Il prévoit un point à la fois et ajuste ses prévisions en fonction des nouvelles données obtenues. Cette technique permet de prévoir les séries dont se dégagent une tendance et/ou des effets saisonniers. Vous pouvez choisir parmi une variété de modèles exponentiels qui diffèrent dans leur manière de traiter la tendance et la saison.

Les modèles **ARIMA** mettent à votre disposition des méthodes de modélisation des composantes de tendance et saisonnières plus élaborées que celles proposées par les modèles de lissage exponentiel. En particulier, ils permettent d'inclure des variables indépendantes (valeurs prédites) dans le modèle. Cela suppose la spécification explicite des ordres autorégressifs et de moyennes mobiles, ainsi que du degré de différenciation. Vous pouvez inclure des variables de prédicteur et définir des fonctions de transfert pour certaines ou pour toutes, et spécifier la détection automatique des valeurs éloignées ou d'un ensemble explicite de valeurs éloignées.

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. Par exemple, une série, dont l'historique montre qu'elle a connu un pic tous les 12 mois, continuera probablement à se comporter ainsi, même si vous ignorez la raison de ce cycle.

Vous disposez également d'un **modélisateur expert**, qui identifie et estime automatiquement le modèle ARIMA ou de lissage exponentiel le mieux adapté pour une ou plusieurs variables cible ; cela évite d'identifier un modèle approprié via une série d'essais et d'erreurs. Dans tous les cas, le modélisateur expert choisit le meilleur modèle pour chacune des variables cible spécifiées. En cas de doute, utilisez le modélisateur expert.

Si des variables de prévision sont spécifiées, le modélisateur expert sélectionne celles qui présentent une relation significative d'un point de vue statistique avec la série dépendante afin de les inclure dans les modèles ARIMA. Les variables de modèle sont transformées le cas échéant à l'aide de transformation par différenciation, par racine carrée ou par log népérien. Par défaut, le modélisateur expert prend en considération tous les modèles de lissage exponentiel et tous les modèles ARIMA, et sélectionne, parmi ceux-ci, le meilleur modèle pour chaque champ cible. Toutefois, vous pouvez faire en sorte que le modélisateur expert ne sélectionne que le meilleur des modèles de lissage exponentiel ou que le meilleur des modèles ARIMA. Vous pouvez également indiquer la détection automatique des valeurs éloignées.

Exemple : Un analyste pour un fournisseur large bande national doit établir des prévisions sur les abonnements des utilisateurs afin de prédire l'utilisation de la bande passante. Les prévisions sont requises pour chacun des marchés locaux qui constitue la base nationale de l'abonné. Vous pouvez recourir à la modélisation des séries temporelles afin de produire des prévisions sur un sous-ensemble des marchés locaux pour les trois prochains mois. [Pour plus d'informations, reportez-vous à la section Prévision avec le noeud Séries temporelles dans le chapitre 14 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Conditions requises

Le noeud Séries temporelles diffère des autres noeuds IBM® SPSS® Modeler, en ce sens que vous ne pouvez pas simplement l'inclure dans un flux, puis exécuter celui-ci. Le noeud Séries temporelles doit toujours être précédé d'un noeud Intervalle de temps qui spécifie des informations telles que l'intervalle de temps à utiliser (années, trimestres, mois, etc.), les données à utiliser pour l'estimation et, le cas échéant, la période sur laquelle doit porter une prévision.

Figure 13-7

Faites toujours précéder un noeud Séries temporelles d'un noeud Intervalle de temps



Les séries temporelles doivent être espacées de manière égale. Les méthodes de modélisation des séries temporelles nécessitent l'utilisation d'un intervalle uniforme entre chaque mesure, chaque valeur manquante étant signalée par des lignes vides. Si les données ne répondent pas à ces exigences, le noeud Intervalle de temps peut transformer les valeurs si nécessaire. [Pour plus d'informations, reportez-vous à la section Noeud Intervalles de temps dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

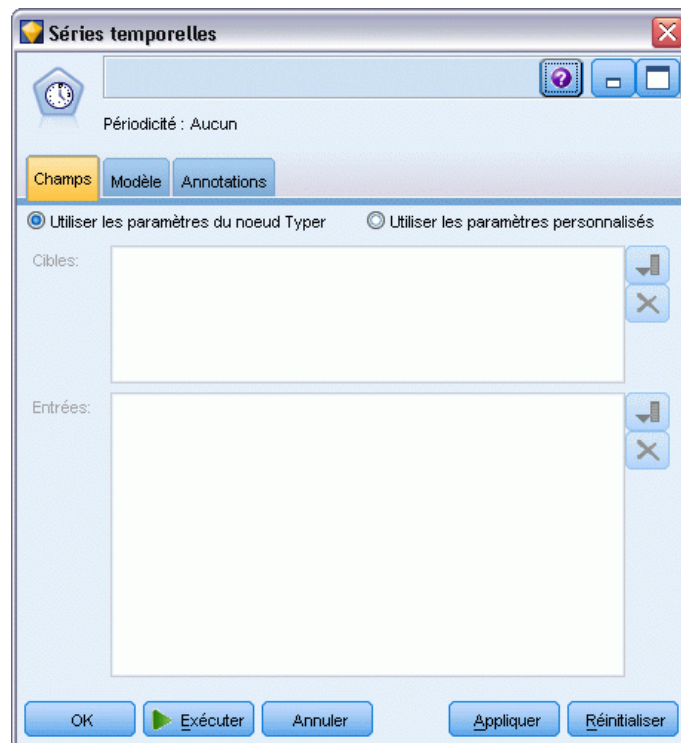
Autres points à noter en relation avec les noeuds Séries temporelles :

- Les champs doivent être numériques
- Les champs de date ne peuvent pas être utilisés comme entrées
- Les partitions sont ignorées

Options de champs

Figure 13-8

Boîte de dialogue du noeud Séries temporelles - Onglet Champs



L'onglet Champs vous permet de spécifier les champs à utiliser lors de la création du modèle. Avant de construire un modèle, vous devez indiquer les champs à utiliser en tant que cibles et en tant qu'entrées. En règle générale, le noeud Séries temporelles utilise les informations du champ à partir d'un noeud Typer en amont. Si vous utilisez un noeud Typer pour sélectionner les champs d'entrée et les champs cible, vous n'avez aucune modification à apporter dans cet onglet.

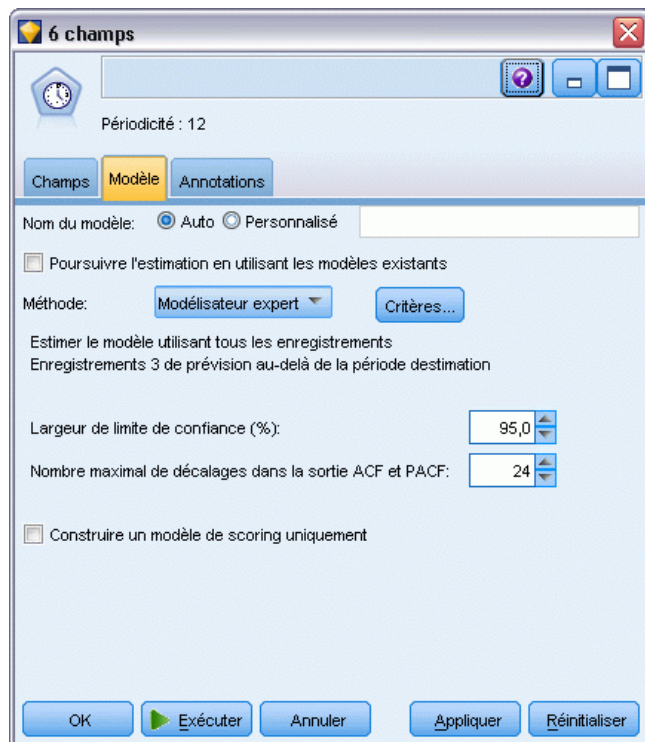
Utiliser les paramètres du noeud Typer. Cette option indique au noeud d'utiliser les informations du champ à partir d'un noeud Typer en amont. Il s'agit de la valeur par défaut.

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. Une fois cette option sélectionnée, renseignez les champs ci-dessous. Les champs stockés sous forme de dates ne sont pas acceptés en tant que champs cible ou d'entrée.

- **Cibles.** Sélectionnez un ou plusieurs champs cible. Cela revient à définir le rôle du champ sur la valeur *Cible* dans un noeud Typer. Les champs cible pour un modèle de séries temporelles ont un niveau de mesure *Continu*. Un modèle distinct est créé pour chaque champ cible. Un champ cible peut avoir comme entrée tous les champs d'*entrée* spécifiés, à l'exception de lui-même. Par conséquent, le même champ peut être inclus dans les deux listes ; un champ de ce type sera utilisé comme entrée possible pour tous les modèles, sauf pour celui où il fait office de cible.
- **Entrées.** Sélectionnez le ou les champs d'entrée. Cela revient à définir le rôle du champ sur la valeur *Entrée* dans un noeud Typer. Les champs d'entrée pour un modèle de séries temporelles doivent être numériques.

Options du modèle de séries temporelles

Figure 13-9
Boîte de dialogue du noeud Séries temporelles - Onglet Modèle



Nom du modèle. Indique le nom attribué au modèle généré 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 d'indiquer un nom personnalisé pour le nugget de modèle.

Poursuivre l'estimation à l'aide d'un ou de plusieurs modèles existants. Si vous avez déjà généré un modèle de séries temporelles, sélectionnez cette option afin de réutiliser l'ensemble des critères spécifié pour ce modèle et de générer un nouveau noeud de modèle dans la palette Modèles, plutôt que de créer totalement un nouveau modèle. Ainsi, vous pouvez gagner du temps en effectuant une nouvelle estimation et en générant une nouvelle prévision à partir des mêmes paramètres de modèle que ceux utilisés auparavant, mais en recourant à des données plus récentes. Par conséquent, par exemple, si le modèle d'origine pour une série temporelle spécifique est la tendance linéaire de Holt, le même type de modèle est utilisé pour effectuer une nouvelle estimation et une prévision portant sur ces données ; le système n'essaie pas de rechercher de nouveau le meilleur type de modèle pour les nouvelles données. La sélection de cette option désactive les commandes Méthode et Critères. [Pour plus d'informations, reportez-vous à la section Nouvelle estimation et prévision sur p. 480.](#)

Méthode : Vous pouvez choisir l'option Modélisateur expert, Lissage exponentiel ou ARIMA. [Pour plus d'informations, reportez-vous à la section Noeud de modélisation Séries temporelles sur p. 465.](#) Sélectionnez l'option Critères pour spécifier les options de la méthode choisie.

- **Modélisateur expert** Sélectionnez cette option pour utiliser le modélisateur expert, qui recherche automatiquement le modèle le mieux adapté pour chaque série dépendante.
- **Lissage exponentiel.** Utilisez cette option pour indiquer un modèle de lissage exponentiel personnalisé.
- **ARIMA.** Utilisez cette option pour indiquer un modèle ARIMA personnalisé.

Informations sur les intervalles de temps

Cette section de la boîte de dialogue contient des informations sur la spécification des estimations et des prévisions effectuées sur le noeud Intervalle de temps. Notez que cette section n'apparaît pas si vous choisissez l'option Poursuivre l'estimation à l'aide d'un ou de plusieurs modèles existants.

La première ligne d'information indique si des enregistrements sont exclus du modèle ou utilisés comme ensemble de rétention. [Pour plus d'informations, reportez-vous à la section Période d'estimation dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

La seconde ligne fournit des informations sur les périodes de prévision spécifiées sur le noeud Intervalle de temps. [Pour plus d'informations, reportez-vous à la section Prévisions dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Si la première ligne indique Aucun intervalle de temps défini, cela signifie qu'aucun noeud Intervalle de temps n'est connecté. Cette situation peut faire échouer l'exécution du flux ; vous devez inclure un noeud Intervalle de temps en amont du noeud Séries temporelles.

Informations diverses

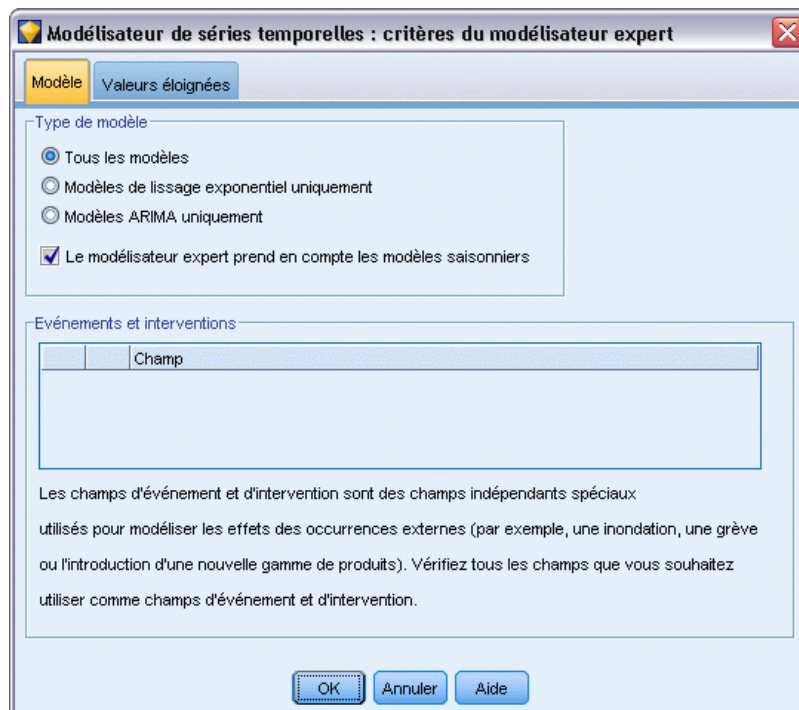
Largeur de limite de confiance (%). Les intervalles de confiance sont calculés pour les prévisions et les autocorrélations résiduelles du modèle. Vous pouvez indiquer toute valeur positive inférieure à 100. Par défaut, un intervalle de confiance de 95 % est utilisé.

Nombre maximal de décalages dans la sortie ACF et PACF. Vous pouvez définir le nombre maximum de décalages affichés dans les tableaux et diagrammes d'autocorrélations et d'autocorrélations partielles.

Construire un modèle de scoring uniquement. Cochez cette case pour réduire la quantité de données stockées dans le modèle. Cela peut améliorer les performances lors de la construction de modèles avec de grands nombres de séries temporelles (dizaines de milliers). Si vous sélectionnez cette option, les onglets Modèles, Paramètres et Résidus n'apparaissent pas dans le nugget de modèle Séries temporelles mais vous pouvez quand même évaluer les données de manière habituelle.

Critères du modélisateur expert de séries temporelles

Figure 13-10
Boîte de dialogue Critères du modélisateur expert, onglet Modèle



Type de modèle. Les options suivantes sont disponibles :

- **Tous les modèles.** Le modélisateur expert considère aussi bien les modèles ARIMA que ceux de lissage exponentiel.
- **Modèles de lissage exponentiel uniquement.** Le modélisateur expert considère uniquement les modèles de lissage exponentiel.
- **Modèles ARIMA uniquement.** Le modélisateur expert considère uniquement les modèles ARIMA.

Le modélisateur expert considère les modèles saisonniers. Cette option est uniquement activée si une périodicité a été définie pour l'ensemble de données actif. Lorsque cette option est sélectionnée, le modélisateur expert prend en considération les modèles saisonniers et non saisonniers. Si cette option n'est pas sélectionnée, le modélisateur expert considère uniquement les modèles non saisonniers.

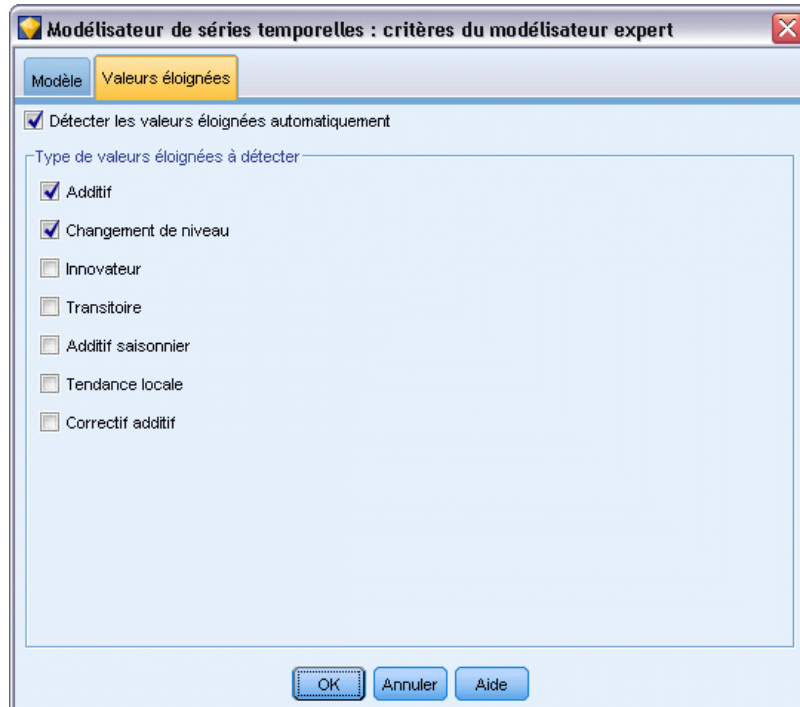
Événements et interventions. Permet de désigner certains champs d'entrée en tant que champs d'événement ou d'intervention. Cela permet d'indiquer qu'un champ contient des séries temporelles affectées par des événements (situations récurrentes prévisibles, telles que les campagnes publicitaires) ou des interventions (incidents occasionnels, tels qu'une coupure de courant ou une grève d'employés). Le modélisateur expert prend uniquement en considération les fonctions de régression simple, et non les fonctions de transfert arbitraire, pour les entrées identifiées en tant que champs d'événement ou d'intervention.

Les champs d'entrée doit avoir un niveau de mesure *Booléen*, *Nominal*, ou *Ordinal* et doivent être numériques (par exemple, 1/0, pas Vrai/Faux, pour un champ booléen) avant qu'ils soient inclus dans cette liste. [Pour plus d'informations, reportez-vous à la section Impulsions et étapes sur p. 461.](#)

Valeurs éloignées

Figure 13-11

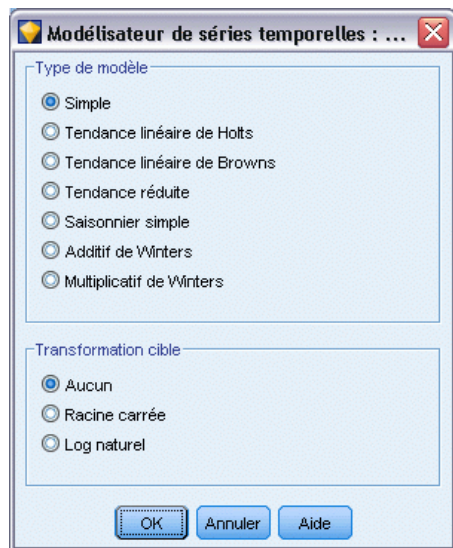
Boîte de dialogue Critères du modélisateur expert, onglet Valeurs éloignées



Détecter les valeurs éloignées automatiquement. Par défaut, la détection automatique de valeurs éloignées n'est pas exécutée. Sélectionnez cette option pour procéder à la détection automatique des valeurs éloignées, puis sélectionnez les types de valeur éloignée qui vous intéressent. [Pour plus d'informations, reportez-vous à la section Valeurs éloignées sur p. 462.](#)

Critères du lissage exponentiel des séries temporelles

Figure 13-12
Boîte de dialogue Critère de lissage exponentiel



Type de modèle. Les modèles de lissage exponentiel sont de type soit saisonnier, soit non saisonnier. Les modèles saisonniers ne sont disponibles que si la périodicité définie à l'aide du noeud Intervalle de temps est saisonnière. Les périodicités saisonnières sont les suivantes : périodes cycliques, années, trimestres, mois, jours par semaine, heures par jour, minutes par jour et secondes par jour. [Pour plus d'informations, reportez-vous à la section Noeud Intervalles de temps dans le chapitre 4 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

- **Simple :** Ce modèle est approprié aux séries dont ne se dégage aucune tendance ou effet saisonnier. Son seul paramètre de lissage pertinent est le niveau. Le lissage exponentiel simple ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, un ordre de différenciation, un ordre de moyenne mobile et aucune constante.
- **Tendance linéaire de Holt.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau et la tendance. Dans ce modèle, la valeur d'un paramètre n'est pas tributaire de celle d'un autre paramètre. Le modèle de Holt est plus général que le modèle de Brown, mais il prend plus de temps pour calculer les estimations pour les grandes séries. Le lissage exponentiel de Holt ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, deux ordres de différenciation et deux ordres de moyenne mobile.
- **Tendance linéaire de Brown.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau et la tendance, mais, dans ce modèle, ils sont supposés égaux. Par conséquent, le modèle de Brown est un cas particulier du modèle de Holt. Le lissage exponentiel de Brown ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, deux ordres de différenciation et deux ordres de moyenne mobile, le coefficient du deuxième ordre de moyenne mobile étant égal à la moitié du coefficient du premier ordre élevé au carré.

- **Tendance amortie.** Ce modèle est approprié aux séries dont se dégage une tendance linéaire qui s'éteint, mais aucun effet saisonnier. Ses paramètres de lissage pertinents sont le niveau, la tendance et la tendance avec amortissement. Le lissage exponentiel amorti ressemble sensiblement à une ARIMA avec un ordre d'autorégression, un ordre de différenciation et deux ordres de moyenne mobile.
- **Saisonnier simple.** Ce modèle est approprié aux séries ne présentant pas de tendance et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau et la saison. Le lissage exponentiel saisonnier ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, un ordre de différenciation, un ordre de différenciation saisonnier, ainsi que 1, p et $p+1$ ordres de moyenne mobile, où p représente le nombre de périodes dans un intervalle saisonnier. Pour les données mensuelles, p a pour valeur 12.
- **Modèle additif de Winters.** Ce modèle est approprié aux séries présentant une tendance linéaire et un effet saisonnier constant dans le temps. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. Le lissage exponentiel additif de Winters ressemble sensiblement à une ARIMA avec aucun ordre d'autorégression, un ordre de différenciation, un ordre de différenciation saisonnière et $p+1$ ordres de moyenne mobile, où p représente le nombre de périodes dans un intervalle saisonnier. Pour les données mensuelles, p a pour valeur 12.
- **Modèle multiplicatif de Winters.** Ce modèle est approprié aux séries qui présentent une tendance linéaire et un effet saisonnier qui varie en fonction de la valeur des séries. Ses paramètres de lissage pertinents sont le niveau, la tendance et la saison. Le lissage exponentiel multiplicatif de Winter n'est pas similaire à un modèle ARIMA.

Transformation cible. Vous pouvez indiquer qu'une transformation doit être effectuée sur chaque variable dépendante avant que celle-ci ne soit modélisée. [Pour plus d'informations, reportez-vous à la section Transformations de série sur p. 464.](#)

- **Aucune.** Aucune transformation n'est effectuée.
- **Racine carrée :** Une transformation par la racine carrée est effectuée.
- **Log népérien :** Une transformation par le logarithme naturel est effectuée.

Critères ARIMA pour les séries temporelles

Le noeud Séries temporelles vous permet de créer des modèles ARIMA saisonniers ou non saisonniers personnalisés (également appelés modèles de Box-Jenkins) avec ou sans le recours à un ensemble fixe de variables d'entrée (valeur prédite). Vous pouvez définir des fonctions de transfert pour la totalité ou une partie des variables d'entrée et spécifier la détection automatique des valeurs éloignées ou d'un ensemble explicite de valeurs éloignées.

Toutes les variables d'entrée spécifiées sont explicitement incluses dans le modèle. Il en va différemment lorsque vous utilisez le modélisateur expert, car les variables d'entrée ne sont incluses que si elles ont une relation significative d'un point de vue statistique avec la variable cible.

Modèle

L'onglet Modèle vous permet de spécifier la structure d'un modèle ARIMA personnalisé.

Figure 13-13
Boîte de dialogue Critères ARIMA, onglet Modèle

Modélisateur de séries temporelles : critères ARIMA

Modèle Fonctions de transfert Valeurs éloignées

Ordres Arima

Structure:

	Non saisonnier	Saisonnier
Autorégressif (p)	0	0
Différence (d)	0	0
Moyenne mobile (q)	0	0

Transformation cible

Aucun
 Racine carrée
 Log naturel

Inclure la constante dans le modèle

OK Annuler Aide

Ordres ARIMA. Entrez des valeurs pour les différentes composantes ARIMA de votre modèle dans les cellules correspondantes de la grille Structure. Toutes les valeurs doivent être des entiers non négatifs. Pour les composants autorégressifs et de moyenne mobile, la valeur représente l'ordre maximum. Tous les ordres inférieurs positifs seront inclus dans le modèle. Par exemple, si vous spécifiez 2, le modèle comprend les ordres 2 et 1. Les cellules de la colonne Saisonnier ne sont activées que si une périodicité a été définie pour l'ensemble de données actif.

- **Autorégressif (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.
- **Différence (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.

Ordres saisonniers. Les composants autorégressifs, de moyenne mobile et de différenciation saisonniers tiennent le même rôle 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.

