IBM SPSS Decision Trees 26

Important
Avant d'utiliser le présent document et le produit associé, prenez connaissance des informations générales figurant à la section «Remarques», à la page 19.

La présente édition s'applique à la version 26.0.0 d'IBM SPSS Statistics et à toutes les éditions et modifications ultérieures sauf mention contraire dans les nouvelles éditions.

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFACON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE.

Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. Les informations qui y sont fournies sont susceptibles d'être modifiées avant que les produits décrits ne deviennent eux-mêmes disponibles. En outre, il peut contenir des informations ou des références concernant certains produits, logiciels ou services non annoncés dans ce pays. Cela ne signifie cependant pas qu'ils y seront annoncés.

Pour plus de détails, pour toute demande d'ordre technique, ou pour obtenir des exemplaires de documents IBM, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial.

Vous pouvez également consulter les serveurs Internet suivants :

- http://www.fr.ibm.com (serveur IBM en France)
- http://www.ibm.com/ca/fr (serveur IBM au Canada)
- http://www.ibm.com (serveur IBM aux Etats-Unis)

Compagnie IBM France
Direction Qualité
17, avenue de l'Europe
92275 Bois-Colombes Cedex

Table des matières

Avis aux lecteurs canadiens

Le présent document a été traduit en France. Voici les principales différences et particularités dont vous devez tenir compte.

Illustrations

Les illustrations sont fournies à titre d'exemple. Certaines peuvent contenir des données propres à la France.

Terminologie

La terminologie des titres IBM peut différer d'un pays à l'autre. Reportez-vous au tableau ci-dessous, au besoin.

IBM France	**IBM Canada**
ingénieur commercial	représentant
agence commerciale	succursale
ingénieur technico-commercial	informaticien
inspecteur	technicien du matériel

Claviers

Les lettres sont disposées différemment : le clavier français est de type AZERTY, et le clavier français-canadien de type QWERTY.

OS/2 et Windows - Paramètres canadiens

Au Canada, on utilise :

- les pages de codes 850 (multilingue) et 863 (français-canadien),
- le code pays 002,
- le code clavier CF.

Nomenclature

Les touches présentées dans le tableau d'équivalence suivant sont libellées différemment selon qu'il s'agit du clavier de la France, du clavier du Canada ou du clavier des États-Unis. Reportez-vous à ce tableau pour faire correspondre les touches françaises figurant dans le présent document aux touches de votre clavier.

France	Canada	Etats-Unis
(Pos1)		Home
Fin	Fin	End
(PgAr)		PgUp
(PgAv)		PgDn
Inser	Inser	Ins
Suppr	Suppr	Del
Echap	Echap	Esc
Attn	Intrp	Break
Impr écran	ImpEc	PrtSc
Verr num	Num	Num Lock
Arrêt défil	Défil	Scroll Lock
(Verr maj)	FixMaj	Caps Lock
AltGr	AltCar	Alt (à droite)

Brevets

Il est possible qu'IBM détienne des brevets ou qu'elle ait déposé des demandes de brevets portant sur certains sujets abordés dans ce document. Le fait qu'IBM vous fournisse le présent document ne signifie pas qu'elle vous accorde un permis d'utilisation de ces brevets. Vous pouvez envoyer, par écrit, vos demandes de renseignements relatives aux permis d'utilisation au directeur général des relations commerciales d'IBM, 3600 Steeles Avenue East, Markham, Ontario, L3R 9Z7.

Assistance téléphonique

Si vous avez besoin d'assistance ou si vous voulez commander du matériel, des logiciels et des publications IBM, contactez IBM direct au 1 800 465-1234.

Arbres de décision

Les fonctions suivantes d'arbre de décision sont incluses dans SPSS Statistics Professional Edition ou l'option Arbres de décision.

Création d'arbres de décisions

La procédure Arbre de décisions crée un modèle de segmentation basée sur un arbre. Elle classe les observations en groupes ou estime les valeurs d'une variable (cible) dépendante à partir des valeurs de variables (prédicteur) indépendantes. Cette procédure fournit des outils de validation pour les analyses de classification d'exploration et de confirmation.

Vous pouvez utiliser cette procédure pour les opérations suivantes :

Segmentation : Identifie les personnes susceptibles d'appartenir à un groupe particulier.

Stratification : Attribue des observations à l'intérieur d'une des catégories telles que les groupes à risques élevé, moyen ou faible.

Prévision : Elabore des règles et les utilise pour prédire des événements futurs, tels que la probabilité qu'une personne manque à ses engagements à l'occasion d'un prêt ou la valeur de revente possible d'un véhicule ou d'une maison.

Réduction de données et balayage des variables : Sélectionne à partir d'un ensemble étendu de variables un sous-ensemble exploitable de prédicteurs utilisé pour construire un modèle paramétrique formel.

Identification des interactions : Identifie les relations relatives uniquement à certains sous-groupes particuliers et spécifie ces relations dans un modèle paramétrique formel.

Fusion des catégories et discrétisation des variables continues : Etablit un nouveau code de groupe des catégories de prédicteur et des variables continues avec une perte d'informations minimum.