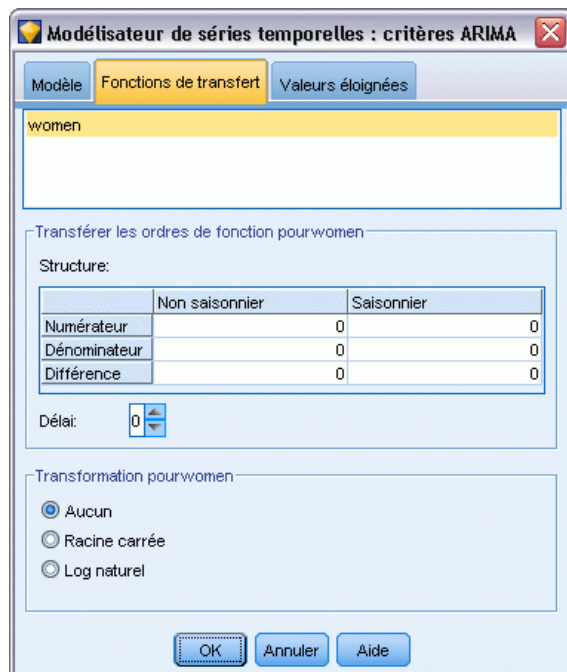
Transformation cible. Vous pouvez indiquer qu'une transformation doit être effectuée sur chaque variable cible avant que celle-ci ne soit modélisée. [Pour plus d'informations, reportez-vous à la section Transformations de série sur p. 464.](#)

- **Aucune.** Aucune transformation n'est effectuée.
- **Racine carrée :** Une transformation par la racine carrée est effectuée.
- **Log népérien :** Une transformation par le logarithme naturel est effectuée.

Inclure terme constant dans le modèle. L'inclusion d'une constante est normale, à moins que vous ne soyez sûr que la valeur de série de la moyenne générale est 0. L'exclusion de la constante est recommandée lorsque la différenciation s'applique.

Fonctions de transfert

Figure 13-14
Boîte de dialogue Critères ARIMA, onglet Fonctions de transfert



L'onglet Fonctions de transfert vous permet de définir des fonctions de transfert pour les champs d'entrée de votre choix. Les fonctions de transfert vous permettent de spécifier la manière dont les valeurs passées de ces champs sont utilisées pour la prévision des valeurs futures des séries cible.

L'onglet apparaît uniquement si les champs d'entrée (avec le rôle défini sur *Entrée*) sont spécifiés soit sur le noeud *Type* soit dans l'onglet *Champs* du noeud *Séries temporelles* (sélectionner *Utiliser les paramètres d'entrées personnalisés*). [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*.](#)

La liste supérieure recense tous les champs d'entrées. Les autres informations de cette boîte de dialogue concernent le champ d'entrée sélectionné dans la liste.

Ordres de fonction de transfert. Entrez des valeurs pour les différents composants de la fonction de transfert dans les cellules correspondantes de la grille *Structure*. Toutes les valeurs doivent être des entiers non négatifs. Pour les composants numérateur et dénominateur, la valeur représente l'ordre maximum. Tous les ordres inférieurs positifs seront inclus dans le modèle. En outre, l'ordre 0 est toujours inclus pour les composants numérateur. Par exemple, si vous spécifiez 2 comme numérateur, le modèle comprend les ordres 2, 1 et 0. Si vous indiquez 3 en guise de dénominateur, le modèle comporte les ordres 3, 2 et 1. Les cellules de la colonne *Saisonnier* ne sont activées que si une périodicité a été définie pour l'ensemble de données actif.

Numérateur. L'ordre du numérateur de la fonction de transfert détermine les valeurs précédentes de la série indépendante sélectionnée (valeur prédite) qui sont utilisées pour la prévision des valeurs actuelles de la série dépendante. Par exemple, un ordre de numérateur de 1 indique que la valeur d'une série indépendante Un point dans le temps dans le passé, ainsi que la valeur courante de la série indépendante, sont utilisées pour prévoir la valeur courante de chaque série dépendante.

Dénominateur. L'ordre du dénominateur de la fonction de transfert détermine la manière dont les écarts par rapport à la moyenne de la série, pour les valeurs précédentes de la série indépendante sélectionnée (valeur prédite), sont utilisés pour la prévision des valeurs actuelles de la série dépendante. Par exemple, un ordre de dénominateur de 1 indique que les écarts de la valeur moyenne d'une série indépendante Un point dans le temps dans le passé doivent être considérés lors de la prévision de la valeur courante de chaque série dépendante.

Différence : Spécifie l'ordre de différenciation appliqué à la série indépendante (explicative) avant d'estimer les modèles. La différenciation est nécessaire lorsque les tendances sont présentes et est utilisée pour supprimer leur effet.

Ordres saisonniers. Les composants numérateur, dénominateur et de différenciation saisonniers tiennent le même rôle 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.

Délai. La définition d'un délai retarde l'influence du champ d'entrée du nombre d'intervalles spécifié. Par exemple, si le délai a pour valeur 5, la valeur du champ d'entrée au moment t n'affectera les prévisions qu'une fois les cinq périodes écoulées ($t + 5$).

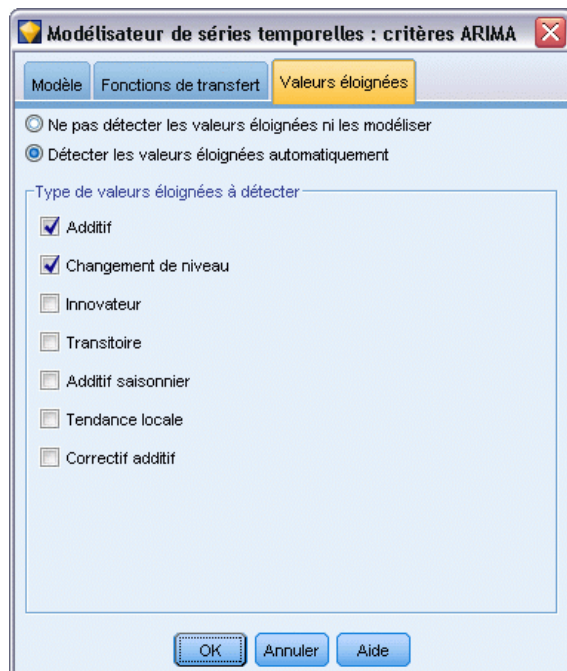
Transformation. La spécification d'une fonction de transfert pour un ensemble de variables indépendantes comprend une transformation facultative à exécuter sur ces variables.

■ **Aucune.** Aucune transformation n'est effectuée.

- **Racine carrée** : Une transformation par la racine carrée est effectuée.
- **Log népérien** : Une transformation par le logarithme naturel est effectuée.

Gestion des valeurs éloignées

Figure 13-15
Boîte de dialogue Critères ARIMA, onglet Valeurs éloignées



L'onglet Valeurs éloignées comprend une série d'options qui permettent de gérer les valeurs éloignées contenues dans les données .

Ne pas détecter ou modéliser les valeurs éloignées. Par défaut, les valeurs éloignées ne sont ni détectées ni modélisées. Sélectionnez cette option pour désactiver la détection ou la modélisation des valeurs éloignées.

Détecter les valeurs éloignées automatiquement. Sélectionnez cette option pour effectuer une détection automatique des valeurs éloignées et sélectionnez un ou plusieurs des types de valeur éloignée proposés.

Type de valeurs éloignées à détecter. Sélectionnez les types de valeur éloignée à détecter. Les types pris en charge sont :

- Additif (sélectionné par défaut)
- Changement de niveau (sélectionné par défaut)
- Innovation
- Transitoire
- Additive saisonnière

- Tendance locale
- Additive corrective

Pour plus d'informations, reportez-vous à la section Valeurs éloignées sur p. 462.

Génération de modèles de séries temporelles

Cette section donne des informations générales à propos de certains aspects de la génération de modèles de séries temporelles :

- Génération de plusieurs modèles
- Utilisation des modèles de séries temporelles à des fins de prévision
- Nouvelle estimation et prévision

Le nugget de modèle généré est décrit dans une rubrique séparée. [Pour plus d'informations, reportez-vous à la section Nugget du modèle Séries temporelles sur p. 480.](#)

Génération de plusieurs modèles

Dans IBM® SPSS® Modeler, la modélisation des séries temporelles génère un modèle unique (ARIMA ou lissage exponentiel) pour chaque champ cible. Par conséquent, en présence de plusieurs champs cible, SPSS Modeler génère plusieurs modèles simultanément, ce qui permet de gagner du temps et de comparer les paramètres de chaque modèle.

Pour comparer un modèle ARIMA et un modèle de lissage exponentiel pour le même champ cible, vous pouvez effectuer des exécutions distinctes du noeud Séries temporelles, en spécifiant un modèle différent à chaque fois.

Utilisation des modèles de séries temporelles à des fins de prévision

La génération d'une série temporelle utilise une série spécifique d'observations ordonnées, appelée amplitude d'estimation, pour créer un modèle de prévision des valeurs futures de la série. Ce modèle contient des informations sur l'intervalle de temps utilisé. Pour effectuer une prévision à l'aide de ce modèle, vous devez utiliser le même intervalle de temps et les mêmes informations d'intervalle avec la même série, à la fois pour la variable cible et pour les variables de prévision.

Par exemple, supposons qu'au début du mois de janvier vous souhaitez prévoir les ventes mensuelles du produit 1 pour les trois premiers mois de l'année. Vous créez un modèle à l'aide des données des ventes mensuelles du produit 1 effectivement réalisées pendant la période allant du mois de janvier au mois de décembre de l'année précédente (que nous appellerons année 1), en attribuant au paramètre Intervalle de temps la valeur « Mois ». Vous pouvez ensuite utiliser le modèle afin de prévoir les ventes du produit 1 pour les trois premiers mois de l'année 2.

En fait, la prévision pourrait porter sur un nombre de mois futurs quelconque, mais il va de soi que l'efficacité du modèle est inversement proportionnelle à la durée de la période considérée. Toutefois, il serait impossible d'effectuer une prévision pour les trois premières semaines de l'année 2, car l'intervalle utilisé pour créer le modèle a pour valeur « Mois ». En outre, cela n'aurait aucun sens d'utiliser ce modèle pour la prévision des ventes du produit 2, puisqu'un

modèle de séries temporelles n'est approprié que pour les données ayant servi à le définir. [Pour plus d'informations, reportez-vous à la section Prévision avec le noeud Séries temporelles dans le chapitre 14 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Nouvelle estimation et prévision

La période d'estimation est codée en dur dans le modèle généré. Cela signifie que toutes les valeurs situées hors de la période d'estimation sont ignorées si vous appliquez le modèle actuel à de nouvelles données. Par conséquent, vous devez de nouveau estimer un modèle de séries temporelles chaque fois que de nouvelles données sont disponibles, contrairement aux autres modèles IBM® SPSS® Modeler, que vous pouvez réappliquer tels quels à des fins de scoring.

Poursuivons l'exemple précédent, en supposant qu'au plus tard au début du mois d'avril de l'année 2, vous disposiez des données de ventes mensuelles réalisées de janvier à mars. Toutefois, si vous réappliquez le modèle que vous avez généré au début du mois de janvier, il effectuera une nouvelle prévision portant sur la période de janvier à mars et ignorera les données de ventes connues relatives à cette période.

La solution consiste à générer un nouveau modèle basé sur les données réelles mises à jour. Dans l'hypothèse où vous ne modifiez pas les paramètres de prévision, vous pouvez utiliser le nouveau modèle pour effectuer une prévision portant sur les trois prochains mois (avril à juin). Si vous avez toujours accès au flux ayant servi à générer le modèle d'origine, il vous suffit, pour obtenir le nouveau modèle, de remplacer la référence au fichier source dans ce flux par une référence au fichier contenant les données mises à jour et de réexécuter le flux. Toutefois, si vous ne disposez que du modèle d'origine enregistré dans un fichier, vous pouvez l'utiliser pour générer un noeud Séries temporelles en vue de l'ajouter à un nouveau flux contenant une référence au fichier source mis à jour. Sous réserve que, dans ce nouveau flux, le noeud Séries temporelles soit précédé d'un noeud Intervalle de temps où l'intervalle a pour valeur « Mois », l'exécution de ce nouveau flux génère le nouveau modèle requis. [Pour plus d'informations, reportez-vous à la section Réapplication d'un modèle de séries temporelles dans le chapitre 14 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Nugget du modèle Séries temporelles

La modélisation des séries temporelles crée un ensemble de nouveaux champs dotés du préfixe \$TS-, comme suit :

\$TS- <i>colname</i>	La valeur prévue par le modèle pour chaque série de cibles.
\$TSLCI- <i>colname</i>	Les intervalles de confiance inférieurs pour chaque série prévue.*
\$TSUCI- <i>colname</i>	Les intervalles de confiance supérieurs pour chaque série prévue.*
\$TSNR- <i>nom_colonne</i>	Valeur des résidus de bruit pour chaque colonne de données du modèle généré.*
\$TS-Total	Total des valeurs de \$TS- <i>colname</i> pour cette ligne.
\$TSLCI-Total	Total des valeurs de \$TSLCI- <i>nom_colonne</i> pour cette ligne.*
\$TSUCI-Total	Total des valeurs de \$TSUCI- <i>nom_colonne</i> pour cette ligne.*
\$TSNR-Total	Total des valeurs de \$TSNR- <i>nom_colonne</i> pour cette ligne.*

* La visibilité de ces champs (par exemple dans la sortie d'un noeud Table attaché) dépend des options dans l'onglet Paramètres du nugget de modèle de séries temporelles. [Pour plus d'informations, reportez-vous à la section Paramètres du modèle de séries temporelles sur p. 487.](#)

Figure 13-16
Onglet Modèle - nugget de modèle de séries temporelles

The screenshot shows a window titled '6 fields' with a menu bar (Fichier, Générer, Aperçu) and a toolbar. Below the menu is a tabbed interface with 'Modèle' selected. A 'Trier par' dropdown is set to 'Sélectionné(e)(s)'. Below that, it says 'Nombre d'enregistrements utilisés dans l'estimation: 60'. The main table lists six targets with their respective models and statistics.

	Cible	Modèle	Valeurs prédites	R**2 stationnaire	Q	ddl	Sig.
<input checked="" type="checkbox"/>	Market_1	Tendance lin...	0	0,264	8,53	16,0	0,931
<input checked="" type="checkbox"/>	Market_2	Tendance lin...	0	0,121	35,9	16,0	0,003
<input checked="" type="checkbox"/>	Market_3	Tendance lin...	0	0,258	15,76	16,0	0,47
<input checked="" type="checkbox"/>	Market_4	Tendance lin...	0	0,25	27,714	16,0	0,034
<input checked="" type="checkbox"/>	Market_5	Additif de Vi...	0	0,544	11,888	15,0	0,688
<input checked="" type="checkbox"/>	Total	Tendance lin...	0	0,049	27,616	16,0	0,035

Below this table is a section for 'Statistiques récapitulatives' with another table:

	Statistique	R**2 stationnaire	Q	ddl	Sig.
SUMMARY	MOYENNE	0,247	21,235	15,833	0,36
SUMMARY	ES	0,169	10,738	0,408	0,396
SUMMARY	MINIMUM	0,049	8,53	15	0,003
SUMMARY	MAXIMUM	0,544	35,9	16	0,931
SUMMARY	PERCENTILE 5	0,049	8,53	15	0,003
SUMMARY	PERCENTILE ...	0,049	8,53	15	0,003
SUMMARY	PERCENTILE ...	0,103	11,048	15,75	0,026
SUMMARY	PERCENTILE ...	0,254	21,688	16	0,252
SUMMARY	PERCENTILE ...	0,334	29,761	16	0,749
SUMMARY	PERCENTILE ...	0,544	35,9	16	0,931
SUMMARY	PERCENTILE ...	0,544	35,9	16	0,931

At the bottom are buttons for 'OK', 'Annuler', 'Appliquer', and 'Réinitialiser'.

Le nugget de modèle Séries temporelles affiche les détails des différents modèles sélectionnés pour chacune des séries entrées dans le noeud de création Séries temporelles. Plusieurs séries (telles que les données relatives aux gammes de produits, aux régions ou aux magasins) peuvent être entrées et un modèle distinct est généré pour chaque série cible. Par exemple, si le chiffre d'affaires réalisé dans la région est s'avère adapté à un modèle ARIMA, alors que celui atteint dans la région ouest est uniquement adapté à une moyenne mobile simple, chaque région est évaluée avec le modèle approprié.

Pour chaque modèle créé, la sortie par défaut indique le type de modèle, le nombre de valeurs prédites spécifiées et la mesure de la qualité d'ajustement (le R -deux stationnaire est la mesure par défaut). Si vous avez spécifié des méthodes de recherche de valeurs éloignées, une colonne indique le nombre de valeurs éloignées détectées. En outre, la sortie par défaut comprend des colonnes pour la statistique Q de Ljung-Box, les degrés de liberté et les valeurs de signification.

Vous pouvez également opter pour une sortie avancée, qui affiche les colonnes supplémentaires suivantes :

- R-deux
- RMSE (erreur moyenne quadratique)
- MAPE (erreur moyenne absolue en pourcentage)
- MAE (nombre moyen d'erreurs absolues)
- MaxAPE (erreur maximale absolue en pourcentage)
- MaxAE (nombre maximum d'erreurs absolues)
- BIC norm (critère d'information de Bayes normalisé)

Générer. Permet de générer un noeud de modélisation Séries temporelles dans le flux ou un nugget de modèle dans la palette.

- **Générer un noeud de modélisation.** Place un noeud de modélisation Séries temporelles dans un flux en lui associant les paramètres ayant servi à créer cet ensemble de modèles. Vous pouvez, par exemple, recourir à cette option si vous trouvez en présence d'un flux dans lequel vous souhaitez utiliser ces paramètres de modèle alors que vous ne disposez plus du noeud de modélisation ayant servi à les générer.
- **Modèle vers palette.** Place, dans le gestionnaire de modèles, un nugget de modèle contenant toutes les cibles.

Modèle

Figure 13-17

Boutons Tout sélectionner et Tout désélectionner



Cases à cocher. Choisissez les modèles à utiliser dans le scoring. Par défaut, toutes les cases sont cochées. Les boutons Tout sélectionner et Tout désélectionner agissent simultanément sur toutes les cases.

Trier par. Permet de trier les lignes de la sortie dans l'ordre croissant ou décroissant de la colonne d'affichage de votre choix. L'option "Sélectionnées" permet de trier les lignes sélectionnées à l'aide des cases à cocher. Cela peut, par exemple, s'avérer utile pour déplacer les champs cible "Marché_1" à "Marché_9" avant le champ "Marché_10", puisque l'ordre de tri par défaut affiche "Marché_10" immédiatement après "Marché_1".

Affichage. L'affichage par défaut (Simple) comprend l'ensemble des colonnes de sortie de base. L'option Options avancées affiche des colonnes supplémentaires pour les mesures de la qualité d'ajustement.

Nombre d'enregistrements utilisés dans l'estimation. Nombre de lignes dans le fichier de données source d'origine.

Cible. Champs identifiés en tant que champs cibles (ceux avec un rôle de *Cible*) dans le noeud Typar.

Modèle : Type de modèle utilisé pour ce champ cible.

Valeurs prédites. Le nombre de valeurs prédites (ceux avec un rôle de type *Entrée*) utilisés pour ce champ cible.

Valeurs éloignées : Cette colonne n'apparaît que si vous avez demandé, dans le modélisateur expert ou les critères ARIMA, la détection automatique des valeurs éloignées. La valeur indiquée représente le nombre de valeurs éloignées détectées.

R-deux stationnaire. Une mesure qui compare la partie stationnaire du modèle à un simple modèle de moyenne. Cette mesure est préférable à un R-deux ordinaire lorsqu'il y a une tendance ou un motif saisonnier. Le R-deux stationnaire peut être négatif avec une plage d'infinité négative de 1. Les valeurs négatives signifient que la sous-considération du modèle est pire que le modèle de la ligne de base. Les valeurs positives signifient que la sous-considération du modèle est meilleure que le modèle de la ligne de base.

R-deux. Mesure de la qualité d'ajustement d'un modèle linéaire, parfois appelée coefficient de détermination. Il s'agit de la proportion de la variation de la variable dépendante, expliquée par le modèle de régression. Elle varie entre 0 et 1. Des valeurs faibles indiquent que le modèle n'est pas bien estimé.

RMSE. Erreur quadratique moyenne. La racine carrée de l'erreur quadratique moyenne. Une mesure de la variation de la série dépendante par rapport au niveau de prédiction, exprimée dans les mêmes unités que la série dépendante.

MAPE. Erreur de pourcentage Moyenne absolue. Une mesure de la variation de la série dépendante par rapport au niveau prévu par le modèle. Elle est indépendante des mesures utilisées et peut donc être utilisée pour comparer les séries comportant des unités différentes.

Erreur absolue moyenne. Erreur Moyenne absolue. Mesure la variation de la série par rapport au niveau prévu par le modèle. La MAE est reportée dans les unités de séries d'origine.

MaxAE. Erreur de pourcentage absolu maximale. La plus grande erreur prévue, exprimée en pourcentage. Cette mesure est utile pour imaginer le pire scénario lors de vos prévisions.

MaxAE. Erreur absolue maximale. La plus grande erreur prévue, exprimée dans les mêmes unités que la série dépendante. Comme MaxAPE, elle est utile pour imaginer le pire scénario lors de vos prévisions. L'erreur Maximum absolue et l'erreur de pourcentage Maximum absolue peuvent se produire à différents moments d'une série, par exemple lorsque l'erreur absolue pour la valeur d'une grande série est légèrement supérieure à l'erreur absolue pour la valeur d'une petite série. Dans ce cas, l'erreur maximum absolue se produira à la valeur de la grande série et l'erreur de pourcentage maximum absolue se produira à la valeur de la petite série.

Critère d'information bayésien normalisé (BIC). Critère d'information bayésien normalisé. Une mesure générale de l'ajustement global d'un modèle qui essaye de prendre en compte la complexité du modèle. C'est un résultat basé sur l'erreur quadratique moyenne et qui inclut une pénalité pour le nombre de paramètres du modèle et la longueur de la série. La pénalité supprime l'avantage des modèles disposant de plus de paramètres, rendant les statistiques plus faciles à comparer parmi les différents modèles d'une même série.

Q. La statistique Q de Ljung-Box. Un test de l'aspect aléatoire des erreurs résiduelles dans ce modèle.

ddl. Degrés de liberté. Le nombre de paramètres du modèle qui peuvent varier librement lors de l'estimation d'une cible particulière.

Sig. Valeur de signification de la statistique Ljung-Box. Une valeur de signification inférieure à 0,05 indique que les erreurs résiduelles ne sont pas aléatoires.

Statistiques récapitulatives Cette section contient diverses statistiques récapitulatives pour les différentes colonnes, notamment pour les valeurs moyenne, minimale, maximale et de centile.

Paramètres du modèle Séries temporelles

Figure 13-18
Modèle Séries temporelles, onglet Paramètres

Cible	Modèle	Champ	Transfor...	Paramètre	Décalage	Estimation	ES	t	Sig.
Market_2	ARIMA(1,...	Market_2	Log naturel	Constante		9,73	2,662	3,655	0,001
				AR	Décalage 1	0,999	0,012	84,65	0,0
				Autorégre...	Décalage 1	0,786	0,154	5,094	0,0

L'onglet Paramètres répertorie les détails des différents paramètres utilisés pour créer un modèle sélectionné.

Afficher les paramètres pour le modèle. Sélectionnez le modèle dont vous souhaitez afficher les paramètres détaillés.

Cible. Le nom du champ cible (avec le rôle de type *Cible*) prévu par ce modèle.

Modèle : Type de modèle utilisé pour ce champ cible.

Champ (Modèles ARIMA uniquement). Contient une entrée pour chacune des variables utilisées dans le modèle, avec la cible en premier lieu, suivie des prédicteurs éventuels.

Transformation. Indiquez quel type de transformation a été spécifié éventuellement pour ce champ avant la création du modèle.

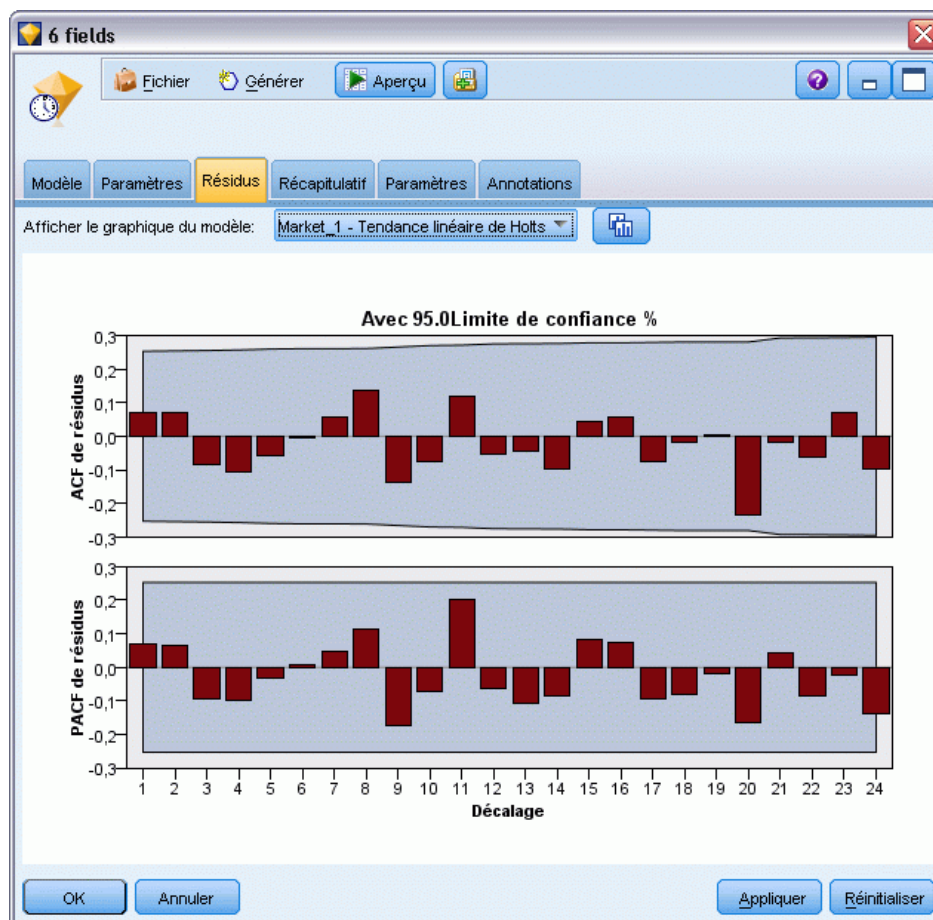
Paramètre. Le paramètre du modèle pour lequel les détails suivants sont affichés :

- **Décalage (Modèles ARIMA uniquement).** Indique les éventuels onglets pris en compte pour ce paramètre du modèle.
- **Estimation.** Le paramètre d'estimation. Cette valeur est utilisée pour calculer la valeur de prévision et les intervalles de confiance pour le champ cible.
- **ES.** L'erreur standard de l'estimation de paramètre.

- **t**. La valeur de l'estimation de paramètre divisée par l'erreur standard.
- **Sig.** Le niveau de signification pour l'estimation du paramètre. Les valeurs supérieures à 0,05 sont considérées comme insignifiantes sur le plan statistique.

Résidus du modèle de séries temporelles

Figure 13-19
Modèle Séries temporelles, onglet Résidus, affichage ACF et PACF

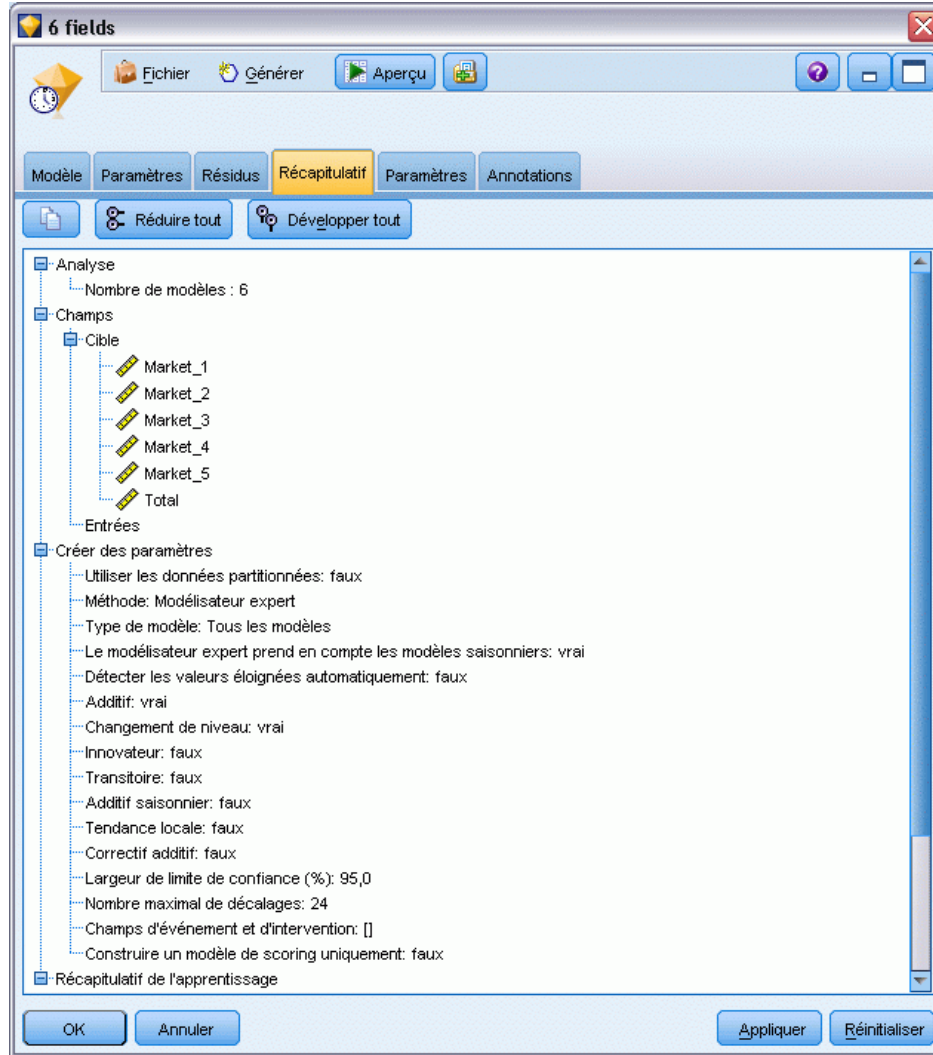


L'onglet Résidus indique la fonction d'auto-corrélation et la fonction d'auto-corrélation partielle des résidus (différences entre les valeurs théoriques et les valeurs réelles) pour chaque modèle créé. [Pour plus d'informations, reportez-vous à la section Fonctions d'autocorrélation et d'autocorrélation partielle sur p. 463.](#)

Afficher le graphique du modèle. Sélectionnez le modèle pour lequel vous souhaitez afficher les fonctions d'auto-corrélation et d'auto-corrélation partielle des résidus.

Récapitulatif du modèle de séries temporelles

Figure 13-20
Modèle Séries temporelles, onglet Résumé



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.

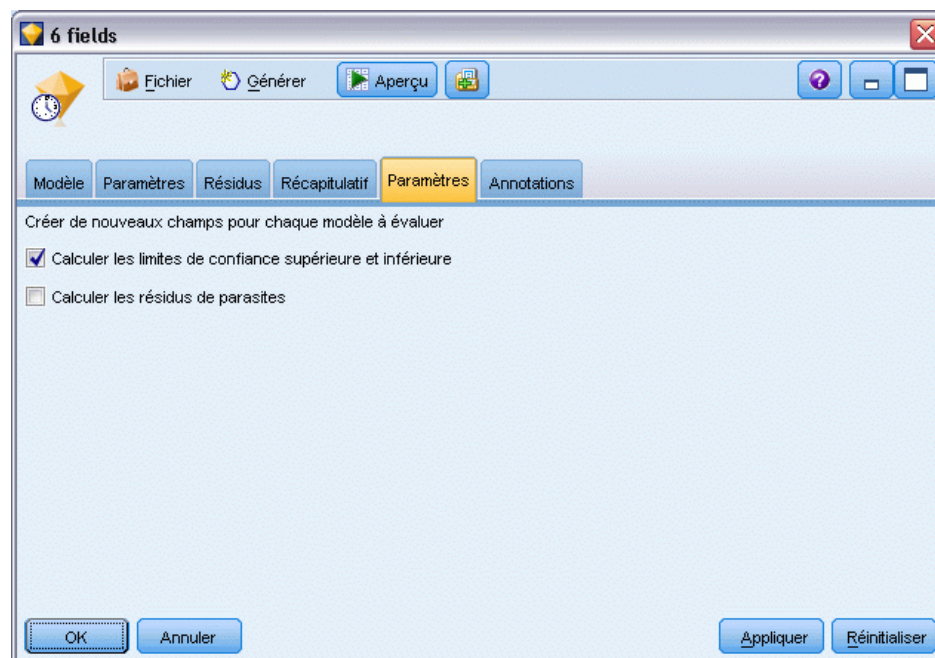
Champs. Répertorie les champs utilisés comme cibles et entrées lors de la création du modèle.

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.

Paramètres du modèle de séries temporelles

Figure 13-21
Modèle Séries temporelles, onglet Paramètres



L'onglet Paramètres vous permet d'indiquer les champs supplémentaires que doit créer l'opération de modélisation.

Créer de nouveaux champs pour chaque modèle à évaluer. Permet de spécifier les nouveaux champs à créer pour chaque modèle à évaluer.

- **Calculer les limites de confiance supérieure et inférieure.** Si cette option est sélectionnée, elle permet de créer de nouveaux champs (dotés des préfixes par défaut \$TSLCI- et \$TSUCI-) respectivement pour les valeurs inférieure et supérieure de l'intervalle de confiance, et ce, pour chaque champ cible, et de calculer les totaux de ces valeurs.
- **Calculer les résidus de parasites.** Si cette option est sélectionnée, elle permet de créer un nouveau champ (doté du préfixe par défaut \$TSNR-) pour les résidus de modèle, et ce, pour chaque champ cible, et de calculer le total de ces valeurs.

Modèles de noeud Réponse en auto-apprentissage

Noeud MRAA

Le noeud **Modèle de réponse en auto-apprentissage** (MRAA) vous permet de créer un modèle que vous pouvez continuellement mettre à jour ou réestimer tout au long de la croissance d'un ensemble de données, sans qu'il soit nécessaire de recréer le modèle chaque fois sur la base de l'ensemble de données complet. Il est par exemple utile lorsque vous avez plusieurs produits et que vous voulez identifier le produit qu'un client est le plus susceptible d'acheter si vous le lui proposez. Ce modèle vous permet de prédire quelles offres sont les plus appropriées pour les clients et la probabilité d'acceptation des offres.

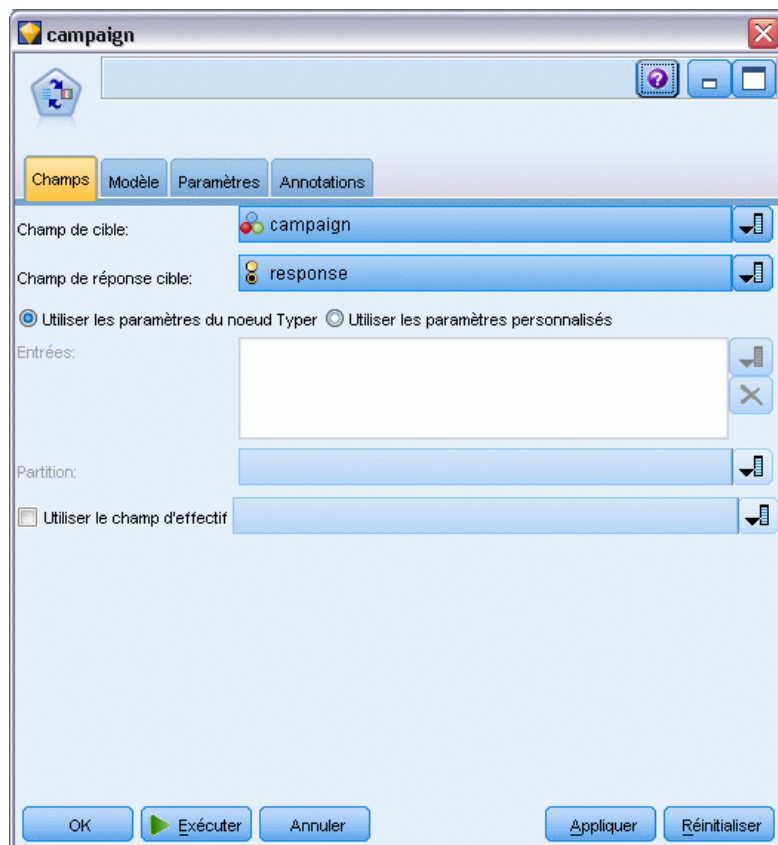
Le modèle peut être initialement créé à l'aide d'un petit ensemble de données avec des offres faites au hasard et les réponses à ces offres. Au fur et à mesure que l'ensemble de données augmente, le modèle peut être mis à jour. Il est donc de plus en plus à même de prédire les offres les mieux adaptées aux clients et la probabilité d'acceptation sur la base des autres champs d'entrée tels que l'âge, le sexe, la profession et le revenu. Les offres disponibles peuvent être modifiées en les ajoutant ou en les supprimant dans la boîte de dialogue du noeud au lieu de devoir modifier le champ cible de l'ensemble de données.

Lorsque vous utilisez IBM® SPSS® Collaboration and Deployment Services, vous pouvez définir des mises à jour automatiques du modèle, à intervalles réguliers. Ce processus, qui ne nécessite aucune supervision ou action humaine, offre une solution souple et économique pour les entreprises et applications où une intervention personnalisée par Data Miner n'est pas possible ou nécessaire.

Exemple : Une institution financière souhaite obtenir des résultats plus rentables en présentant à chaque client l'offre qui est le plus susceptible d'être acceptée. Vous pouvez utiliser un modèle d'auto-apprentissage pour identifier les caractéristiques des clients les plus susceptibles de répondre favorablement, sur la base des anciennes promotions, et pour mettre à jour le modèle en temps réel sur la base des dernières réponses des clients. [Pour plus d'informations, reportez-vous à la section Propositions aux clients \(auto-apprentissage\) dans le chapitre 16 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Noeud MRAA - Options de champs

Figure 14-1
Boîte de dialogue du noeud MRAA - Onglet Champs



Avant d'exécuter un noeud MRAA, vous devez indiquer les champs cible et les champs de réponse cible dans l'onglet Champs du noeud.

Champ cible. Sélectionnez le champ cible dans la liste, il peut s'agir, par exemple, d'un champ nominal (ensemble) contenant les différents produits que vous voulez proposer aux clients.

Remarque : Le champ cible doit avoir un stockage chaîne et non un stockage numérique.

Champ de réponse cible. Sélectionnez le champ de réponse cible dans la liste. Par exemple, Accepté ou Rejeté.

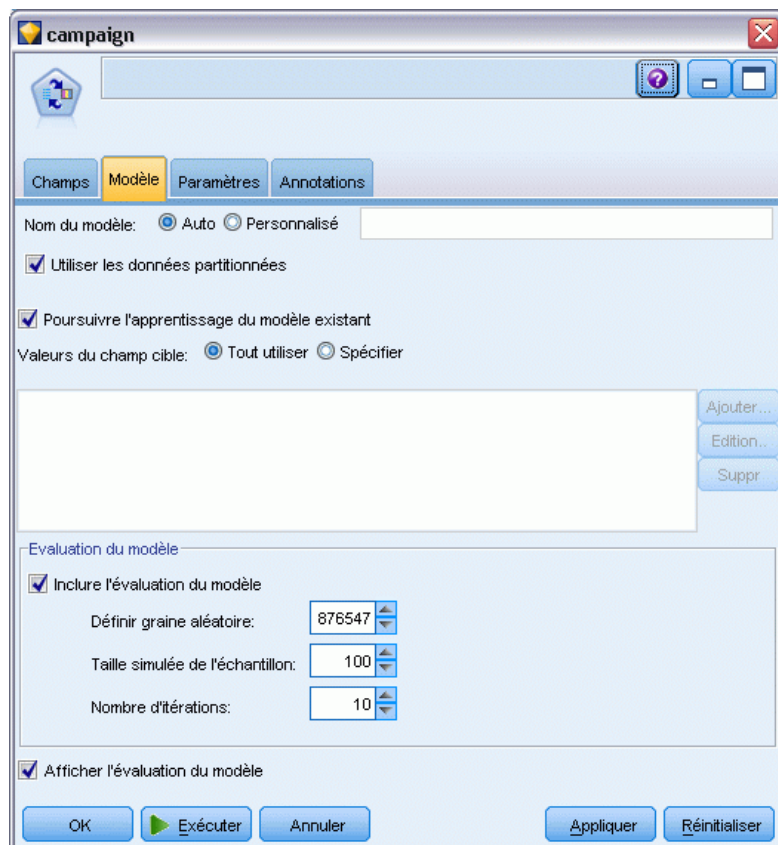
Remarque : Ce champ doit être booléen. La valeur vraie (True) du booléen indique l'acceptation de l'offre et la valeur fausse (False) indique son refus.

Les autres champs de cette boîte de dialogue sont les champs habituels de IBM® SPSS® Modeler. [Pour plus d'informations, reportez-vous à la section Options de champs des noeuds de modélisation dans le chapitre 3 sur p. 39.](#)

Remarque : si les données source incluent des intervalles qui doivent être utilisés en tant que champs d'entrée continus (intervalle numérique), vous devez vérifier que les métadonnées incluent les détails minimum et maximum pour chaque intervalle.

Noeud MRAA - Options du modèle

Figure 14-2
Boîte de dialogue du noeud MRAA - 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.](#)

Poursuivre l'apprentissage du modèle existant. Par défaut, un modèle complètement nouveau est créé chaque fois qu'un noeud de modélisation est exécuté. Si vous sélectionnez cette option, l'apprentissage se poursuit sur le dernier modèle généré par le noeud. Cela permet de mettre à jour ou d'actualiser un modèle existant sans avoir à accéder aux données d'origine et peut permettre des performances nettement plus rapides car *seuls* les enregistrements nouveaux ou mis à jour sont acheminés dans le flux. Les détails relatifs au modèle précédent sont stockés avec le noeud de modélisation, ce qui permet d'utiliser cette option même si le nugget du modèle précédent n'est plus disponible dans le flux ou dans la palette de modèles.

Valeurs du champ cible Par défaut, ce champ a la valeur Utiliser tout, ce qui signifie que le modèle créé contient chaque offre associée à la valeur du champ cible sélectionné. Si vous souhaitez générer un modèle uniquement pour certaines offres de champ cible, cliquez sur Spécifier et utilisez les boutons Ajouter, Modifier et Supprimer pour entrer ou modifier le nom des offres pour lesquelles vous souhaitez créer des modèles. Par exemple, si vous choisissez une cible qui répertorie tous les produits que vous fournissez, vous pouvez utiliser ce champ pour limiter les produits proposés à ceux que vous avez entrés ici.

Evaluation du modèle. Les champs de ce panneau sont indépendants du modèle car ils n'influencent pas sur le scoring. Ils vous permettent par contre de créer une représentation visuelle de la qualité de prévision des résultats par le modèle.

Remarque : Pour afficher les résultats de l'évaluation dans le nugget de modèle, vous devez également sélectionner la case Afficher l'évaluation du modèle.

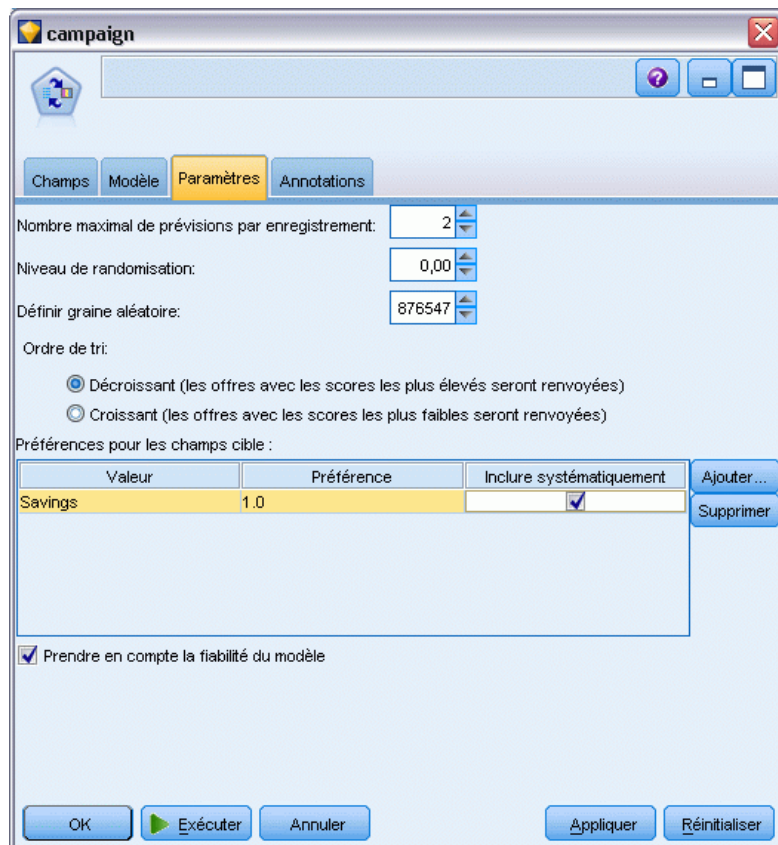
- **Inclure l'évaluation du modèle.** Cochez cette case pour créer des graphiques qui affichent la précision prédite du modèle de chaque offre sélectionnée.
- **Définir graine aléatoire.** Lors de l'estimation de la précision d'un modèle basé sur un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.
- **Taille simulée de l'échantillon.** Indiquez le nombre d'enregistrements à utiliser dans l'échantillon lors de l'évaluation du modèle. La valeur par défaut est 100.
- **Nombre d'itérations.** Cela vous permet d'arrêter la création de l'évaluation du modèle une fois que le nombre d'itérations indiqué a été atteint. Précisez le nombre maximal d'itérations ; la valeur par défaut est 20.

Remarque : gardez à l'esprit que les échantillons volumineux et les nombres élevés d'itérations augmentent le temps nécessaire à la création du modèle.

Afficher l'évaluation du modèle. Sélectionnez cette option pour afficher une représentation graphique des résultats dans le nugget de modèle.

Noeud MRAA - Options de paramètres

Figure 14-3
Boîte de dialogue du noeud MRAA, onglet Paramètres



Les options de paramètres de noeud vous permettent d'affiner le processus de création de modèle.

Nombre maximal de prévisions par enregistrement. Cette option permet de limiter le nombre de prévisions effectuées pour chaque enregistrement d'un ensemble de données. La valeur par défaut est 3.

Par exemple, vous pouvez disposer de six offres (épargne, prêt hypothécaire, prêt automobile, retraite, carte de crédit et assurance) mais ne souhaiter connaître que les deux meilleures recommandations possibles, auquel cas il convient de définir la valeur 2 pour ce champ. Une fois le modèle créé et relié à une table, deux colonnes de prévisions (plus la confiance associée dans la probabilité d'acceptation de l'offre) seront alors visibles pour chaque enregistrement. Les prévisions peuvent comprendre l'une des six offres possibles.

Niveau de randomisation. Pour éviter tout biais, notamment dans le cas d'un ensemble de données peu volumineux ou incomplet, et traiter toutes les offres potentielles de la même manière, vous pouvez ajouter un niveau de randomisation aux offres proposées, ainsi que la probabilité de leur affichage sous forme de recommandations. La randomisation est exprimée en pourcentage sous la forme d'une valeur décimale comprise entre 0,0 (sans randomisation) et 1,0 (randomisation complète). La valeur par défaut est 0,0.

Définir graine aléatoire. Lorsque vous ajoutez un niveau de randomisation à la sélection d'une offre, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Lorsque vous utilisez l'option Définir graine aléatoire avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la graine aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. [Pour plus d'informations, reportez-vous à la section Noeud Trier dans le chapitre 3 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Ordre de tri. Sélectionnez l'ordre d'affichage des offres dans le modèle créé :

- **Décroissant.** Le modèle affiche les offres avec le score le plus élevé en premier. Il s'agit des offres ayant la plus forte probabilité d'être acceptées.
- **Croissant.** Le modèle affiche les offres avec le score le plus faible en premier. Il s'agit des offres ayant la plus forte probabilité d'être rejetées. Cette option peut se révéler utile lorsque vous décidez des clients à exclure d'une campagne de marketing pour une offre spécifique.

Préférences des champs cible. Pendant la création d'un modèle, vous pouvez être amené à favoriser ou à ignorer certains aspects des données. Si, par exemple, vous créez un modèle de sélection de la meilleure offre financière à proposer à un client, il peut être judicieux de veiller à toujours inclure une offre particulière quel que soit son score par client.

Pour inclure une offre dans ce panneau et modifier ses préférences, cliquez sur le bouton Ajouter, saisissez le nom de l'offre (par exemple, Epargne ou Prêt hypothécaire), puis cliquez sur le bouton OK.

- **Valeur.** Il s'agit du nom de l'offre que vous venez d'ajouter.
- **Préférence.** Il s'agit du niveau de préférence à appliquer à l'offre. La préférence est exprimée en pourcentage sous la forme d'une valeur décimale comprise entre 0,0 (sans préférence) et 1,0 (préférence maximale). La valeur par défaut est 0,0.
- **Inclure systématiquement.** Cochez cette case pour inclure systématiquement une offre spécifique dans les prévisions.