Exemple : Les banques cherchent à classer les demandeurs de crédit selon le risque de crédit, raisonnable ou pas, qu'ils représentent. A partir de plusieurs facteurs, dont le classement de solvabilité connue des anciens clients, vous pouvez construire un modèle estimant les futurs clients susceptibles de manquer à leurs engagements de remboursement de leur prêt.

Une analyse sous forme d'arbre présente des fonctions intéressantes :

- Elle vous permet d'identifier des groupes homogènes présentant un risque élevé ou faible.
- Cela facilite l'élaboration de règles de prédiction pour chaque observation.

Remarques sur les données

Données : Les variables dépendantes et indépendantes peuvent être les suivantes :

- *Nominal*. Une variable peut être traitée comme étant nominale si ses valeurs représentent des catégories sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le code postal ou l'appartenance religieuse sont des exemples de variables nominales.
- *Ordinal*. Une variable peut être traitée comme étant ordinale si ses valeurs représentent des catégories associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.

- *Echelle*. Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des catégories ordonnées avec une mesure significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

Pondération de fréquence Si le calcul des pondérations est activé, les pondérations fractionnelles sont arrondies à l'entier le plus proche ; ainsi, les observations ayant une valeur de pondération inférieure à 0,5 ont un pondération de 0 et sont donc exclues de l'analyse.

Hypothèses : Cette procédure considère qu'un niveau de mesure adéquat a été attribué à toutes les variables d'analyse, et certaines fonctions considèrent que toutes les valeurs de la variable dépendante incluses dans l'analyse ont des libellés de valeur définis.

- **Niveau de mesure :** Le niveau de mesure a une influence sur les trois calculs ; le bon niveau de mesure doit donc être attribué à chaque variable. Par défaut, on considère que les variables numériques sont des variables d'échelle et que les variables de chaîne sont nominales, ce qui risque de ne pas refléter correctement les niveaux de mesure. Dans la liste des variables, une icône indique le type de chaque variable.

Pour modifier de manière temporaire le niveau de mesure d'une variable, cliquez sur la variable dans la liste des variables source avec le bouton droit de la souris et sélectionnez un niveau de mesure dans le menu contextuel.

- **Libellés de valeurs :** L'interface de la boîte de dialogue associée à cette procédure suppose que, soit toutes les valeurs non manquantes d'une variable catégorielle (nominale, ordinale) disposent de libellés de valeurs définis, soit qu'aucune n'en dispose. Certaines fonctions ne sont disponibles que si au moins deux valeurs non manquantes de la variable dépendante catégorielle disposent de libellés de valeur. Si au moins deux valeurs non manquantes disposent de libellés de valeur définies, toutes les observations contenant d'autres valeurs ne disposant pas de libellés de valeur seront exclues de l'analyse.

Obtention d'arbres de décisions

1. A partir des menus, sélectionnez :

   Analyse > **Classifier** > **Arborescence**.
2. Sélectionnez une variable dépendante.
3. Sélectionnez une ou plusieurs variables indépendantes.
4. Sélectionnez une méthode de croissance.

Sinon, vous pouvez :

- Modifiez le niveau de mesure de toutes les variables de la liste source.
- Introduisez de force la première variable de la liste des variables indépendantes dans le modèle en tant que première variable de scission.
- Sélectionnez une variable d'influence définissant le degré d'influence d'une observation sur le processus de croissance de l'arbre. Les observations ayant des valeurs d'influence faibles ont le moins d'influence ; les observations ayant des valeurs élevées en ont le plus. Les valeurs de variables d'influence doivent être positives.
- Validez l'arbre.
- Personnalisez les critères de croissance de l'arbre.
- Enregistrez les numéros des noeuds terminaux, les prévisions et les probabilités prévues en tant que variables.
- Enregistrez le modèle au format XML (PMML).

Champs avec un niveau de mesure inconnu

L'alerte du niveau de mesure s'affiche lorsque le niveau de mesure d'une ou plusieurs variables (champs) du jeu de données est inconnu. Le niveau de mesure ayant une incidence sur le calcul des résultats de cette procédure, toutes les variables doivent avoir un niveau de mesure défini.

Analyser les données
Lit les données dans le jeu de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si le jeu de données est important, cette action peut prendre un certain temps.

Affecter manuellement
Répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez affecter un niveau de mesure à ces champs. Vous pouvez également affecter un niveau de mesure dans le panneau Liste de variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas exécuter celle-ci avant que tous les champs n'aient des niveaux de mesure définis.

Modification du niveau de mesure

1. Cliquez avec le bouton droit sur la variable dans la liste source.
2. Sélectionnez un niveau de mesure dans le menu contextuel.

Le niveau de mesure est alors modifié de manière temporaire pour être utilisé dans la procédure Arbre de décisions.

Méthodes de croissance

Les méthodes de croissance disponibles sont :

CHAID
Chi-squared Automatic Interaction Detection. A chaque étape, CHAID choisit la variable indépendante (prédicteur) dont l'interaction avec la variable dépendante est la plus forte. Les catégories de chaque prédicteur sont fusionnées si elles ne présentent pas de différences significatives avec la variable dépendante.