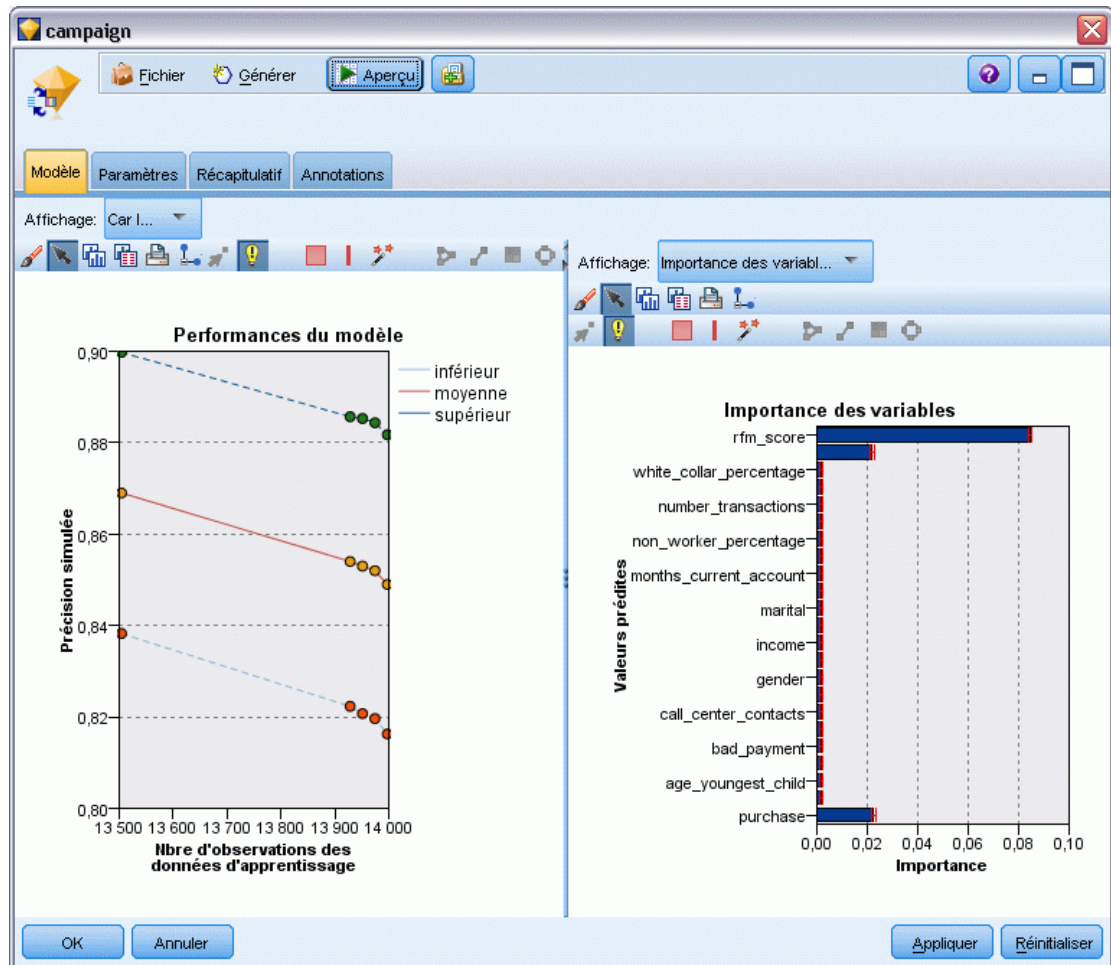
Remarque : Si l'option Préférence a pour valeur 0,0, le paramètre Inclure systématiquement n'est pas pris en compte.

Prendre en compte la fiabilité du modèle. Un modèle riche en données, bien structuré et affiné par plusieurs régénérations est censé produire des résultats plus précis qu'un tout nouveau modèle contenant peu de données. Cochez cette case pour bénéficier de la plus grande fiabilité du modèle plus abouti.

Nuggets de modèle MRAA

Remarque : les résultats sont uniquement affichés dans cet onglet si vous sélectionnez Inclure l'évaluation du modèle et Afficher l'évaluation du modèle dans l'onglet Options de modèle.

Figure 14-4
Affichage des nuggets de modèle MRAA



Lorsque vous exécutez un flux qui contient un modèle MRAA, le noeud estime la précision des prédictions de chaque valeur de champ cible (offre) et l'importance de chaque valeur prédite utilisée.

Remarque : si vous avez sélectionné Poursuivre l'apprentissage du modèle existant dans l'onglet Modèle du noeud de modélisation, les informations affichées dans cet onglet Modèle du nugget de modèle sont mises à jour chaque fois que vous régénérez le modèle.

Lors de la création de modèles avec IBM® SPSS® Modeler 12.0 ou une version ultérieure, l'onglet Modèle du nugget de modèle est divisé en deux colonnes :

Colonne de gauche.

- **Affichage.** Lorsque vous avez plusieurs offres, sélectionnez celle dont vous voulez afficher les résultats.
- **Performances du modèle.** Cette option permet d'afficher la précision de modèle estimée de chaque offre. L'ensemble de test est généré par la simulation.

Colonne droite.

- **Affichage.** Choisissez l’affichage des détails de Association avec réponse ou Importance des variables.
- **Association avec réponse.** Affiche l’association (corrélation) de chaque valeur prédite avec la variable cible.
- **Importance des valeurs prédites.** Indique l’importance relative de chaque variable prédite pour l’estimation du modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique peut être interprété de la même façon que pour les autres modèles qui affichent l’importance des variables prédites, même si dans le cas de MRAA, le graphique est généré par simulation de l’algorithme MRAA. Il suffit de supprimer tout à tour chaque valeur prédite du modèle et d’observer l’impact sur la précision du modèle. [Pour plus d’informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Paramètres du modèle MRAA

L’onglet Paramètres d’un nugget de modèle MRAA affiche les options de modification du modèle créé. Par exemple, vous pouvez utiliser le noeud MRAA pour créer plusieurs modèles différents en utilisant les mêmes données et paramètres, puis utiliser cet onglet dans chaque modèle pour modifier légèrement les paramètres et voir en quoi cela influe sur les résultats.

Remarque : Cet onglet n’est disponible qu’une fois le nugget de modèle ajouté à un flux.

Figure 14-5
Boîte de dialogue du nugget de modèle MRAA, onglet Paramètres

Modèle Paramètres Récapitulatif Annotations

Nombre maximal de prévisions par enregistrement: 2

Niveau de randomisation: 0,00

Définir graine aléatoire: 876547

Ordre de tri:

Décroissant (les offres avec les scores les plus élevés seront renvoyées)

Croissant (les offres avec les scores les plus faibles seront renvoyées)

Préférences pour les champs cible :

Valeur	Préférence	Inclure systématiquement	Ajouter...	Supprimer
Savings	1.0	<input checked="" type="checkbox"/>		

Prendre en compte la fiabilité du modèle

OK Annuler Appliquer Réinitialiser

Nombre maximal de prévisions par enregistrement. Cette option permet de limiter le nombre de prévisions effectuées pour chaque enregistrement d'un ensemble de données. La valeur par défaut est 3.

Par exemple, vous pouvez disposer de six offres (épargne, prêt hypothécaire, prêt automobile, retraite, carte de crédit et assurance) mais ne souhaitez connaître que les deux meilleures recommandations possibles, auquel cas il convient de définir la valeur 2 pour ce champ. Une fois le modèle créé et relié à une table, deux colonnes de prévisions (plus la confiance associée dans la probabilité d'acceptation de l'offre) seront alors visibles pour chaque enregistrement. Les prévisions peuvent comprendre l'une des six offres possibles.

Niveau de randomisation. Pour éviter tout biais, notamment dans le cas d'un ensemble de données peu volumineux ou incomplet, et traiter toutes les offres potentielles de la même manière, vous pouvez ajouter un niveau de randomisation aux offres proposées, ainsi que la probabilité de leur affichage sous forme de recommandations. La randomisation est exprimée en pourcentage sous la forme d'une valeur décimale comprise entre 0,0 (sans randomisation) et 1,0 (randomisation complète). La valeur par défaut est 0,0.

Définir graine aléatoire. Lorsque vous ajoutez un niveau de randomisation à la sélection d'une offre, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les

mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.

Remarque : Lorsque vous utilisez l'option Définir graine aléatoire avec des enregistrements lus à partir d'une base de données, il peut s'avérer nécessaire d'exécuter un noeud Trier avant de procéder à l'échantillonnage afin de garantir le même résultat à chaque exécution du noeud. Cela s'explique par le fait que la graine aléatoire dépend de l'ordre des enregistrements, et qu'il n'est pas garanti que cet ordre reste inchangé dans une base de données relationnelle. [Pour plus d'informations, reportez-vous à la section Noeud Trier dans le chapitre 3 dans Noeuds source, exécution et de sortie de IBM SPSS Modeler 15.](#)

Ordre de tri. Sélectionnez l'ordre d'affichage des offres dans le modèle créé :

- **Décroissant.** Le modèle affiche les offres avec le score le plus élevé en premier. Il s'agit des offres ayant la plus forte probabilité d'être acceptées.
- **Croissant.** Le modèle affiche les offres avec le score le plus faible en premier. Il s'agit des offres ayant la plus forte probabilité d'être rejetées. Cette option peut se révéler utile lorsque vous décidez des clients à exclure d'une campagne de marketing pour une offre spécifique.

Préférences des champs cible. Pendant la création d'un modèle, vous pouvez être amené à favoriser ou à ignorer certains aspects des données. Si, par exemple, vous créez un modèle de sélection de la meilleure offre financière à proposer à un client, il peut être judicieux de veiller à toujours inclure une offre particulière quel que soit son score par client.

Pour inclure une offre dans ce panneau et modifier ses préférences, cliquez sur le bouton Ajouter, saisissez le nom de l'offre (par exemple, Epargne ou Prêt hypothécaire), puis cliquez sur le bouton OK.

- **Valeur.** Il s'agit du nom de l'offre que vous venez d'ajouter.
- **Préférence.** Il s'agit du niveau de préférence à appliquer à l'offre. La préférence est exprimée en pourcentage sous la forme d'une valeur décimale comprise entre 0,0 (sans préférence) et 1,0 (préférence maximale). La valeur par défaut est 0,0.
- **Inclure systématiquement.** Cochez cette case pour inclure systématiquement une offre spécifique dans les prévisions.

Remarque : Si l'option Préférence a pour valeur 0,0, le paramètre Inclure systématiquement n'est pas pris en compte.

Prendre en compte la fiabilité du modèle. Un modèle riche en données, bien structuré et affiné par plusieurs régénérations est censé produire des résultats plus précis qu'un tout nouveau modèle contenant peu de données. Cochez cette case pour bénéficier de la plus grande fiabilité du modèle plus abouti.

Modèles Support Vector Machine

A propos de SVM

Support Vector Machine (SVM) est une technique robuste de classification et de régression qui optimise la précision prédictive d'un modèle sans surajustement des données d'apprentissage. SVM est particulièrement adapté à l'analyse de données avec de très grands nombres (par exemple, des milliers) de champs prédictifs.

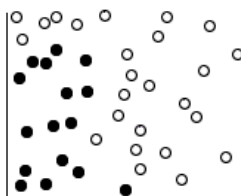
SVM comporte des applications dans de nombreuses disciplines, dont la gestion de la relation client (CRM), reconnaissance faciale et d'autres images, bio-informatique, extraction du concept Text Mining, détection de l'intrusion, prédiction de la structure des protéines et reconnaissance vocale.

Fonctionnement de SVM

SVM fonctionne par mappage des données à un espace d'attributs haute dimension pour que les points de données puissent être classés, même lorsque les données ne sont pas séparables sur un plan linéaire. Un séparateur entre les catégories est identifié. Ensuite, les données sont transformées de sorte que le séparateur puisse être défini comme un hyperplan. Ensuite, les caractéristiques des nouvelles données peuvent être utilisées pour prédire le groupe auquel un nouvel enregistrement doit appartenir.

Par exemple, prenez la figure suivante, dans laquelle les points de données rentrent dans deux catégories différentes :

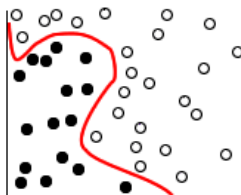
Figure 15-1
Ensemble de données d'origine



Les deux catégories peuvent être séparées par une courbe :

Figure 15-2

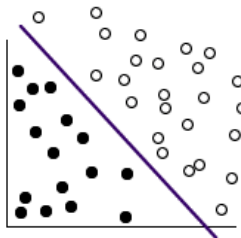
Données avec séparateur ajouté



Une fois la transformation effectuée, la limite entre les deux catégories peut être définie par un hyperplan :

Figure 15-3

Données transformées



La fonction mathématique utilisée pour la transformation est appelée fonction **noyau**. SVM dans IBM® SPSS® Modeler prend en charge les types de noyaux suivants :

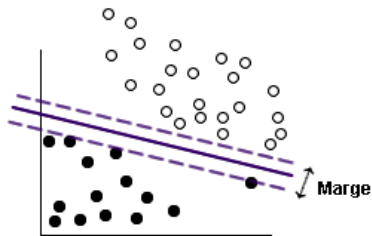
- Linéaire
- Polynomial
- Fonction radiale de base (RBF)
- Sigmoid

Une fonction noyau linéaire est recommandée lorsque la séparation linéaire des données est simple. Dans les autres cas, l'une des autres fonctions doit être utilisée. Vous devrez tester les différentes fonctions pour obtenir le meilleur modèle dans chaque cas, comme ils utilisent des algorithmes et des paramètres différents.

Affinement d'un modèle SVM

Outre la ligne de séparation entre les catégories, un modèle SVM de classification recherche également les lignes marginales qui définissent l'espace entre deux catégories :

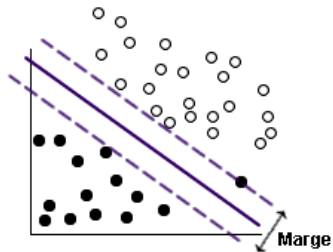
Figure 15-4
Données avec modèle préliminaire



Les points de données qui figurent sur les marges sont appelés **vecteurs de prise en charge**.

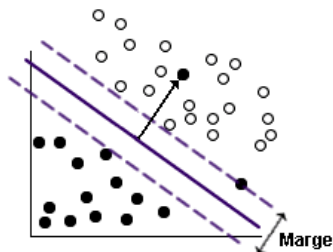
Plus la marge est large entre les deux catégories, plus le modèle sera à même de prédire la catégorie pour les nouveaux enregistrements. Dans l'exemple précédent, la marge n'est pas large, et le modèle est dit **surajusté**. Un classement légèrement incorrect peut être accepté afin d'élargir la marge, par exemple :

Figure 15-5
Données avec modèle amélioré



Dans certains cas, la séparation linéaire est plus difficile, par exemple :

Figure 15-6
Un problème pour la séparation linéaire



Dans un tel cas, l'objectif est de trouver l'équilibre optimal entre une marge large et un petit nombre de points de données mal classés. La fonction noyau possède un **paramètre de régularisation** (appelé C) qui contrôle le compromis entre ces deux valeurs. Vous devrez

probablement tester différentes valeurs et autres paramètres du noyau pour trouver le meilleur modèle.

Noeud SVM

Le noeud SVM permet d'utiliser un modèle support vector machine pour classer les données. SVM est particulièrement adapté à de grands ensembles de données, c'est à dire, ceux avec un grand nombre de champs prédicteurs. Vous pouvez utiliser les paramètres par défaut sur le noeud pour générer relativement rapidement un modèle de base. Vous pouvez aussi utiliser les paramètres Expert pour tester les différents types de modèles SVM.

Lorsque le modèle a été créé, vous pouvez :

- parcourir le nugget du modèle pour afficher l'importance relative des champs d'entrée dans la création du modèle ;
- ajouter un noeud Table au nugget de modèle pour afficher la sortie du modèle.

Exemple : Un chercheur en médecine a obtenu un ensemble de données contenant les caractéristiques d'un certain nombre d'échantillons de cellules humaines supposées favoriser le développement du cancer. L'analyse des données originales indiquait que de nombreuses caractéristiques différaient considérablement entre les échantillons bénins et malins. Le chercheur en médecine souhaite développer un modèle SVM qui peut utiliser les valeurs de caractéristiques de cellules semblables dans des échantillons d'autres patients pour savoir au plus tôt si leurs échantillons peuvent être bénins ou malins. [Pour plus d'informations, reportez-vous à la section Classification des échantillons de cellules \(SVM\) dans le chapitre 25 dans *Guide d'applications de IBM SPSS Modeler 15*.](#)

Noeud SVM - Options modèle

Figure 15-7
Noeud SVM - Options 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.

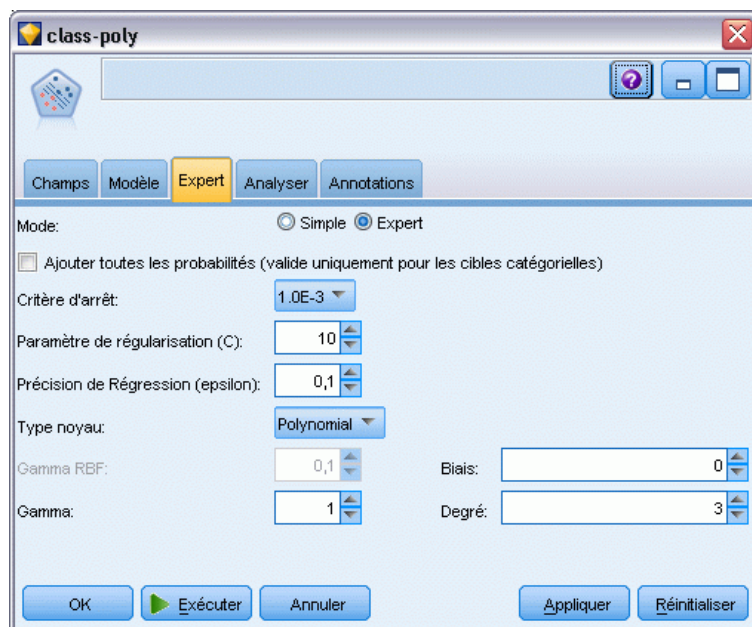
Utiliser les données partitionnées. Si un champ de partition est défini, cette option assure que seules les données provenant de la partition d'apprentissage sont utilisées pour générer le modèle. [Pour plus d'informations, reportez-vous à la section Noeud Partitionner dans le chapitre 4 dans Noeuds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

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Noeud SVM - Options expert

Si vous êtes familier des modèles support vector machines, utilisez les options expert pour affiner le processus d'apprentissage. Pour accéder aux options expert, paramétrez le mode sur Expert dans l'onglet Expert.

Figure 15-8
Noeud SVM - Options expert



Ajouter toutes les probabilités (valide uniquement pour les cibles catégorielles). Si cette option est sélectionnée (cochée), cela indique que les probabilités pour chaque valeur possible d'un champ cible Nominal ou Booléen sont affichées pour chaque enregistrement traité par le noeud. Sinon, la probabilité de la valeur prédite uniquement est affichée pour les champs cible Nominal ou Booléen. Le paramétrage de cette case à cocher détermine l'état par défaut de la case à cocher correspondante sur l'affichage du nugget de modèle.

Critère d'arrêt. Détermine le moment de l'arrêt de l'algorithme d'optimisation. Les valeurs se situent entre $1.0E-1$ et $1.0E-6$; la valeur par défaut est $1.0E-3$. La réduction de la valeur entraîne un modèle plus précis. Toutefois, l'apprentissage de ce dernier est plus long

Paramètre de régularisation (C). Contrôle le compromis entre l'optimisation de la marge et la réduction du terme de l'erreur d'apprentissage. La valeur doit généralement se situer entre 1 et 10 inclus ; la valeur par défaut est 10. L'augmentation de la valeur améliore la précision du classement (ou réduit l'erreur de régression) pour les données d'apprentissage. Cependant, ceci peut également conduire à un surajustement.

Précision de régression (epsilon). Option uniquement utilisée si le niveau de mesure du champ cible est *Continu*. Permet l'acceptation d'erreurs à condition qu'elles soient inférieures à la valeur spécifiée ici. L'augmentation de la valeur risque d'entraîner une modélisation plus rapide, mais aux dépens de la précision.

Type noyau. Détermine le type de fonction noyau utilisé pour la transformation. Des types noyaux différents permettent le calcul du séparateur de différentes manières. Par conséquent, il est recommandé de tester les diverses options. La valeur par défaut est RBF (Fonction radiale de base).

RBF gamma. Option activée uniquement si le type noyau est défini sur RBF. La valeur doit généralement se situer entre $3/k$ et $6/k$, où k est le nombre de champs d'entrée. Par exemple, avec 12 champs d'entrée, les valeurs entre 0,25 et 0,5 valent la peine d'être essayées. L'augmentation de la valeur améliore la précision du classement (ou réduit l'erreur de régression) pour les données d'apprentissage. Cependant, ceci peut également conduire à un surajustement.

Gamma. Option activée uniquement si le type noyau est défini sur Polynomial ou Sigmoidal. L'augmentation de la valeur améliore la précision du classement (ou réduit l'erreur de régression) pour les données d'apprentissage. Cependant, ceci peut également conduire à un surajustement.

Biais. Option activée uniquement si le type noyau est défini sur Polynomial ou Sigmoidal. Définit la valeur `coef0` dans la fonction noyau. La valeur par défaut 0 est appropriée dans la plupart des cas.

Degré. Option activée uniquement si le type noyau est défini sur Polynomial. Contrôle la complexité (dimension) de l'espace de mappage. Généralement, vous n'utilisez pas une valeur supérieure à 10.

Nugget de modèle SVM

Le modèle SVM crée un certain nombre de nouveaux champs. Le plus important de ceux-ci est le champ `$$fieldname`, qui indique la valeur de champ cible prédite par le modèle.

Le nombre et le nom des nouveaux champs créés par le modèle dépendent du niveau de mesure du champ cible (ce champ est indiqué dans les tableaux suivants par *fieldname*).

Pour consulter ces champs et les valeurs correspondantes, ajoutez un noeud Table au nugget du modèle SVM et exécutez le noeud Table.

Table 15-1

Le niveau de mesure du champ cible est « Nominal » ou « Booléen »

Nouveau nom de champ	Description
<code>\$\$fieldname</code>	Valeur prédite du champ cible.
<code>\$\$SP-fieldname</code>	Probabilité de la valeur prédite.
<code>\$\$SP-value</code>	Probabilité de chaque valeur possible de type nominal ou booléen (uniquement affichée si Ajouter toutes les probabilités est cochée dans l'onglet Paramètres du nugget de modèle).
<code>\$\$SRP-value</code>	(Champs cible de type booléen uniquement) Scores de propension brute (SRP) et ajustée (SAP), indiquant la probabilité d'un résultat « true (vrai) » pour le champ cible. Ces scores sont affichés uniquement si les cases correspondantes sont cochées dans l'onglet Analyser du noeud de modélisation SVM avant que le modèle ne soit généré. Pour plus d'informations, reportez-vous à la section Options d'analyse des noeuds de modélisation dans le chapitre 3 sur p. 44.
<code>\$\$SAP-value</code>	

Table 15-2

Le niveau de mesure du champ cible est « Continu »

Nouveau nom de champ	Description
<code>\$\$fieldname</code>	Valeur prédite du champ cible.

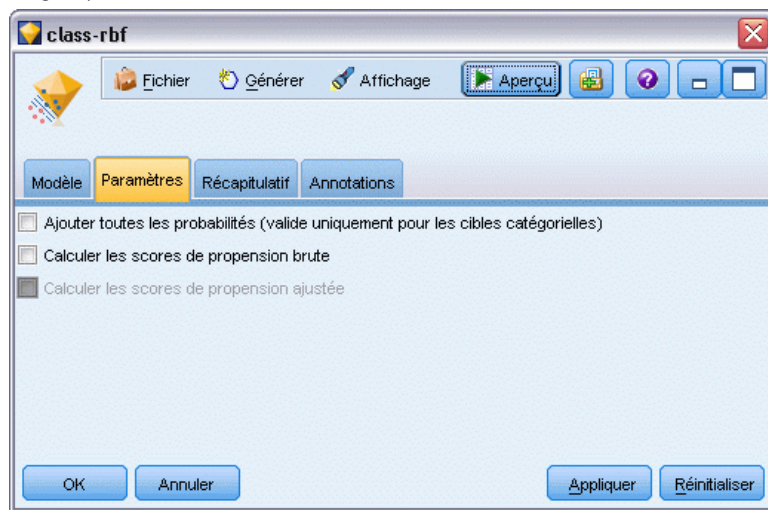
Importance des valeurs prédites

Un graphique illustrant l'importance relative de chaque valeur prédite dans l'estimation du modèle peut également être affiché dans l'onglet Modèle. En général, vous préférerez concentrer vos efforts de modélisation sur les valeurs prédites les plus importantes et abandonner ou ignorer les moins importantes. Ce graphique n'est disponible que si Calculer l'importance des valeurs prédites a été sélectionné dans l'onglet Analyse avant la génération du modèle. [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Remarque : Le calcul de l'importance des valeurs prédites risque de prendre plus longtemps pour SVM que pour d'autres types de modèles. En outre, cette importance n'est pas sélectionnée par défaut dans l'onglet Analyser. Le choix de cette option risque de ralentir les performances, surtout avec de grands ensembles de données.

Paramètres du modèle SVM

Figure 15-9
Onglet paramètres du modèle SVM



L'onglet Paramètres vous permet de spécifier des champs supplémentaires à afficher lors de la consultation des résultats (par exemple en exécutant un noeud Table relié au nugget). Vous pouvez voir l'effet de chacune de ces options en les sélectionnant et en cliquant sur le bouton Aperçu : faites défiler vers la droite de l'aperçu pour voir les champs supplémentaires.

Ajouter toutes les probabilités (valide uniquement pour les cibles catégorielles). Si cette option est cochée, les probabilités pour chaque valeur possible d'un champ cible Nominal ou Booléen sont affichées pour chaque enregistrement traité par le noeud. Sinon, seule la valeur prédite et la probabilité correspondante sont affichées pour les champs cible de type nominal ou booléen.

Le paramètre par défaut de cette case à cocher est déterminé par la case correspondante sur le noeud de modélisation.

Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Modèles d'agrégation suivant le saut minimum

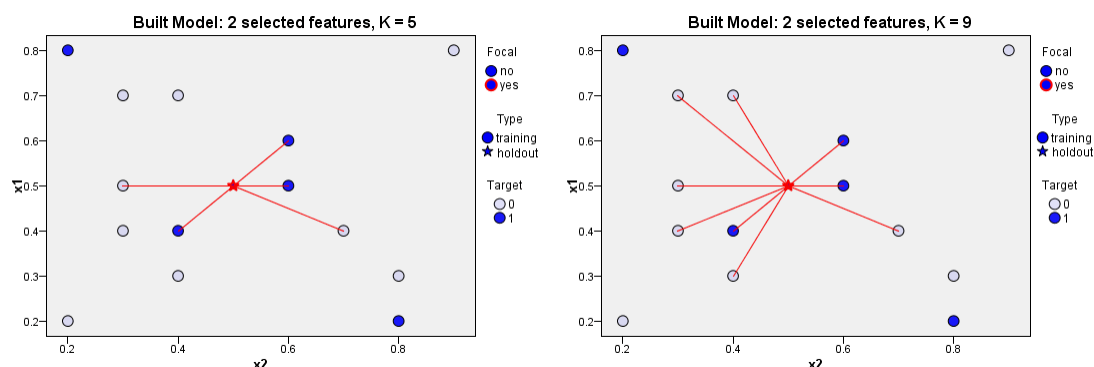
Noeud KNN

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.

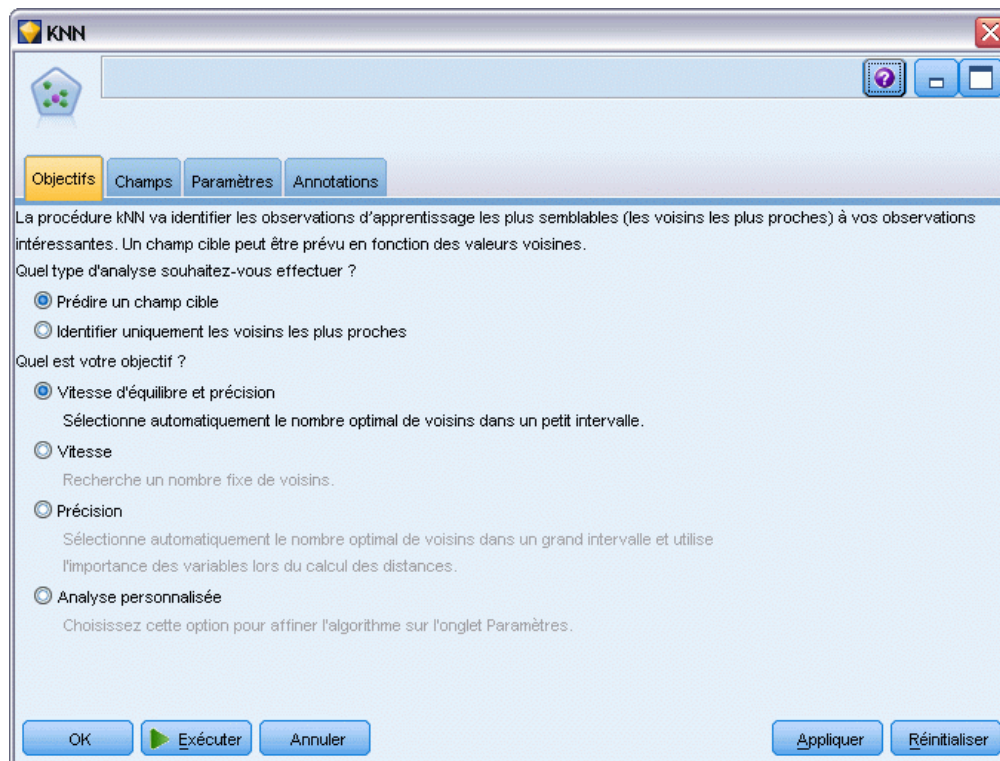
Figure 16-1
Les effets de k sur la classification



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.

Noeud KNN - Options des objectifs

Figure 16-2
Noeud KNN - Options des objectifs



Dans l'onglet Objectifs vous pouvez choisir de créer un modèle qui prédit la valeur d'un champ cible dans vos données d'entrée en fonction des valeurs de ses voisins les plus proches ou de simplement rechercher les voisins les plus proches pour une observation intéressante particulière.

Quel type d'analyse souhaitez-vous effectuer ?

Prédire un champ cible. Choisissez cette option si vous voulez prédire la valeur d'un champ cible en fonction des valeurs de ses voisins les plus proches.

Identifier uniquement les voisins les plus proches. Choisissez cette option si vous souhaitez simplement connaître les voisins les plus proches d'un champ d'entrée particulier.

Si vous choisissez de n'identifier que les voisins les plus proches, les options restantes de cet onglet en rapport avec la précision et la vitesse sont désactivées car elles ne sont pertinentes que pour des cibles de prévision.

Quel est votre objectif ?

Lors de la prévision de champs cibles, ce groupe d'options vous permet de décider si la vitesse, la précision, ou un mélange des deux, sont les facteurs les plus importants lors de la prévision d'un champ cible. Vous pouvez également choisir de personnaliser vous-même les paramètres.

Si vous sélectionnez l'option Equilibrer, Vitesse ou Précision, l'algorithme présélectionne la combinaison de paramètres la plus appropriée pour cette option. Les utilisateurs avancés peuvent remplacer ces sélections ; cela est possible via les divers panneaux de l'onglet Paramètres.

Équilibrer la vitesse et la précision. Sélectionne le nombre le plus approprié de voisins dans un intervalle réduit.

Vitesse. Recherche un nombre fixe de voisins.

Précision. Sélectionne le nombre le plus approprié de voisins au sein d'un intervalle important et utilise l'importance des valeurs prédites lors du calcul des distances.

Analyse personnalisée. Choisissez cette option pour affiner l'algorithme sur l'onglet Paramètres.

Remarque : La taille du modèle KNN obtenu, à la différence de la plupart des autres modèles, augmente de manière linéaire en fonction de la quantité de données d'apprentissage. Si vous voyez une erreur due à une insuffisance de mémoire, lors d'une tentative de création d'un modèle KNN, essayez d'augmenter la mémoire système maximum utilisée par IBM® SPSS® Modeler. Pour ce faire, sélectionnez :

Outils > Options > Options du système

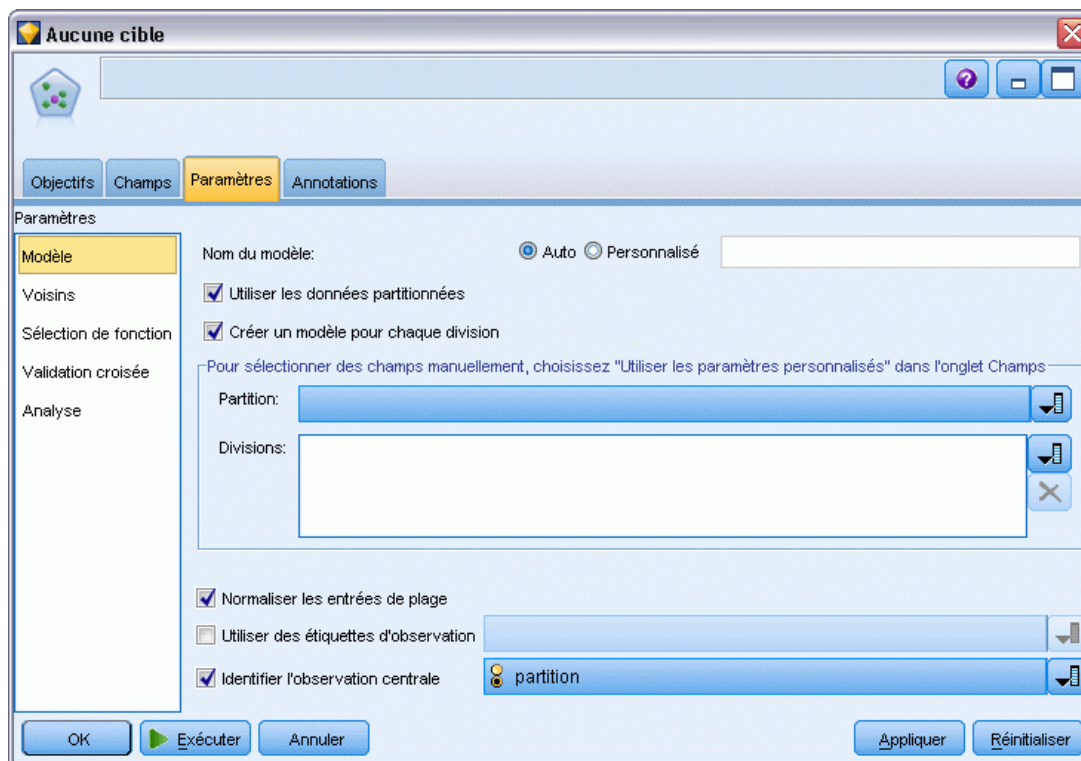
et saisissez la taille dans le champ Mémoire maximale. Les changements effectués dans la boîte de dialogue Options système ne prennent effet qu'une fois SPSS Modeler redémarré.

Paramètres de noeud KNN

Vous pouvez spécifier, dans l'onglet Paramètres, les options spécifiques à l'analyse d'agrégation suivant le saut minimum. La barre latérale à gauche de l'écran répertorie les panneaux que vous utilisez pour spécifier les options.

Modèle

Figure 16-3
Noeud KNN - Options du modèle



Le panneau Modèle fournit des options qui commandent la façon dont le modèle doit être créé. Par exemple, l'utilisation ou non de la partition ou de la découpe de modèles, la transformation ou non de champs d'entrée numériques afin qu'ils tombent tous dans le même intervalle et la façon de gérer les observations intéressantes. Vous pouvez aussi choisir un nom personnalisé pour le 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.](#)

Créer des modèles divisés. Crée un modèle séparé pour chaque valeur possible des champs d'entrée spécifiés en tant que champs de découpage. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Pour sélectionner des champs manuellement... Par défaut, le noeud utilise les paramètres de partition et de découpage de champ (le cas échéant) du noeud Typer, mais vous pouvez remplacer ces paramètres à ce stade. Pour activer les champs Partition et Découpage, sélectionnez l'onglet Champs puis sélectionnez Utiliser les paramètres personnalisés, puis revenez ici.

- **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 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.)
- **Découpage.** Pour des modèles découpés, sélectionnez le ou les champs de découpage. Cela revient à définir le rôle du champ sur la valeur *Découper* dans un noeud Typer. Vous pouvez désigner uniquement les champs de type Booléen, Nominal ou Ordinal comme champs de découpage. Les champs sélectionnés en tant que champ de découpage ne peuvent pas être utilisés comme champs de cible, d'entrée, de partition, de fréquence ou de pondération. [Pour plus d'informations, reportez-vous à la section Création de modèles découpés dans le chapitre 3 sur p. 33.](#)

Normaliser les entrées de plage. Cochez cette case pour normaliser les valeurs des champs d'entrée continus. Les fonctions normalisées ont le même intervalle de valeur, ce qui peut améliorer la performance de l'algorithme d'estimation. La normalisation ajustée, $[2*(x-\min)/(\max-\min)]-1$, est utilisée. Les valeurs normalisées ajustées sont comprises entre -1 et 1.

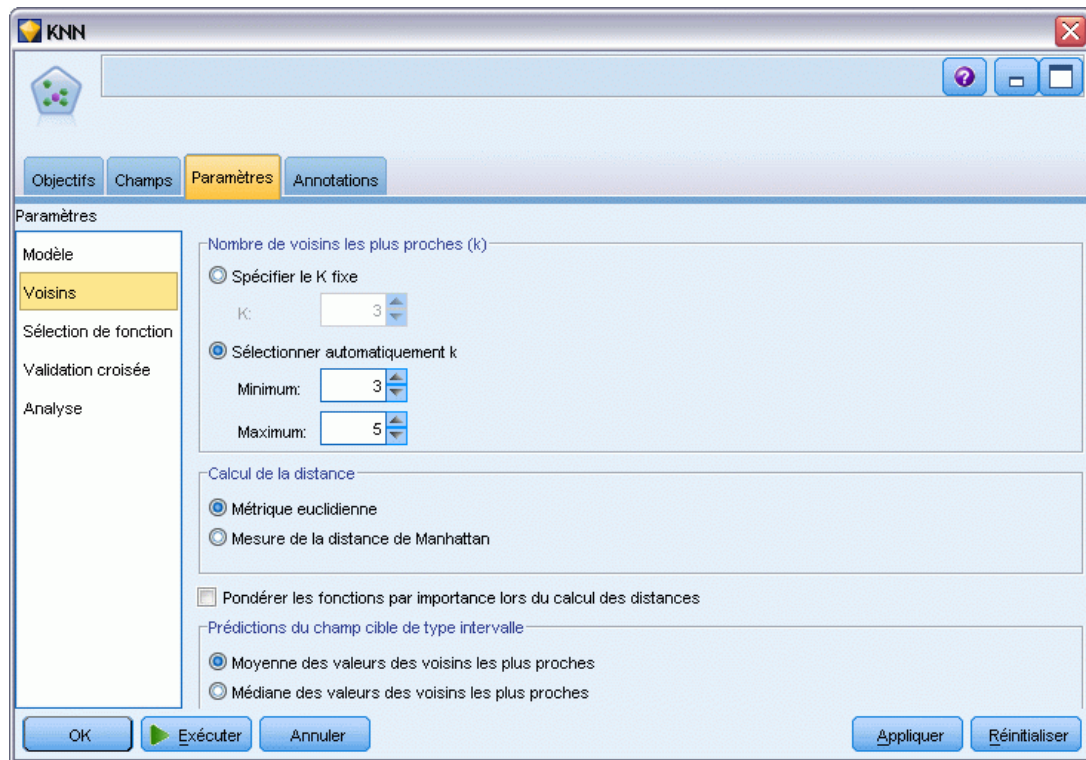
Utiliser des étiquettes d'observation. Cochez cette case pour activer la liste déroulante, à partir de laquelle vous pouvez choisir un champ dont les valeurs seront utilisées comme étiquettes d'identification des observations intéressantes dans le diagramme de l'espace des prédicteurs, le diagramme des pairs, et la carte de cadran dans le Visualiseur de modèles. Vous pouvez choisir un des champs ayant un niveau de mesure *Nominal*, *Ordinal* ou *Booléen* à utiliser en tant que le champ d'étiquette. Si vous ne sélectionnez aucun champ, les enregistrements sont affichés dans les diagrammes du Visualiseur de modèles en fonction des plus proches voisins identifiés par numéro de ligne dans les données source. Si vous devez manipuler les données après la construction du modèle, utilisez des étiquettes d'observation afin d'éviter d'avoir à vous reporter aux données source à chaque fois afin d'identifier les observations affichées.

Identifier les enregistrements centraux. Cochez cette case pour activer la liste déroulante qui vous permet de marquer un champ d'entrée intéressant (uniquement pour les champs booléens). Si vous spécifiez un champ à ce stade, les points représentant ce champ sont sélectionnés à l'origine dans le Visualiseur de modèles lorsque le modèle est construit. La sélection d'un enregistrement

central à ce stade est facultative ; tout point peut temporairement devenir un enregistrement central lorsqu'il est sélectionné manuellement dans le Visualiseur de modèles.

Voisins

Figure 16-4
Noeud KNN - Options des voisins



Le panneau Voisins dispose d'un ensemble d'options qui commande la façon dont le nombre de voisins les plus proches est calculé.

Nombre de voisins les plus proches (k). Spécifiez le nombre de voisins les plus proches pour une observation particulière. Remarque : l'utilisation d'un nombre élevé de voisins ne garantit pas forcément un modèle plus précis.

Si l'objectif consiste à prévoir une cible, deux choix s'offrent à vous :

- **Spécifier le k fixe.** Utilisez cette option si vous souhaitez spécifier un nombre fixe de voisins les plus proches à rechercher.
- **Sélectionner automatiquement k.** Vous pouvez aussi utiliser les champs Minimum et Maximum afin de spécifier un intervalle de valeurs et d'autoriser la procédure à sélectionner le nombre "le plus approprié" de voisins dans cet intervalle. La méthode de détermination du nombre de

voisins les plus proches varie selon que la sélection de fonction a été demandée ou non sur le panneau Sélection de fonction :

Si oui, la sélection des caractéristiques sera alors exécutée pour chaque valeur de k dans l'intervalle requis, et le k ainsi que l'ensemble des caractéristiques l'accompagnant, avec le taux d'erreur le plus faible (ou l'erreur de la somme des carrés la plus faible si la cible est continue), seront sélectionnés.

Si la sélection des caractéristiques n'est pas activée, alors la validation croisée de niveau V sera utilisée pour sélectionner le nombre de voisins "optimal". Consultez le panneau Validation croisée pour contrôler l'affectation des validations croisées.

Calcul de la distance. Il s'agit de la métrique employée pour spécifier la distance métrique utilisée dans la mesure de la similarité des observations.

- **Métrique euclidienne.** La distance entre deux observations, x et y , est la racine carrée de la somme, sur toutes les dimensions, des carrés des différences entre les valeurs de ces observations.
- **Mesure de la distance de Manhattan.** La distance entre deux observations est la somme, sur toutes les dimensions, des différences absolues entre les valeurs de ces observations. Appelée également distance City Block.

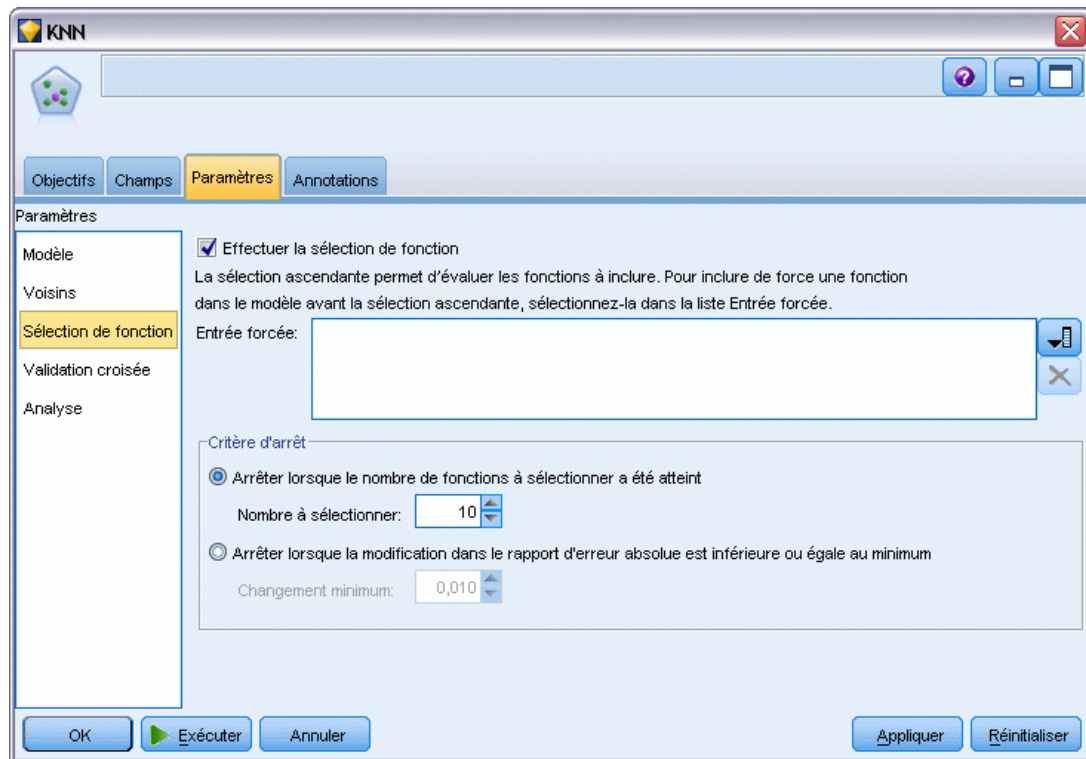
Si l'objectif consiste à prévoir une cible, vous pouvez également pondérer les fonctions par leur importance normalisée lors du calcul des distances. L'importance des fonctions pour une valeur prédite est calculée par le rapport du taux d'erreur ou de la somme d'erreurs des carrés du modèle sans la valeur prédite, sur le taux d'erreur ou la somme d'erreurs des carrés du modèle complet. L'importance normalisée est calculée par nouvelle pondération des valeurs d'importance des caractéristiques de sorte que leur somme soit égale à 1.

Pondérer les fonctions par importance lors du calcul des distances. (Affichées uniquement si l'objectif consiste à prévoir une cible.) Cochez cette case pour que l'importance des valeurs prédites soit utilisée lors du calcul des distances entre voisins. L'importance des valeurs prédites sera ensuite affichée dans le nugget de modèle, et utilisée dans les prévisions (et affectera ainsi le scoring). [Pour plus d'informations, reportez-vous à la section Importance des valeurs prédites dans le chapitre 3 sur p. 57.](#)

Prédictions du champ cible de type intervalle. (Affichées uniquement si l'objectif consiste à prévoir une cible.) Si une cible continue (intervalle numérique) est spécifiée, celle-ci définit si la valeur prédite est calculée en fonction de la moyenne ou de la valeur médiane des voisins les plus proches.

Sélection des caractéristiques

Figure 16-5
Noeud KNN - Options de sélection de fonction



Ce panneau est activé uniquement si l'objectif consiste à prévoir une cible. Il vous permet de demander et spécifier des options pour la sélection de fonction. Par défaut, toutes les caractéristiques sont prises en compte pour la sélection de caractéristiques, mais vous pouvez également sélectionner un sous-ensemble de caractéristiques à introduire de force dans le modèle.

Exécuter la sélection des caractéristiques. Cochez cette case pour activer les options de la sélection de fonction.

- **Entrée forcée.** Cliquez sur le sélecteur de champ en regard de cette zone et choisissez une ou plusieurs fonctions à inclure de force dans le modèle.

Critère d'arrêt. À chaque étape, la caractéristique dont l'addition au modèle entraîne l'erreur la plus faible (calculée comme le taux d'erreur pour une cible qualitative et l'erreur de la somme des carrés pour une cible continue) est prise en compte afin d'être incluse dans l'ensemble de modèle. La sélection ascendante se poursuit jusqu'à la rencontre de la condition spécifiée.

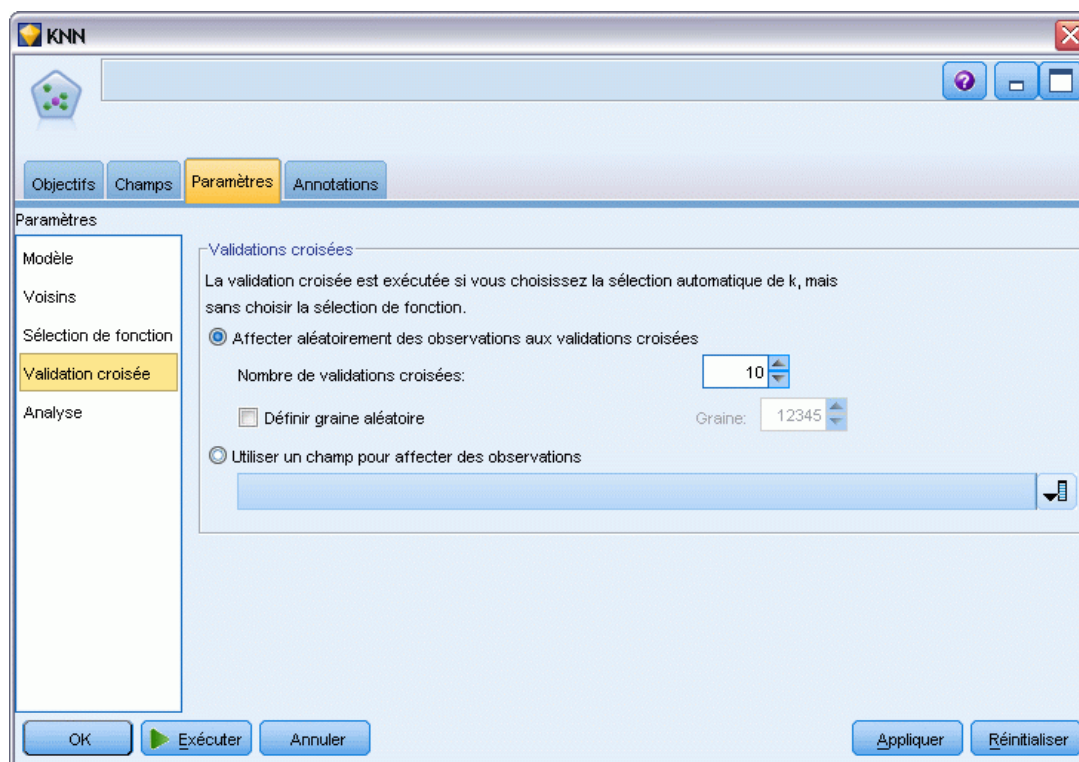
- **Arrêter lorsque le nombre de fonctions à sélectionner a été atteint.** L'algorithme ajoute un nombre fixe de caractéristiques en plus de celles introduites de force dans le modèle. Spécifiez un nombre entier positif. La diminution des valeurs du nombre à sélectionner produit un modèle plus réduit, au risque d'un manque de caractéristiques importantes. L'augmentation

des valeurs du nombre à sélectionner capturera toutes les caractéristiques importantes, au risque d'ajouter des caractéristiques qui en réalité alimentent l'erreur du modèle.

- **Arrêter lorsque la modification dans le rapport d'erreur absolue est inférieure ou égale au minimum.** L'algorithme prend fin lorsque le changement dans le ratio d'erreur absolue indique que le modèle ne peut pas être davantage amélioré par l'ajout de nouvelles caractéristiques. Indiquez un nombre positif. La diminution de la valeur de l'incrément minimal aura tendance à inclure davantage de fonctions, au risque d'inclure des fonctions qui n'ajoutent pas beaucoup de valeur au modèle. L'augmentation de la valeur du changement minimal aura tendance à exclure davantage de caractéristiques, au risque de perdre des caractéristiques importantes pour le modèle. La valeur "optimale" du changement minimal dépendra de vos données et de l'application. Reportez-vous au Journal d'erreur de sélection des caractéristiques pour pouvoir déterminer quelles sont les caractéristiques les plus importantes. [Pour plus d'informations, reportez-vous à la section Journal d'erreur de sélection des valeurs prédites sur p. 526.](#)

Validation croisée

Figure 16-6
Noeud KNN - Options de validation croisée



Ce panneau est activé uniquement si l'objectif consiste à prévoir une cible. Les options de ce panneau contrôlent l'utilisation ou non de la validation croisée lors du calcul des voisins les plus proches.

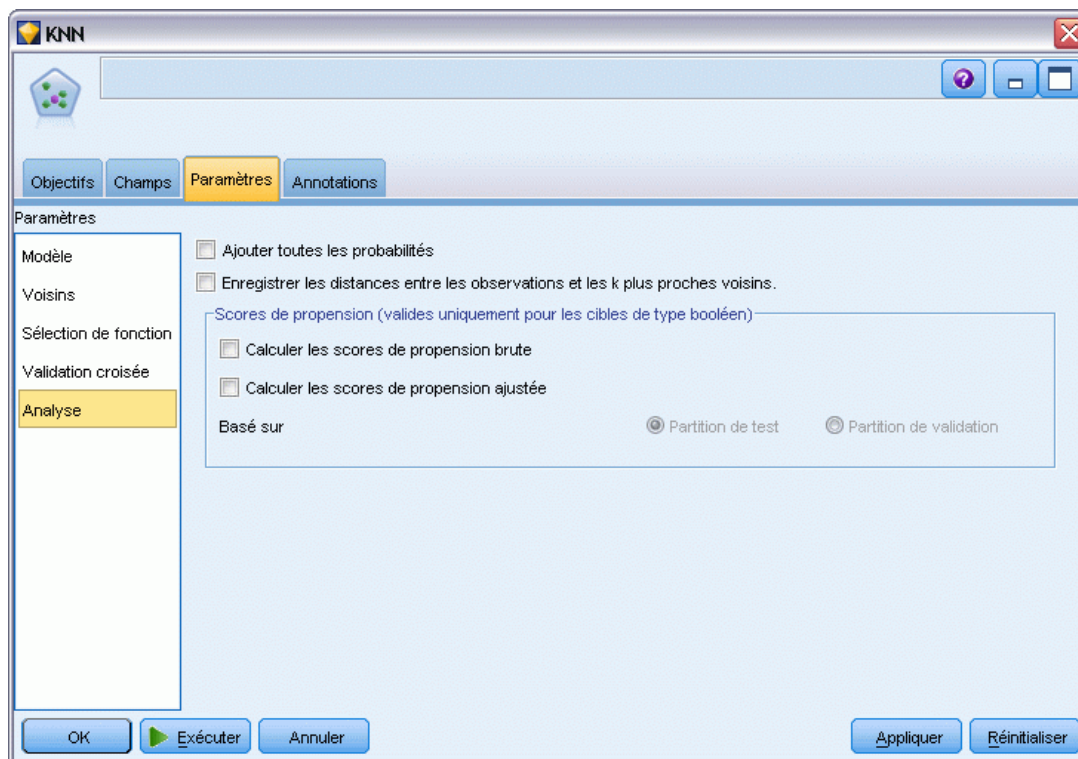
La validation croisée divise l'échantillon en plusieurs sous échantillons, ou **niveaux**. Les modèles du voisin le plus proche sont générés en excluant à tour de rôle les données de chaque sous-échantillon. Le premier modèle est basé sur toutes les observations à l'exception de celles du premier sous-échantillon, le deuxième modèle est basé sur toutes les observations à l'exception de celles du deuxième sous-échantillon, etc. L'erreur est estimée pour chaque modèle en appliquant le modèle au sous-échantillon exclu lors de la génération du modèle. Le "meilleur" nombre des voisins les plus proches est celui qui produit l'erreur la plus faible sur les sous-échantillons.

Validations croisées. Le Niveau V de validation croisée est utilisé pour déterminer le "meilleur" nombre de voisins. Il n'est pas disponible en association avec la sélection de caractéristiques pour des raisons de performance.

- **Affecter aléatoirement des observations aux niveaux.** Spécifier le nombre de niveaux à utiliser pour la validation croisée. Cette procédure affecte aléatoirement des observations aux sous-échantillons, numérotés de 1 à V , le nombre de sous-échantillons.
- **Définir graine aléatoire.** Lors de l'estimation de la précision d'un modèle basé sur un pourcentage aléatoire, cette option vous permet de dupliquer les mêmes résultats dans une autre session. Indiquez la valeur de départ utilisée par le générateur de nombres aléatoires pour vous assurer que les mêmes enregistrements sont affectés à chaque exécution du noeud. Entrez la valeur de graine souhaitée. Si cette option n'est pas sélectionnée, un échantillon différent est généré à chaque exécution du noeud.
- **Utiliser un champ pour affecter des observations.** Indiquez un champ numérique qui affecte chaque observation de l'ensemble de données actif à un niveau. Le champ doit être numérique et prendre des valeurs de 1 à V . Si des valeurs de cet intervalle sont manquantes, et sont situées sur des découpages si des modèles de découpage sont appliqués, cela provoquera une erreur.

Analyse

Figure 16-7
Noeud KNN - Options d'analyse



Le panneau Analyser est activé uniquement si l'objectif consiste à prévoir une cible. Vous pouvez l'utiliser pour spécifier si le modèle doit inclure des variables supplémentaires pour contenir :

- des probabilités pour chaque valeur de champ cible possible
- des distances entre une observation et ses voisins les plus proches
- des scores de propension brute et ajustée (uniquement pour des cibles de type booléen).

Ajouter toutes les probabilités. Si cette option est cochée, les probabilités pour chaque valeur possible d'un champ cible Nominal ou Booléen sont affichées pour chaque enregistrement traité par le noeud. Sinon, seule la valeur prédite et la probabilité correspondante sont affichées pour les champs cible de type nominal ou booléen.

Enregistrer les distances entre les observations et les k plus proches voisins. Pour chaque enregistrement central, une variable séparée est créée pour chacun des k plus proches voisins des enregistrements centraux (à partir de l'échantillon d'apprentissage) et les k plus proches distances correspondantes.

Scores de propension

Des scores de propension peuvent être établis dans le noeud de modélisation et dans l'onglet Paramètres du nugget de modèle. Cette fonctionnalité n'est disponible que lorsque la cible sélectionnée est un champ booléen. [Pour plus d'informations, reportez-vous à la section Scores de propension dans le chapitre 3 sur p. 47.](#)

Calculer les scores de propension brute. Les scores de propension brute sont calculés à partir du modèle uniquement en fonction des données d'apprentissage. Si le modèle prédit la valeur *true (vrai)* (que cette valeur va être la réponse), alors la propension est identique à P, où P est la probabilité de la prédiction. Si le modèle prédit la valeur *false (faux)*, alors la propension est calculée comme $(1 - P)$.

- Si vous choisissez cette option lors de la construction du modèle, des scores de propension sont activés par défaut dans le nugget du modèle. Cependant, vous pouvez toujours choisir d'activer des scores de propension brute dans le nugget du modèle que vous les sélectionnez ou non dans le noeud de modélisation.
- Lors du scoring du modèle, des scores de propension brute seront ajoutés dans un champ avec les lettres *RP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RRP-churn*.

Calculer les scores de propension ajustée. Les propensions brutes sont uniquement basées sur des estimations données par le modèle, qui peut être surajusté, ce qui entraîne des estimations trop optimistes de la propension. Les propensions ajustées tentent de compenser en vérifiant comment le modèle se comporte sur des partitions de test ou de validation et en ajustant les propensions pour donner une meilleure estimation en conséquence.

- Ce paramètre nécessite qu'un champ de partition valide soit présent dans le flux. [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*.](#)
- A la différence des scores de confiance brute, les scores de propension ajustée doivent être calculés lors de la construction du modèle ; sinon, ils ne seront pas disponibles lors du scoring du nugget du modèle.
- Lors du scoring du modèle, des scores de propension ajustée seront ajoutés dans un champ avec les lettres *AP* ajoutées au préfixe standard. Par exemple, si les prédictions se trouvent dans un champ nommé *\$R-churn*, le nom du champ de score de propension est *\$RAP-churn*. Les scores de propension ajustée ne sont pas disponibles pour des modèles de régression logistique.
- Lors du calcul des scores de propension ajustée, la partition de test ou de validation utilisée pour le calcul ne doit pas avoir été équilibrée. Pour éviter cela, assurez-vous que l'option Equilibrer uniquement les données d'apprentissage est sélectionnée dans tous les noeuds Equilibrer en amont. [Pour plus d'informations, reportez-vous à la section Définition des options du noeud Equilibrer dans le chapitre 3 dans *Noeuds source, exécution et de sortie de IBM SPSS Modeler 15*.](#) De plus, si un échantillon complexe a été pris en amont, cela invalide les scores de propension ajustée.
- Les scores de propension ajustée ne sont pas disponibles pour les modèles d'arbre "amélioré" et les modèles d'ensemble de règles. [Pour plus d'informations, reportez-vous à la section Modèles C5.0 améliorés dans le chapitre 6 sur p. 193.](#)

En fonction de. Pour que les scores de propension ajustée soient calculés, un champ de partition doit être présent dans le flux. Vous pouvez spécifier s'il faut utiliser la partition de test ou de validation pour ce calcul. Pour de meilleurs résultats, la partition de test ou de validation doit comprendre au moins autant d'enregistrements que la partition utilisée pour l'apprentissage du modèle original. [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.](#)

Modèle de nugget KNN

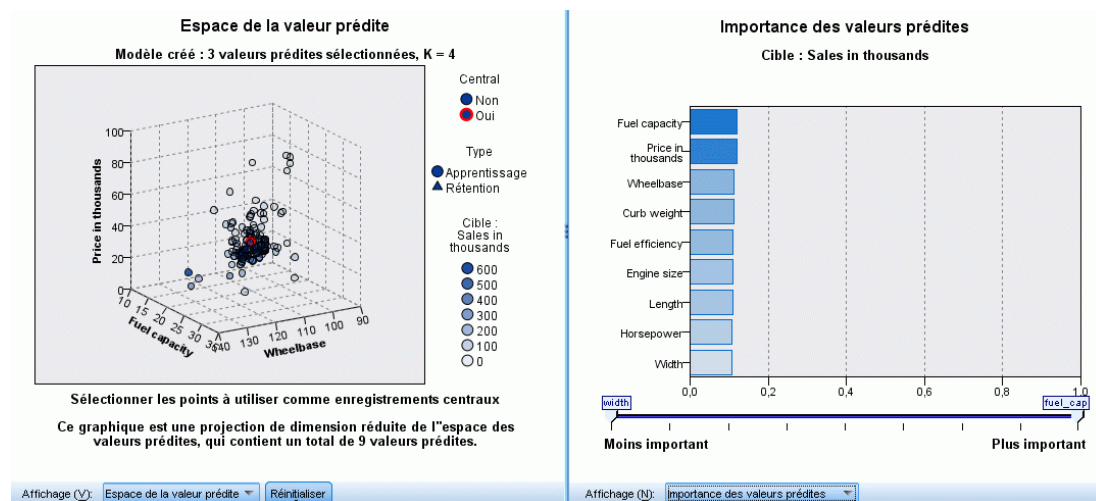
Le modèle KNN crée plusieurs nouveaux champs, comme illustré dans la table suivante. Pour consulter ces champs et les valeurs correspondantes, ajoutez un noeud Table au nugget du modèle KNN et exécutez le noeud Table ou cliquez sur le bouton Aperçu du nugget.

Table 16-1
Champs de modèle KNN

Nouveau nom de champ	Description
\$KNN-fieldname	Valeur prédite du champ cible.
\$KNNP-fieldname	Probabilité de la valeur prédite.
\$KNNP-value	Probabilité de chaque valeur possible d'un champ nominal ou booléen. Inclus uniquement si Ajouter toutes les probabilités est coché dans l'onglet Paramètres du nugget de modèle.
\$KNN-neighbor-n	Le nom du <i>nième</i> voisin le plus proche de l'enregistrement central. Inclus uniquement si Afficher le plus proche de l'onglet Paramètres du nugget de modèle est configuré sur une valeur différente de zéro.
\$KNN-distance-n	La distance relative de l'enregistrement central au <i>nième</i> voisin le plus proche de l'enregistrement central. Inclus uniquement si Afficher le plus proche de l'onglet Paramètres du nugget de modèle est configuré sur une valeur différente de zéro.

Vue du modèle

Figure 16-8
Vue du modèle de l'analyse du voisin le plus proche



Le modèle présente une fenêtre à double panels :

- Le premier affiche une présentation du modèle, appelée vue principale.
- Le second affiche un des deux types de vues :

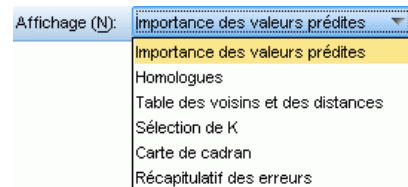
Une vue de modèle auxiliaire affiche davantage d'informations sur le modèle, mais n'est pas focalisée sur le modèle lui-même.

Une vue liée est un affichage montrant les détails d'une caractéristique du modèle lorsque l'utilisateur fait défiler une partie de la vue principale.

Par défaut, le premier panel affiche l'espace du prédicteur et le second le diagramme d'importance des variables prédites. Si le diagramme d'importance des valeurs prédites n'est pas disponible ; c'est-à-dire, lorsque l'option Pondérer les fonctions par importance n'a pas été sélectionnée sur le panneau Voisins de l'onglet Paramètres, la première vue disponible dans la liste déroulante Affichage est affichée.

Figure 16-9

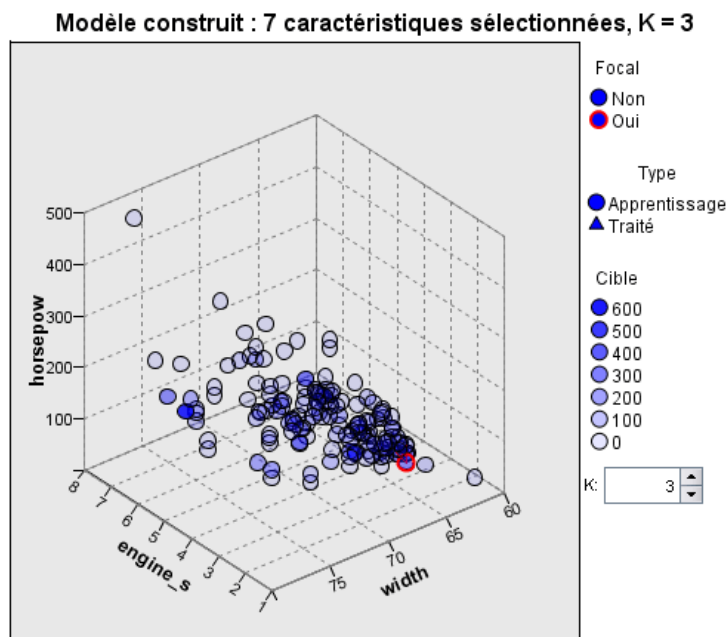
Liste déroulante Affichage de l'analyse du voisin le plus proche



Lorsqu'une vue n'a pas d'informations disponibles, elle est omise de la liste déroulante Affichage.

Espace du prédicteur

Figure 16-10
Espace du prédicteur



Le diagramme d'espace du prédicteur est un diagramme interactif de l'espace du prédicteur (ou un sous-espace, s'il existe plus de 3 variables prédites). Chaque axe représente une variable prédite du modèle, et l'emplacement des points dans le diagramme montre la valeur de ces variables prédites pour des observations dans les partitions de formation et traitée.

Clés. En plus des valeurs des variables prédites, les points du diagramme contiennent d'autres informations.

- La forme indique la partition à laquelle appartient un point, Formation ou Traité.
- La couleur/ombrage d'un point indique la valeur de la cible pour cette observation, avec les valeurs de couleur distinctes correspondant aux modalités d'une cible qualitative, et les ombres indiquant l'intervalle des valeurs d'une cible continue. La valeur indiquée pour la partition de formation est la valeur observée. Pour la partition traitée, il s'agit de la valeur prévue. Si aucune cible n'est spécifiée, la clé ne s'affiche pas.
- Les contours plus épais indiquent que l'observation est focale. Les enregistrements centraux sont affichés avec un lien vers les k voisins les plus proches.

Commandes et interactivité. Un certain nombre de commandes dans le graphique vous permettent d'explorer l'espace du prédicteur.

- Vous pouvez choisir quel sous-ensemble de variables prédites vous souhaitez afficher dans le diagramme et modifier les variables prédites à représenter dans les dimensions.

- Les “ enregistrements centraux ” sont tout simplement des points sélectionnés dans le diagramme de l’espace du prédicteur. Si vous avez spécifié une variable d’enregistrement central, les points représentant les enregistrements centraux seront sélectionnés dès le début. Cependant, tous les points peuvent devenir temporairement un enregistrement central si vous les sélectionnez. Les commandes “habituelles” pour la sélection des points sont appliquées. Cliquer sur un point permet de sélectionner ce point et de désélectionner tous les autres. Cliquer sur un point avec la touche Ctrl enfoncée ajoute ce point à l’ensemble des points sélectionnés. Les vues liées, tels que le diagramme des pairs, sont automatiquement mis à jour en fonction des observations sélectionnées dans l’espace du prédicteur.
- Vous pouvez modifier le nombre de voisins les plus proches (k) à afficher pour les enregistrements centraux.
- Positionner le curseur sur un point du diagramme affiche une note d’aide avec la valeur de l’étiquette d’observation, ou le nombre d’observations si les étiquettes d’observation ne sont pas définies, et les valeurs des cibles observées et prévues.
- Un bouton “Réinitialiser” vous permet de revenir à l’Espace du prédicteur à son état d’origine.

Modification des axes du diagramme d’espace du prédicteur

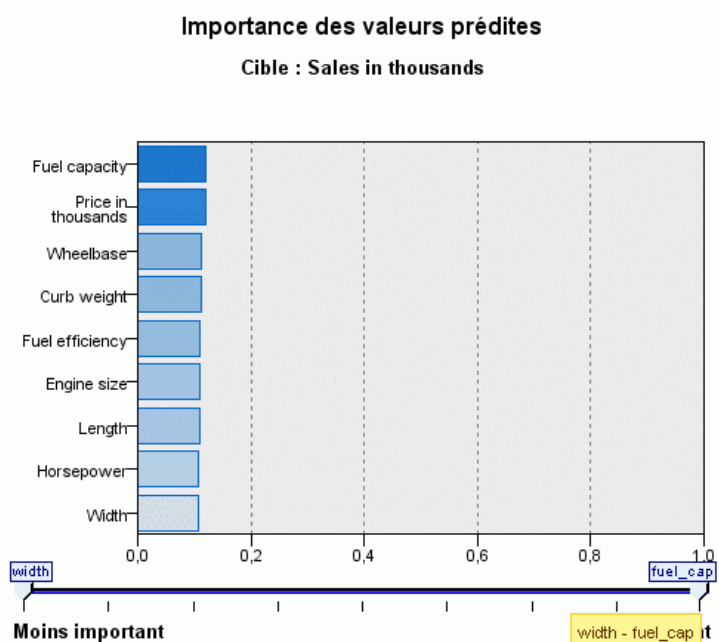
Vous pouvez contrôler les fonctions à afficher sur les axes du diagramme de l’espace du prédicteur.

Pour modifier les paramètres des axes :

- ▶ Cliquez sur le bouton Mode d’édition (icône du pinceau) dans le panneau de gauche pour sélectionner le mode d’édition de l’espace du prédicteur.
- ▶ Modifiez la vue (sur ce que vous voulez) dans le panneau de droite. Le panneau Afficher les zones apparaît entre les deux panneaux principaux.
- ▶ Cochez la case Afficher les zones.
- ▶ Cliquez sur un point de données dans l’espace du prédicteur.
- ▶ Pour remplacer un axe par une variable prédite du même type de données :
Faites glisser la nouvelle variable prédite sur l’étiquette de zone (celle avec le petit bouton X) de la fonction que vous souhaitez remplacer.
- ▶ Pour remplacer un axe par une variable prédite d’un type de données différent :
 - Dans l’étiquette de zone de la variable prédite à remplacer, cliquez sur le petit bouton X. L’espace du prédicteur passe en vue bidimensionnelle.
 - Faites glisser la nouvelle variable prédite sur l’étiquette de zone Ajouter une dimension.
- ▶ Cliquez sur le bouton du mode d’interaction (icône de pointe de flèche) dans le panneau de gauche pour quitter le mode d’édition.

Importance des valeurs prédites

Figure 16-11
Importance des valeurs prédites



En général, vous préférerez concentrer vos efforts de modélisation sur les champs prédictifs les plus importants et abandonner ou ignorer les moins importants. Le diagramme d'importance des valeurs prédites peut vous y aider en indiquant l'importance relative de chaque valeur prédite en estimant le modèle. Etant donné que les valeurs sont relatives, la somme des valeurs pour l'ensemble des valeurs prédites affichée est 1,0. L'importance des des valeurs prédites n'a aucun rapport avec la précision du modèle. Elle est juste rattachée à l'importance de chaque valeur prédite pour la réalisation d'une prévision, peu importe si cette dernière est précise ou non.

Distances du voisin le plus proche

Figure 16-12
Distances du voisin le plus proche

k voisins les plus proches et distances
Affiché pour les enregistrements centraux initiaux

Enregistrement central	Voisins les plus proches				Distances les plus proches			
	1	2	3	4	1	2	3	4
66	104	2	71	139	0,042	0,057	0,079	0,080

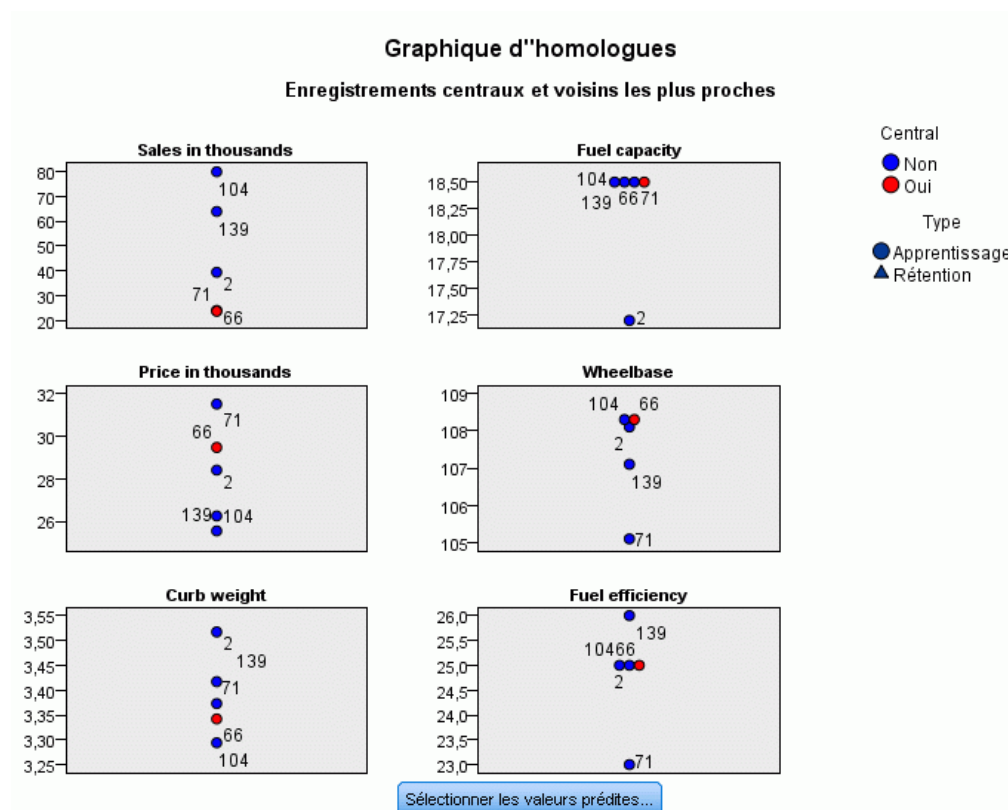
Ce tableau affiche les k voisins les plus proches et les distances pour les enregistrement centraux uniquement. Il est disponible si un identificateur d'enregistrement central est spécifié sur le nœud de modélisation, et n'affiche que les enregistrement centraux identifiés par cette variable.

Chaque ligne de :

- La colonne Enregistrement central contient la valeur de la variable d'étiquetage des observations pour l'enregistrement central. Si les étiquettes d'observations ne sont pas définies, cette colonne contient le nombre d'observations de l'enregistrement central.
- La i ème colonne sous le groupe Voisins les plus proches contient la valeur de la variable étiquetant le i ème voisin le plus proche de l'enregistrement central ; s'il n'y a pas d'étiquettes définies, cette colonne contient le numéro du i ème voisin le plus proche de l'enregistrement central.
- La i ème colonne sous le groupe Distances les plus proches contient la distance du i ème voisin le plus proche de l'enregistrement central

Pairs

Figure 16-13
Diagramme des paires



Ce diagramme affiche les observations focales et leurs k voisins les plus proches sur chaque variable prédite et sur la cible. Il est disponible si une observation focale est sélectionnée dans l'espace du prédicteur.

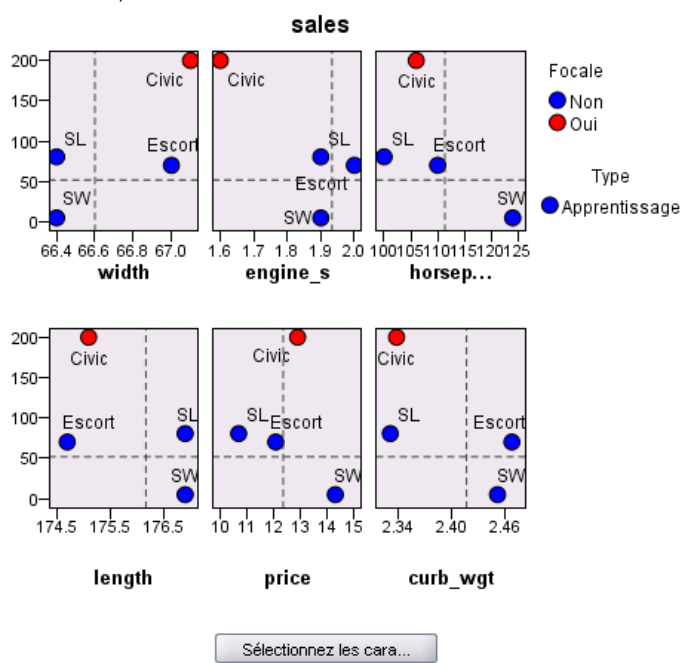
Le diagramme des paires est lié à l'espace du prédicteur de deux façons.

- Les observations sélectionnées (focales) dans l'espace du prédicteur sont affichées dans le diagramme des paires, ainsi que leurs k voisins les plus proches.
- La valeur de k sélectionnée dans l'espace du prédicteur est utilisée dans le diagramme des paires.

Sélectionner les valeurs prédites. Vous permet de sélectionner les variables prédites à afficher dans le diagramme des paires.

Carte des quadrants

Figure 16-14
Carte des quadrants



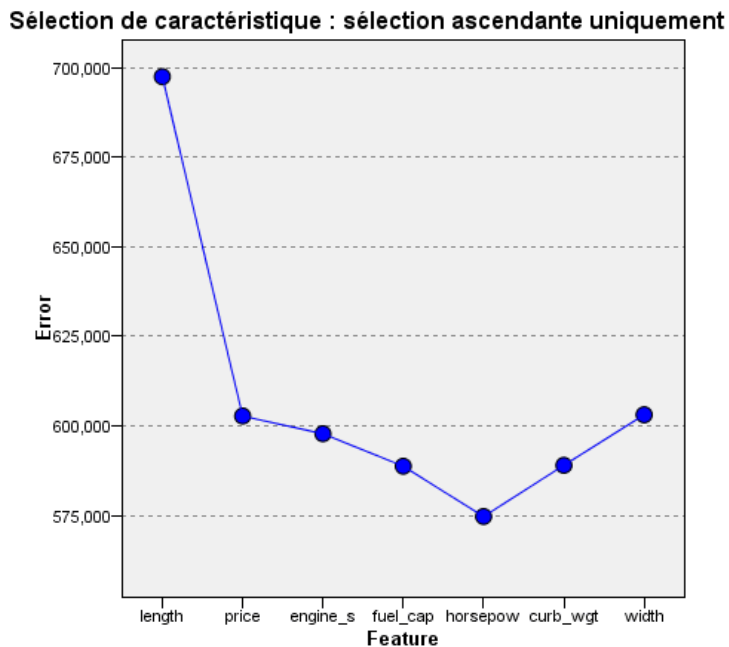
Ce diagramme affiche les observations focales et leur k voisins les plus proches sur un diagramme de dispersion (ou nuage de points, selon le niveau de mesure de la cible) avec la cible sur l'axe y et une valeur prédite d'échelle sur l'axe x, affichés sous forme de panel par variable prédite. Il est disponible si une cible existe et si une observation focale est sélectionnée dans l'espace du prédicteur.

- Les lignes de référence sont tracées pour des variables continues, aux moyennes variables dans la partition de formation.

Sélectionner les valeurs prédites. Vous permet de sélectionner les variables prédites à afficher dans la carte de cadran.

Journal d'erreur de sélection des valeurs prédites

Figure 16-15
Sélection de valeurs prédites



Les points du diagramme affichent l'erreur (le ratio du taux d'erreur ou l'erreur de la somme des carrés, selon le niveau de mesure de la cible) sur l'axe y pour le modèle avec la variable prédite sur l'axe x (plus toutes les caractéristiques à gauche sur l'axe x). Ce diagramme est disponible si une cible existe et si la sélection des caractéristiques est activée.

Le tableau de classification

Figure 16-18
Tableau de classement

Partition		Prévisions		
		0	1	Pourcentage correct
Apprentissage	0	111	1	99.11%
	1	7	33	82.50%
	Pourcentage global	77.64%	22.37%	94.74%

Ce tableau affiche par partition la classification croisée des valeurs prévues de la cible observées contre celles prévues. Il est disponible s'il y a une cible et si elle est catégorielle (booléen, nominal ou ordinal).

- La ligne (Manquante) dans la partition traitée contient des observations traitées contenant des valeurs manquantes sur la cible. Ces observations contribuent à l'échantillon traité : Les valeurs de pourcentage global mais pas les valeurs de pourcentage correct.

Récapitulatif d'erreur

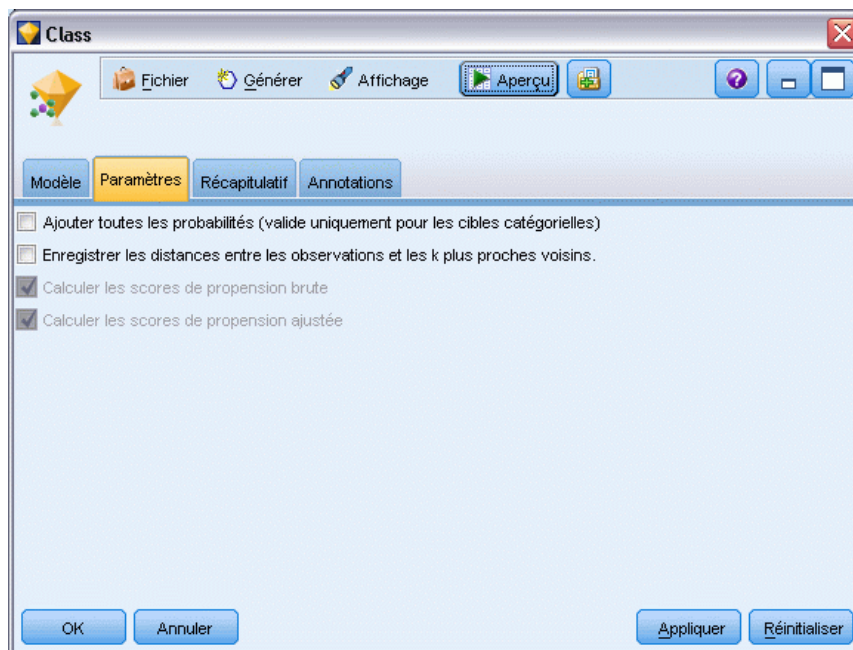
Figure 16-19
Récapitulatif des erreurs

Partition	Sum-of-Squares Error
Apprentissage	622043

Ce tableau est disponible si une variable cible existe. Il affiche l'erreur associée au modèle, la somme des carrés pour une cible continue et le taux d'erreur (100% – pourcentage général correct) pour une cible qualitative.

Paramètres du modèle KNN

Figure 16-20
Paramètres des nuggets de modèles KNN



L'onglet Paramètres vous permet de spécifier des champs supplémentaires à afficher lors de la consultation des résultats (par exemple en exécutant un noeud Table relié au nugget). Vous pouvez voir l'effet de chacune de ces options en les sélectionnant et en cliquant sur le bouton Aperçu : faites défiler vers la droite de l'aperçu pour voir les champs supplémentaires.

Ajouter toutes les probabilités (valide uniquement pour les cibles catégorielles). Si cette option est cochée, les probabilités pour chaque valeur possible d'un champ cible Nominal ou Booléen sont affichées pour chaque enregistrement traité par le noeud. Sinon, seule la valeur prédite et la probabilité correspondante sont affichées pour les champs cible de type nominal ou booléen.

Le paramètre par défaut de cette case à cocher est déterminé par la case correspondante sur le noeud de modélisation.

Calculer les scores de propension brute. Pour les modèles disposant d'un booléen cible (retournant une prédiction oui ou non), vous pouvez demander des scores de propension qui indiquent la probabilité du résultat vrai spécifié pour le champ cible. Ces scores viennent s'ajouter aux autres valeurs de prédiction et de confiance qui peuvent être générées pendant le scoring.

Calculer les scores de propension ajustée. Les scores de propension brute sont uniquement basés sur les données d'apprentissage et peuvent être exagérément optimiste en raison de la tendance de nombreux modèles à sur-ajuster ces données. Les propensités ajustées tentent de compenser cet écart en évaluant les performances par rapport à un test ou à une partition de validation. Cette option nécessite la définition d'un champ de partition dans le flux et les scores de propension ajustée doivent être activés dans le noeud de modélisation avant la génération du modèle.

Afficher le plus proche. Si vous configurez cette valeur sur n , où n représente un entier positif différent de zéro, les n voisins les plus proches de l'enregistrement central sont inclus dans le modèle, ainsi que leur distance relative à l'enregistrement central.

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.



- ajouter des règles de modèle, 253
- ajustement de Bonferroni
 - nœud CHAID, 176
- algorithmes, 49, 129
- amélioration des performances, 301, 387, 392, 419
- analyse de la variance
 - dans les modèles mixtes linéaires généralisés, 342
- analyse des composantes principales. *Consultez les modèles ACP*, 310, 314
- Analyse du voisin le plus proche
 - vue du modèle, 519
- analyse en arbre
 - emplois généraux, 130
- analyse log linéaire
 - dans les modèles mixtes linéaires généralisés, 342
- analyse probit
 - modèles mixtes linéaires généralisés, 342
- ANOVA
 - dans des modèles linéaires, 283
- antécédent
 - règles dépourvues, 426
- aperçu
 - contenu de modèle, 56
- apprentissage non supervisé, 382–383
- arbres de régression, 158–160
- arbres de segmentation, 158–160, 180
- arbres interactifs, 128, 131, 133, 136, 138
 - bénéfices, 143
 - divisions personnalisées, 134
 - exportation des résultats, 153
 - gains, 140–141, 144, 147
 - génération de diagrammes, 195
 - génération de modèles, 149–150
 - Retour sur investissement, 143
 - valeurs de substitution, 137
- autorégression
 - modèles ARIMA, 475

- bagging, 164
 - dans des modèles linéaires, 270
 - dans les réseaux de neurones, 215
- bénéfices
 - gains d'arbre décision, 143
- boosting, 164, 181, 193
 - dans des modèles linéaires, 270
 - dans les réseaux de neurones, 215

- carte d'arbre
 - génération de diagrammes, 195
 - modèles d'arbre décision, 190
- carte des quadrants
 - dans l'analyse du voisin le plus proche, 525
- carte territoriale
 - Noeud discriminant, 321
- cartes auto-organisatrices, 383

- catégorie de base
 - Noeud Logistique, 294
- Champ ID
 - noeud CARMA, 422
 - Noeud Séquence, 444
- Champ Temps
 - noeud CARMA, 422
 - Noeud Séquence, 444
- champs d'analyse
 - noeud CARMA, 422
 - Noeud Séquence, 444
- champs de pondération, 42
- champs d'effectif, 42
- champs d'entrée
 - filtrage, 80
 - sélection pour analyse, 80
- champs disponibles, 251
- changer la valeur cible, 257
- chargement
 - nuggets de modèle, 52
- chi-deux normalisé
 - mesure d'évaluation a priori, 421
- classement des variables indépendantes, 78, 81–83, 85
- Classification, 382–383, 390, 393–395, 398–399
 - affichage de classes, 400
 - affichage général, 400
 - détection des anomalies, 89
 - Nombre de classes, 396
- coefficient de variance
 - filtrage des champs, 80
- conditions de règle
 - modèles Liste de décision, 229
- confiance
 - modèles d'arbre décision, 189
 - noeud Apriori, 418
 - noeud CARMA, 425
 - Noeud Séquence, 446
 - pour les séquences, 451
 - règles d'association, 429, 431, 453
- confiances
 - ensembles de règles, 191
 - modèles d'arbre décision, 191
 - modèles de régression logistique, 306
- conséquence
 - conséquences multiples, 426
- copier des liens de modèle, 50
- corrélations asymptotiques
 - modèles de régression logistique, 299, 308
- correspondances
 - gains d'arbre décision, 140
- coûts
 - arbres décision, 169, 171
- coûts de mauvaise réaffectation
 - arbres de décision, 106, 170, 172
 - nœud C5.0, 181

- covariance asymptotique
 - modèles de régression logistique, 299
- Critère d'Akaike
 - dans des modèles linéaires, 273
- critère de prévention de surajustement
 - dans des modèles linéaires, 273
- critères
 - dans des modèles linéaires, 273
- cycles non saisonniers, 460

- décalage
 - ACF et PACF, 463
- démarrage, 240
- détection de fraudes
 - détection des anomalies, 86
- détection de séquences, 415, 443
- diagramme de l'espace du prédicteur
 - dans l'analyse du voisin le plus proche, 521
- différence de confiance
 - mesure d'évaluation a priori, 420
- différence de confiance absolue a priori
 - mesure d'évaluation a priori, 420
- différence d'informations
 - mesure d'évaluation a priori, 421
- différence du quotient de confiance par rapport à 1
 - mesure d'évaluation a priori, 420
- directives
 - arbres décision, 153
- directives d'arbre, 164
 - arbres décision, 153
 - Noeud Arbre C&RT, 151
 - nœud CHAID, 151, 153
 - nœud QUEST, 151
- distances du voisin le plus proche
 - dans l'analyse du voisin le plus proche, 523
- divisions personnalisées
 - arbres décision, 134, 136–137
- documentation, 4
- données de table de valeurs vraies, 416, 439–440
- données du panier, 416, 439–440
- données manquantes
 - série de valeurs prédites, 465
- données tabulaires, 416, 439
 - noeud Apriori, 40
 - noeud CARMA, 422
 - Noeud Séquence, 444
 - Transposition, 440
- données till-roll (déroulantes), 416, 439–440
- données transactionnelles, 416, 439–440
 - noeud Apriori, 40
 - noeud CARMA, 422
 - Noeud Règles d'association MS, 40
 - Noeud Séquence, 444
- DTD, 73

- éditer
 - paramètres avancés, 250

- effets principaux
 - modèles de régression logistique, 295
- effets saisonniers, 460
 - identification, 459
- élagage d'arbres décision, 158, 166
- enregistrements centraux, 511
- ensemble de règles, 154, 191, 198, 200, 434, 437–438
 - génération à partir d'arbres décision, 154
- ensemble de règles de séquence généré, 438
- ensemble de règles de vote, 198
- ensemble de règles premier résultat, 198
- ensembles
 - dans des modèles linéaires, 275
 - dans les réseaux de neurones, 219
- Epsilon pour la convergence
 - nœud CHAID, 177
- estimation des risques
 - gains d'arbre décision, 149
- Estimations des paramètres
 - modèles de régression logistique, 308
 - modèles linéaires généralisés, 337
- Étiquettes
 - value, 73
 - variable, 73
- évaluation dans Excel, 260
- évaluation des données, 70
- évaluer un modèle, 258
- événements
 - identification, 461
- exécuter une tâche d'exploration, 247
- exemples
 - Aperçu, 6
- Exemples
 - Guide des applications, 4
- exemples d'application, 4
- exhaustive CHAID, 129, 131, 165
- exportation
 - nuggets de modèle, 52
 - PMML, 73, 75
 - SQL, 56

- factorisation, 130
 - modèles ACP/Analyse factorielle, 310
- fichiers d'adresses
 - modèles Liste de décision, 229
- filtrage des champs d'entrée, 80
- filtrage des règles, 430, 453
 - règles d'association, 431
- filtrage des variables indépendantes, 78, 83, 85
- fonction d'autocorrélation
 - série, 463
- fonction d'autocorrélation partielle
 - série, 463
- fonction de lien
 - modèles mixtes linéaires généralisés, 345
- fonction générale estimée
 - modèles linéaires généralisés, 337

- fonction radiale de base (RBF)
 - dans les réseaux de neurones, 217
- fonctions de transfert, 476
 - délai, 476
 - ordres de dénominateur, 476
 - ordres de différence, 476
 - ordres de numérateur, 476
 - ordres saisonniers, 476
- fonctions du noyau
 - modèles support vector machine, 498
- format de configuration de l'intégration à MS Excel, 261
- fréquences
 - modèles d'arbre décision, 189
- fusion de catégories, 130

- gains
 - arbres décision, 140–141, 144
 - diagramme, 264
 - exportation, 153
- gains de classification
 - arbres décision, 141, 144
- gains de régression
 - arbres décision, 144, 147
- générateur d'arbres, 131, 133, 138
 - bénéfices, 143
 - divisions personnalisées, 134
 - exportation des résultats, 153
 - gains, 140–141, 144, 147
 - génération de diagrammes, 195
 - génération de modèles, 149–150
 - Retour sur investissement, 143
 - valeurs de substitution, 137
 - variables indépendantes, 136
- génération de diagrammes
 - règles d'association, 432
- génération de règles de segment, 247
- générer un nouveau modèle, 258
- questionnaires
 - Onglet Modèles, 52
- graphiques de réponses
 - gains d'arbre décision, 140, 146
- graphiques d'évaluation
 - à partir de modèles de classificateur automatique, 126
 - à partir de modèles de numérisation automatique, 126
- graphiques Evaluation
 - à partir de modèles de classificateur automatique, 125
 - à partir de modèles de classification automatique, 125
 - à partir de modèles de numérisation automatique, 125
- graphiques Lift
 - gains d'arbre décision, 146
- groupes de paires
 - détection des anomalies, 89

- historique
 - modèles d'arbre décision, 189
- Historique des itérations
 - modèles de régression logistique, 299

- modèles linéaires généralisés, 337

- IBM InfoSphere Warehouse (ISW)
 - exportation en PMML, 75
- IBM SPSS Modeler, 1
 - documentation, 4
- ID de règle, 430
- identification d'interaction, 130
- importance
 - classement des variables indépendantes, 78, 81–83, 85
 - filtrage de champs, 59
 - variables indépendantes dans les modèles, 44, 57, 59
- importance des champs
 - classement de champs, 78, 81–83, 85
 - filtrage de champs, 59
 - Résultats de modèles, 44, 57, 59
- importance des valeurs prédites
 - dans l'analyse du voisin le plus proche, 523
- importance des variables
 - modèles de réponse en auto-apprentissage., 493
- importance des variables indépendantes
 - modèles d'arbre décision, 186
 - modèles discriminants, 324
 - modèles linéaires généralisés, 339
- importance des variables prédites
 - filtrage de champs, 59
 - modèles de régression logistique, 303
 - modèles linéaires, 279
 - réseaux de neurones, 223
 - Résultats de modèles, 44, 57, 59
- importation
 - PMML, 52, 73, 75
- impulsions
 - séries, 461
- index
 - gains d'arbre décision, 140
 - induction de règle, 128, 158–160, 180, 417
 - informations sur le modèle
 - modèles linéaires généralisés, 337
 - instances, 429, 452
 - modèles d'arbre décision, 189
 - instantané
 - Création, 245
 - intégration
 - modèles ARIMA, 475
 - interactions
 - modèles de régression logistique, 295
 - Intervalle de confiance
 - modèles de régression logistique, 299
 - interventions
 - identification, 461
 - interventions d'étape
 - identification, 461
 - interventions ponctuelles
 - identification, 461

- Khi-deux
 - nœud CHAID, 177
 - sélection des caractéristiques, 82
- Khi-deux de Pearson
 - nœud CHAID, 177
 - sélection des caractéristiques, 82
- Khi-deux du rapport de vraisemblance
 - nœud CHAID, 177
 - sélection des caractéristiques, 82
- KNN. *Voir* modèles d'agrégation suivant le saut minimum, 507

- Lambda
 - sélection des caractéristiques, 82
- liens
 - Modèle, 49
- liens de modèle, 49
 - Copie et collage, 50
 - définition et suppression, 49
 - et super noeuds, 51
- lift, 429
 - gains d'arbre décision, 140
 - règles d'association, 431
- lissage exponentiel, 465
 - critères dans le modélisateur de séries temporelles, 473
- log-odds
 - modèles de régression logistique, 303

- marques commerciales, 530
- Matrice de corrélation
 - modèles linéaires généralisés, 337
- Matrice de covariance
 - modèles linéaires généralisés, 337
- matrice des coefficients de contraste
 - modèles linéaires généralisés, 337
- Matrice L
 - modèles linéaires généralisés, 337
- meilleurs sous-ensembles
 - dans des modèles linéaires, 273
- mentions légales, 529
- mesure de la capacité de déploiement, 430
- mesure d'impureté Associer par paire, 173
- mesure d'impureté Associer par paire ordonnée, 173
- mesure d'impureté Gini, 173
- mesures de modèle
 - Définition, 259
 - rafraîchir, 260
- mesures d'évaluation
 - noeud Apriori, 419
- mesures d'impureté
 - arbres décision, 172
 - Noeud Arbre C&RT, 173
- MLP (perceptron multicouche)
 - dans les réseaux de neurones, 217
- Modalité de référence
 - Noeud Logistique, 294
- modèle linéaire général
 - modèles mixtes linéaires généralisés, 342
- modèle linéaire généralisé
 - dans les modèles mixtes linéaires généralisés, 342
- Modèles
 - ARIMA, 475
 - découpage, 33, 36–38
 - importation, 52
 - Onglet Récapitulatif, 56
 - remplacement, 51
- modèles ACP
 - équations, 315
 - Facteurs, 312
 - gestion des valeurs manquantes, 312
 - itérations, 312
 - noeud de modélisation, 310
 - nombre de facteurs, 312
 - nugget de modèle, 314–315, 317
 - options de modèle, 311
 - options expert, 312
 - rotation, 313
 - sorties avancées, 317
 - Valeurs propres, 312
- modèles alternatifs, 256
- modèles apriori
 - données tabulaires et données transactionnelles, 40
 - mesures d'évaluation, 419
 - noeud de modélisation, 417
 - options du noeud de modélisation, 418
 - options expert, 419
- modèles ARIMA, 466
 - constante, 475
 - critères dans le modélisateur de séries temporelles, 474
 - fonctions de transfert, 476
 - ordres autorégressifs, 475
 - ordres de différenciation, 475
 - ordres de moyenne mobile, 475
 - ordres saisonniers, 475
 - valeurs éloignées, 478
- modèles bruts, 77, 83, 85, 415
- Modèles C5.0, 129
 - boosting, 181, 193
 - coûts de mauvaise réaffectation, 181
 - élagage, 181
 - génération de diagrammes à partir d'un nugget de modèle, 195
 - noeud de modélisation, 180–181, 190–191, 193
 - nugget de modèle, 184, 198, 200
 - options, 181
 - performance, 181, 184
 - traitement parallèle, 181, 184
- Modèles CARMA
 - Champ ID, 422
 - Champ Temps, 422
 - champs d'analyse, 422
 - conséquences multiples, 440
 - données tabulaires et données transactionnelles, 426

- formats de données, 422
- noeud de modélisation, 421
- options de champs, 422
- options du noeud de modélisation, 425
- options expert, 426
- Modèles CHAID, 129
 - coûts de mauvaise réaffectation, 171
 - ensembles, 168
 - exhaustive CHAID, 165
 - génération de diagrammes à partir d'un nugget de modèle, 195
 - noeud de modélisation, 131, 157, 159, 190–191
 - nugget de modèle, 184
 - objectifs, 162
 - options d'arrêt, 167
 - options de champs, 161
 - Profondeur d'arborescence, 165
- modèles d'agrégation suivant le saut minimum
 - à propos de, 507
 - noeud de modélisation, 507
 - options d'analyse, 517
 - options de modèle, 510
 - options de paramètres, 509
 - options de sélection des caractéristiques, 514
 - options de validation croisée, 515
 - options des objectifs, 508
 - options des voisins, 512
- Modèles d'arbre C&RT
 - coûts de mauvaise réaffectation, 169
 - élagage, 166
 - ensembles, 168
 - génération de diagrammes à partir d'un nugget de modèle, 195
 - mesures d'impureté, 172
 - noeud de modélisation, 131, 157–158, 190–191
 - nugget de modèle, 184
 - objectifs, 162
 - options d'arrêt, 167
 - options de champs, 161
 - pondérations d'effectifs, 42
 - pondérations d'observation, 42
 - Probabilités a priori, 170
 - Profondeur d'arborescence, 165
 - valeurs de substitution, 166
- modèles d'arbre décision, 128, 131, 133, 138, 157–161, 180, 184, 190, 195
 - afficheur, 190
 - bénéfices, 143
 - coûts de mauvaise réaffectation, 169, 171
 - divisions personnalisées, 134
 - effectifs de la règle, 189
 - exportation des résultats, 153
 - gains, 140–141, 144, 147
 - génération, 149–150
 - génération de diagrammes, 195
 - importance des variables indépendantes, 186
 - noeud de modélisation, 154
 - nugget de modèle, 186
 - options d'arrêt, 167
 - panneau d'informations supplémentaire, 189
 - règles d'arbre, 186
 - Retour sur investissement, 143
 - valeurs de substitution, 137, 189
 - variables indépendantes, 136
- modèles de classificateur automatique, 96
 - fenêtre du navigateur des résultats, 122
 - génération de noeuds de modélisation et de nuggets, 125
 - graphiques d'évaluation, 126
 - graphiques Evaluation, 125
 - Introduction, 99
 - modèles de classement, 101
 - noeud de modélisation, 99, 101
 - nugget de modèle, 122
 - Paramètres, 108
 - paramètres d'algorithme, 97
 - partitions, 103
 - règles d'arrêt, 98
 - suppression de modèles, 107
 - types de modèle, 103
- modèles de classification automatique, 96
 - fenêtre du navigateur des résultats, 122
 - génération de noeuds de modélisation et de nuggets, 125
 - graphiques Evaluation, 125
 - modèles de classement, 118
 - noeud de modélisation, 118
 - nugget de modèle, 122
 - paramètres d'algorithme, 97
 - partitions, 120
 - règles d'arrêt, 98
 - suppression de modèles, 121
 - types de modèle, 120
- Modèles de classification automatique
 - noeud de modélisation, 117
- Modèles de classification TwoStep, 382, 396, 398
 - Classification, 398
 - génération de diagrammes à partir d'un nugget de modèle, 412
 - noeud de modélisation, 395
 - Nombre de classes, 396
 - nugget de modèle, 398
 - options, 396
 - standardisation des champs, 396
 - traitement des valeurs éloignées, 396
- modèles de détection des anomalies, 92
 - champs d'anomalie, 88, 94
 - coefficient de correction, 89
 - groupes de pairs, 89, 93
 - index d'anomalie, 88
 - niveau de bruit, 89
 - noeud de modélisation, 86
 - scoring, 91, 94
 - valeur de césure, 88, 93
 - Valeurs manquantes, 89

- Modèles de nuées dynamiques
 - génération de diagrammes à partir d'un nugget de modèle, 412
- modèles de numérisation automatique, 96
 - fenêtre du navigateur des résultats, 122
 - génération de noeuds de modélisation et de nuggets, 125
 - graphiques d'évaluation, 126
 - graphiques Evaluation, 125
 - noeud de modélisation, 110–111
 - nugget de modèle, 122
 - options de modélisation, 111
 - Paramètres, 116
 - paramètres d'algorithme, 97
 - règles d'arrêt, 98, 114
 - types de modèle, 114
- modèles de rafraîchissement
 - modèles de réponse en auto-apprentissage., 490
- modèles de règle non affinée, 427, 436–437
- modèles de règles d'association, 191, 198, 200, 415, 449, 451, 453–454
 - a priori, 417
 - CARMA, 421
 - définition des filtres, 431
 - déploiement, 440
 - détails des nuggets de modèles, 427
 - génération de diagrammes, 432
 - génération d'un ensemble de règles, 437
 - génération d'un modèle filtré, 438
 - IBM InfoSphere Warehouse, 40
 - nugget de modèle, 427
 - Paramètres, 434
 - pour les séquences, 443
 - récapitulatif des nuggets de modèles, 436
 - règles de scoring, 439
 - transposition des scores, 440
- modèles de régression
 - noeud de modélisation, 268
- Modèles de régression de Cox, 381
 - Critères de convergence, 377
 - critères de l'analyse pas à pas, 378
 - noeud de modélisation, 371
 - nugget de modèle, 380
 - options de champs, 372
 - options de modèle, 373
 - options de paramètres, 379
 - options expert, 376
 - sorties avancées, 377, 381
- modèles de régression linéaire, 267
 - moindres carrés pondérés, 42
 - noeud de modélisation, 268
- modèles de régression logistique, 267
 - ajouter des termes, 295
 - effets principaux, 295
 - équations de modèle, 303
 - importance des variables prédites, 303
 - interactions, 295
 - noeud de modélisation, 288
 - nugget de modèle, 302–303, 305–306
 - options binomiales, 290
 - options de convergence, 298
 - options de l'analyse pas à pas, 301
 - options expert, 297
 - options multinomiales, 290
 - sorties avancées, 299, 308
- modèles de régression logistique binomiale, 288, 290
- modèles de régression logistique multinomiale, 288, 290
- modèles de réponse en auto-apprentissage.
 - importance des variables, 493
 - noeud de modélisation, 488
 - nugget de modèle, 493
 - options de champs, 489
 - Paramètres, 492, 495–496
 - préférences des champs cible, 493, 497
 - rafraîchissement de modèle, 490
 - randomisation des résultats, 492, 496
- modèles de réseau Bayésien
 - noeud de modélisation, 203
 - nugget de modèle, 209
 - options de modèle, 205
 - options expert, 207
 - paramètres des nuggets de modèles, 210
 - récapitulatif des nuggets de modèles, 212
- modèles de réseau de neurones
 - options de champs, 39
- modèles de séquences
 - Champ ID, 444
 - Champ Temps, 444
 - champs d'analyse, 444
 - détails des nuggets de modèles, 451
 - données tabulaires et données transactionnelles, 447
 - formats de données, 444
 - génération d'un super noeud Règle, 455
 - navigateur de séquence, 454
 - noeud de modélisation, 443
 - nugget de modèle, 449, 451, 453–454
 - options, 446
 - options de champs, 444
 - options expert, 447
 - paramètres des nuggets de modèles, 453
 - prévisions, 449
 - récapitulatif des nuggets de modèles, 454
 - Tri, 454
- modèles de séries temporelles
 - conditions requises, 466
 - Critères ARIMA, 474
 - critères du lissage exponentiel, 473
 - critères du modélisateur expert, 471
 - fonctions de transfert, 476
 - lissage exponentiel, 465
 - modèles ARIMA, 466
 - noeud de modélisation, 465
 - nugget de modèle, 480
 - Paramètres de modèle, 484
 - périodicité, 476

- Résidus, 485
- transformation de séries, 476
- valeurs éloignées, 472, 478
- modèles découpés, 511
 - création, 33
 - et partitionnement, 36
 - fonctions concernées par, 38
 - noeuds de modélisation, 37
- modèles discriminants
 - critère de l'analyse pas à pas (sélection des champs), 323
 - Critères de convergence, 320
 - forme du modèle, 319
 - noeud de modélisation, 318
 - nugget de modèle, 324–326
 - options expert, 320
 - scores de propension, 325
 - scoring, 324
 - sorties avancées, 321, 325
- modèles factoriels
 - équations, 315
 - Facteurs, 312
 - gestion des valeurs manquantes, 312
 - iterations, 312
 - noeud de modélisation, 310
 - nombre de facteurs, 312
 - nugget de modèle, 314–315, 317
 - options de modèle, 311
 - options expert, 312
 - rotation, 313
 - sorties avancées, 317
 - Valeurs propres, 312
- modèles hiérarchiques
 - modèles mixtes linéaires généralisés, 342
- modèles IBM SPSS Statistics, 33
- modèles K-Means, 382, 390–392
 - champ de distance, 391
 - Classification, 390, 394
 - codage de valeurs pour les ensembles, 392
 - critère d'arrêt, 392
 - nugget de modèle, 393–394
 - options expert, 392
- Modèles Kohonen, 382–383, 385, 387
 - critère d'arrêt, 385
 - génération de diagrammes à partir d'un nugget de modèle, 412
 - noeud de modélisation, 383
 - nugget de modèle, 389
 - option des codages des ensembles binaires (supprimée), 385
 - options expert, 387
 - représentation graphique, 385
 - réseaux de neurones, 383, 389
 - taux d'apprentissage, 387
 - voisinage, 383, 387
- modèles linéaires, 269
 - choix du modèle, 273
 - coefficients, 284
 - critère d'information, 277
 - duplication des résultats, 276
 - ensembles, 275
 - importance des variables prédites, 279
 - moyennes estimées, 286
 - niveau de confiance, 272
 - objectifs, 270
 - options de modèle, 276
 - paramètres des nuggets, 288
 - préparation automatique des données, 272, 278
 - récapitulatif de création de modèle, 287
 - récapitulatif du modèle, 277
 - règles de combinaison, 275
 - Résidus, 281
 - Statistique R-deux, 277
 - Tableau ANOVA, 283
 - valeurs éloignées, 282
 - valeurs prédites en fonction des valeurs observées, 280
- modèles linéaires généralisés
 - champs, 329
 - forme du modèle, 330
 - noeud de modélisation, 327
 - nugget de modèle, 339, 341
 - options de convergence, 335
 - options expert, 332
 - scores de propension, 340
 - sorties avancées, 337, 340
- modèles Liste de décision
 - configuration requise, 234
 - espace de travail de l'afficheur, 240
 - exclusion de segments, 230
 - fichiers d'adresses, 229
 - génération SQL, 240
 - largeur de recherche, 237
 - méthode de création d'intervalles, 237
 - noeud de modélisation, 229
 - onglet alternatives, 243
 - onglet instantanés, 245
 - options de modèle, 235
 - options expert, 237
 - panneau du modèle de travail, 241
 - paramètres, 240
 - PMML, 238
 - scoring, 230, 238
 - segments, 238
 - sens de la recherche, 235
 - utilisation de l'afficheur, 247
 - valeur cible, 235
- modèles longitudinaux
 - modèles mixtes linéaires généralisés, 342
- modèles mixtes
 - modèles mixtes linéaires généralisés, 342
- modèles mixtes linéaires généralisés, 342
 - bloc d'effets aléatoires, 352
 - coefficients fixes, 365
 - covariances à effet aléatoire, 367
 - décalage, 354

- distribution de la cible, 345
- Effets aléatoires, 351
- effets fixes, 348, 363
- fonction de lien, 345
- moyennes estimées, 369
- moyennes marginales estimées, 357
- options de scoring, 356
- paramètres, 370
- paramètres de covariance, 368
- pondération d'analyse, 354
- récapitulatif du modèle, 359
- structure des données, 360
- tableau de classification, 362
- termes personnalisés, 350
- valeurs prédites en fonction des valeurs observées, 361
- vue du modèle, 358
- modèles multi-niveaux
 - modèles mixtes linéaires généralisés, 342
- Modèles QUEST, 129
 - coûts de mauvaise réaffectation, 169
 - élagage, 166
 - ensembles, 168
 - génération de diagrammes à partir d'un nugget de modèle, 195
 - noeud de modélisation, 131, 157, 160, 190–191
 - nugget de modèle, 184
 - objectifs, 162
 - options d'arrêt, 167
 - options de champs, 161
 - Probabilités a priori, 170
 - Profondeur d'arborescence, 165
 - valeurs de substitution, 166
- modèles sélection de fonction, 83, 85
 - classement des variables indépendantes, 78–79, 83
 - filtrage des variables indépendantes, 78–79, 83
 - génération de noeuds Filtrer, 85
 - importance, 78–79, 83
- modèles Statistics, 33
- modèles statistiques, 267
- modèles support vector machine
 - à propos de, 498
 - affinement, 500
 - fonctions du noyau, 498
 - noeud de modélisation, 501
 - nugget de modèle, 504, 519
 - options de modèle, 502
 - options expert, 502
 - paramètres, 505
 - surajustement, 500
- modélisateur expert
 - critères dans le modélisateur de séries temporelles, 471
 - valeurs éloignées, 472
- moindres carrés pondérés, 42
- moyenne mobile
 - modèles ARIMA, 475
- navigateur de séquence, 454
- niveaux de signification
 - de la division, 175–176
 - pour la fusion, 176
- niveaux, validation croisée, 515
- noeud Bâtir règle, 184
- Noeud Filtrer
 - génération à partir d'arbres décision, 154
- noeud linéaire, 269
- noeud neuralnetwork, 213
- noeud nodeName, 342
- noeud Sélectionner
 - génération à partir d'arbres décision, 154
- noeuds de modélisation, 26, 86, 180, 203, 383, 390, 395, 417, 443, 488
- Noeuds de modélisation automatisés
 - modèles de classificateur automatique, 96
 - modèles de classification automatique, 96
 - modèles de numérisation automatique, 96
- noyau linéaire
 - modèles support vector machine, 498
- nuggets de modèle, 48, 77, 184, 191, 193, 198, 200, 341
 - enregistrement, 55
 - enregistrement et chargement, 52
 - évaluation des données avec, 70
 - exportation, 52, 55
 - génération de noeuds de traitement, 70
 - Impression, 55
 - menus, 55
 - modèles découpés, 68–69
 - modèles d'ensemble, 60
 - Onglet Récapitulatif, 56
 - utilisation dans les flux, 70
- nuggets de modèle découpé, 68
 - afficheur, 69
 - Onglet Récapitulatif, 56
- onglet Afficheur
 - génération de diagrammes, 195
 - modèles d'arbre décision, 190
- onglet Alternatives, 243
- Onglet Instantanés, 245
- optimisation des performances, 387, 392, 419
- options d'arrêt
 - arbres de décision, 167
- options de champs
 - Noeud de Cox, 372
 - noeud MRAA, 489
 - noeuds de modélisation, 39
- options de convergence
 - Modèles de régression de Cox, 377
 - modèles de régression logistique, 298
 - modèles linéaires généralisés, 335
 - noeud CHAID, 177
- Options de diagramme, 265
- options de l'analyse pas à pas
 - Modèles de régression de Cox, 378
 - modèles de régression logistique, 301

- options de modèle
 - Modèles de régression de Cox, 373
 - noeud MRAA, 490
 - Noeud Réseau Bayésien, 205
- options de paramètres
 - Modèles de régression de Cox, 379
 - noeud MRAA, 492
- options expert
 - Modèles de régression de Cox, 376
 - modèles *K*-Means, 392
 - Modèles Kohonen, 387
 - noeud Apriori, 419
 - noeud CARMA, 426
 - Noeud Réseau Bayésien, 207
 - Noeud Séquence, 447
- ordres saisonniers
 - modèles ARIMA, 475
- organiser les sélections de données, 252

- pairs
 - dans l'analyse du voisin le plus proche, 524
- palette de modèles, 48, 52
- Panneau des règles alternatives, 253
- panneau d'informations supplémentaire
 - modèles d'arbre décision, 189
- panneau du modèle de travail, 241
- paramètres
 - dans les modèles de séries temporelles, 484
- paramètres avancés, 250
- partitions, 41, 424, 445, 511
 - création de modèles, 102, 112, 118, 182, 205, 235, 290, 311, 319, 330, 373, 385, 391, 397, 446, 490, 502, 510
 - Sélection, 41, 424, 445, 511
- pas à pas ascendant
 - dans des modèles linéaires, 273
- perceptron multicouche (MLP)
 - dans les réseaux de neurones, 217
- performance
 - Modèles C5.0, 181, 184
- périodicité
 - Time Series Modeler, 476
- personnaliser un modèle, 256
- PMML
 - exportation de modèles, 52, 73, 75
 - importation de modèles, 52, 73, 75
- préparation automatique des données
 - dans des modèles linéaires, 278
- prévention de surajustement
 - dans les réseaux de neurones, 220
- prévision
 - Aperçu, 458
 - série de valeurs prédites, 465
- prise en charge
 - noeud Apriori, 418
 - noeud CARMA, 425–426
 - Noeud Séquence, 446
 - pour les séquences, 451
 - prise en charge de la règle, 429, 453
 - prise en charge des antécédents, 429, 453
 - règles d'association, 431
- probabilités
 - modèles de régression logistique, 303
- probabilités a priori, 170
- Probabilités a priori
 - arbres décision, 170
- Profondeur d'arborescence, 165
- pseudo *R*-deux
 - modèles de régression logistique, 308

- Qualité d'ajustement Hosmer-Lemeshow
 - modèles de régression logistique, 308
- qualité de l'ajustement
 - modèles de régression logistique, 308

- R-deux
 - dans des modèles linéaires, 277
- R-deux ajusté
 - dans des modèles linéaires, 273
- rafraîchir les mesures, 260
- rafraîchissement de modèle
 - modèles de réponse en auto-apprentissage., 490
- ratio de confiance
 - mesure d'évaluation a priori, 420
- récapitulatif d'erreur
 - dans l'analyse du voisin le plus proche, 527
- réduction dimensionnelle, 383
- règles
 - prise en charge de la règle, 429, 453
 - règles d'association, 417, 421
- règles de combinaison
 - dans des modèles linéaires, 275
 - dans les réseaux de neurones, 219
- règles doubles, 426
- Régression de Poisson
 - modèles mixtes linéaires généralisés, 342
- Régression logistique
 - modèles mixtes linéaires généralisés, 342
- régression logistique multinomiale
 - modèles mixtes linéaires généralisés, 342
- régression nominale, 288
- remplacement de modèles, 51
- réseaux de neurones, 213
 - classification, 224
 - couches masquées, 217
 - duplication des résultats, 220
 - ensembles, 219
 - fonction radiale de base (RBF), 217
 - importance des variables prédites, 223
 - objectifs, 215
 - options de modèle, 221
 - paramètres des nuggets, 227
 - perceptron multicouche (MLP), 217
 - prévention de surajustement, 220

- récapitulatif du modèle, 222
- règles d'arrêt, 218
- règles de combinaison, 219
- réseau, 226
- Valeurs manquantes, 220
- valeurs prédites en fonction des valeurs observées, 224
- Résidus
 - dans les modèles de séries temporelles, 485
- résultat expert.
 - Modèles de régression de Cox, 377
- Retour sur investissement
 - gains d'arbre décision, 143
- risques
 - exportation, 153
- rotation
 - modèles ACP/Analyse factorielle, 313
- Rotation equamax
 - modèles ACP/Analyse factorielle, 313
- Rotation oblimin directe
 - modèles ACP/Analyse factorielle, 313
- Rotation promax
 - modèles ACP/Analyse factorielle, 313
- Rotation quartimax
 - modèles ACP/Analyse factorielle, 313
- Rotation Varimax
 - modèles ACP/Analyse factorielle, 313

- scores de confiance, 47
- scores de propension
 - équilibrage des données, 47
 - modèles discriminants, 325
 - modèles linéaires généralisés, 340
 - modèles Liste de décision, 240
- scores de propension brute, 47
- scores de propensions ajustés
 - équilibrage des données, 47
 - modèles discriminants, 325
 - modèles linéaires généralisés, 340
 - modèles Liste de décision, 240
- segmentation, 130
- segments
 - classement par ordre de priorité, 256
 - copier, 255
 - exclusion, 257
 - Insertion, 253
 - modèles Liste de décision, 229
 - modification, 254
 - Suppression, 257
 - suppression des conditions de règle, 255
- sélection basée sur les gains, 147
- sélection de champs pas à pas
 - Noeud discriminant, 323
- sélection de valeurs prédites
 - dans l'analyse du voisin le plus proche, 526
- sélections de création
 - Définition, 248
- série
 - transformation, 464
- série de valeurs prédites, 465
 - données manquantes, 465
- SLRM *Consultez* les modèles de réponse en auto-apprentissage., 488
- sorties avancées
 - Modèles de régression de Cox, 377
 - noeud Analyse factorielle/ACP, 314
- splits
 - arbres décision, 134, 136–137
- SPSS Modeler Server, 2
- SQL
 - ensembles de règles, 192
 - exportation, 56
 - modèles de régression logistique, 307
- Statistique de significativité, 299, 301
- Statistique de Wald, 299, 301
- statistique F
 - dans des modèles linéaires, 273
 - sélection des caractéristiques, 82
- statistiques de qualité de l'ajustement
 - modèles de régression logistique, 308
 - modèles linéaires généralisés, 337
- Statistiques descriptives
 - modèles linéaires généralisés, 337
- statistiques *T*
 - sélection des caractéristiques, 82
- stratification, 130
- Super noeud Règle
 - génération à partir d'un noeud Règles de séquence, 455
- Super noeuds
 - et liens de modèle, 51
- Suppression
 - liens de modèle, 49
- suppression de liens de modèle, 49
- surajustement d'un modèle SVM, 500
- SVM. *Voir* les modèles support vector machine, 498

- tableau de classement
 - dans l'analyse du voisin le plus proche, 526
- tableau de classification
 - modèles de régression logistique, 299
- tâche d'exploration
 - démarrage, 248
- tâches d'exploration, 247
 - Création, 248
 - modèles Liste de décision, 229
 - modification, 249
- tendances
 - identification, 459
- tendances linéaires
 - identification, 459
- tendances non linéaires
 - identification, 459
- test de rapport de vraisemblance
 - modèles de régression logistique, 299, 308

- Test M de Box
 - Noeud discriminant, 321
- Test multiplicateur Lagrange
 - modèles linéaires généralisés, 338
- traitement parallèle
 - Modèles C5.0, 181, 184
- transformation de série, 464
- transformation de type Différenciation, 464
 - modèles ARIMA, 475
- transformation de type Différenciation saisonnière, 464
 - modèles ARIMA, 475
- transformation de type Stabilisation de niveau, 464
- transformation de type Stabilisation de variance, 464
- transformation fonctionnelle, 464
- Transformation log :, 464
 - Time Series Modeler, 476
- transformation par log népérien, 464
 - Time Series Modeler, 476
- transformation racine carrée, 464
 - Time Series Modeler, 476
- transposition de sortie tabulaire, 440

- V de Crameré
 - sélection des caractéristiques, 82
- valeur p , 82
- valeurs de substitution
 - arbres décision, 137, 166
 - modèles d'arbre décision, 189
- valeurs éloignées, 462
 - additives saisonnières, 463
 - correctifs additifs, 462
 - dans les modèles de séries temporelles, 478
 - décalage de niveau, 463
 - déterministes, 462
 - identification, 86
 - innovatrices, 463
 - modèles ARIMA, 478
 - modélisateur expert, 472
 - modification passagère, 463
 - séries, 461
 - tendance locale, 463
- valeurs éloignées additives, 462
 - correctifs, 462
 - Time Series Modeler, 478
- valeurs éloignées additives saisonnières, 463
 - Time Series Modeler, 478
- valeurs éloignées avec décalage de niveau, 463
 - Time Series Modeler, 478
- valeurs éloignées avec modification passagère, 463
 - Time Series Modeler, 478
- valeurs éloignées de tendance locale, 463
 - Time Series Modeler, 478
- valeurs éloignées innovatrices, 463
 - Time Series Modeler, 478
- Valeurs manquantes
 - arbres CHAID, 136
 - exclusion de SQL, 192
 - filtrage des champs, 80
- Valeurs propres
 - modèles ACP/Analyse factorielle, 312
- Variables
 - filtrage, 130
- variables continues
 - segmentation, 130
- variables continues en bandes, 130
- variables indépendantes
 - arbres décision, 136
 - classement par importance, 78, 81–83, 85
 - filtrage, 78, 83, 85
 - sélection pour analyse, 78, 81, 83, 85
 - valeurs de substitution, 137
- viewer d'ensemble, 60
 - détails d'un modèle de composant, 67
 - fréquence des valeurs prédites, 64
 - importance des valeurs prédites, 63
 - précision d'un modèle de composant, 65
 - préparation automatique des données, 68
 - récapitulatif du modèle, 62
- visualisation
 - arbres décision, 190
 - génération de diagrammes, 195, 412, 432
 - modèles de classification, 400
- visualiser un modèle, 264
- visualiseur de classes
 - à propos des modèles de classe, 399
 - affichage du contenu des cellules, 405
 - Aperçu, 400
 - comparaison des classes, 409
 - distribution des cellules, 408
 - faire basculer les classes et les caractéristiques, 404
 - génération de diagrammes, 412
 - importance des variables indépendantes, 406
 - récapitulatif du modèle, 402
 - taille des classes, 407
 - transposer les classes et les caractéristiques, 404
 - trier l'affichage des caractéristiques, 405
 - trier l'affichage des classes, 405
 - trier le contenu des cellules, 405
 - trier les caractéristiques, 405
 - trier les classes, 405
 - Utilisation, 410
 - vue de base, 406
 - vue de comparaison des classes, 409
 - vue de la distribution des cellules, 408
 - vue de la taille des classes, 407
 - vue de l'importance des valeurs prédites de classe, 406
 - vue des centres de classes, 403
 - vue des classes, 403
 - vue récapitulative, 402
- vue du modèle
 - dans l'analyse du voisin le plus proche, 519
 - dans les modèles mixtes linéaires généralisés, 358