CHAID exhaustif
Version modifiée de CHAID qui examine toutes les scissions possibles pour chaque prédicteur.

CRT Classification and Regression Trees (arbres de segmentation et de régression). CRT divise les données en segments aussi homogènes que possible par rapport à la variable dépendante. Un noeud terminal dans lequel toutes les observations ont la même valeur de variable dépendante est un noeud homogène et "pur".

QUEST
Quick, Unbiased, Efficient Statistical Tree (arbre statistique rapide, impartial et efficace). Méthode rapide qui favorise les prédicteurs avec de nombreuses catégories par rapport au biais des autres méthodes. La méthode QUEST ne peut être spécifiée que si la variable dépendante est nominale.

Chaque méthode présente des avantages et des limites, qui sont les suivantes :

Tableau 1. Fonctions de la méthode de croissance.

Fonction	CHAID*	CRT	QUEST
Calculé à partir du khi-carré**	X		
Variables (prédicteur) indépendantes de substitution		X	X
Elagage des arbres		X	X
Scission de noeud multiple	X		
Scission de noeud binaire		X	X

Tableau 1. Fonctions de la méthode de croissance (suite).

Fonction	CHAID*	CRT	QUEST
Variables d'influence	X	X	
Probabilités a priori		X	X
Coûts de classification erronée	X	X	X
Calcul rapide	X		X

*Inclut CHAID exhaustif.

**QUEST utilise également une mesure du khi-carré pour les variables indépendantes nominales.

Sélection de catégories

Pour les variables dépendantes catégorielles (nominales, ordinales), vous pouvez effectuer les opérations suivantes :

- Contrôler les catégories à inclure dans l'analyse.
- Identifier les catégories cible qui vous intéressent.

Inclusion/exclusion de catégories

Vous pouvez limiter l'analyse à certaines catégories de la variable dépendante.

- Les observations dont les valeurs de la variable dépendante figurent dans la liste Exclure ne sont pas incluses dans l'analyse.
- Pour les variables dépendantes nominales, vous pouvez également inclure des catégories manquantes spécifiées par l'utilisateur dans l'analyse. (Par défaut, les catégories manquantes spécifiées par l'utilisateur s'affichent dans la liste Exclure.)

Catégories cible

Les catégories sélectionnées (qui sont cochées) sont traitées comme les catégories ayant le plus grand intérêt dans l'analyse. Par exemple, si l'identification des personnes les plus susceptibles de manquer à leurs engagements envers un prêt est la catégorie qui vous intéresse le plus, sélectionnez la catégorie « mauvais » classement de solvabilité en tant que catégorie cible.

- Aucune catégorie cible n'a été définie. Si aucune catégorie n'est sélectionnée, certaines options de règle de classification et certaines sorties liées aux gains ne sont pas disponibles.
- Si plusieurs catégories sont sélectionnées, vous obtenez des tableaux et des graphiques de gains séparés pour chaque catégorie cible.
- La désignation de plusieurs catégories en tant que catégories cible n'a aucun effet sur le modèle de l'arbre, sur l'estimation des risques ou sur les résultats de classification erronée.

Catégories

Cette boîte de dialogue requiert que soient définis des libellés de valeurs pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante catégorielle disposent de libellés de valeur définis.

Inclusion/exclusion de catégories et sélection de catégories cible

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante catégorielle (nominale, ordinale) avec deux libellés de valeur définis, ou plus.
2. Cliquez sur **Catégories**.

Validation

La validation vous permet d'évaluer si votre arbre est généralisable à une plus grande population. Deux méthodes de validation sont disponibles : la validation croisée et la validation par scission d'échantillon.

Validation croisée

La validation croisée consiste à fractionner l'échantillon en plusieurs sous-échantillons ou **niveaux**. Les arbres sont générés en excluant à tour de rôle les données de chaque sous-échantillon. Le premier arbre est basé sur toutes les observations excepté celles du premier sous-échantillon, le deuxième arbre est basé sur toutes les observations excepté celles du deuxième sous-échantillon, etc. Le risque de mauvaise réaffectation est estimé pour chaque arbre en appliquant l'arbre au sous-échantillon exclu lors de la génération de l'arbre.

Important : La validation croisée n'est pas disponible pour les méthodes CRT et Quest si l'élagage est sélectionné.

- Vous pouvez indiquer un maximum de 25 niveaux d'échantillon. Plus la valeur est élevée, moins les observations exclues de chaque modèle d'arbre sont nombreuses.
- La validation croisée obtient un modèle d'arbre final unique. L'estimateur de risque en validation croisée pour l'ensemble de l'arbre est calculé en faisant la moyenne des risques de tous les arbres.

Validation par partition d'échantillon

Pour la validation par partition d'échantillon, le modèle est créé à partir d'un échantillon d'apprentissage et est testé sur un échantillon traité.

- Vous pouvez indiquer une taille d'échantillon d'apprentissage, exprimée sous forme de pourcentage de la taille d'échantillon totale, ou une variable de scission de l'échantillon en échantillons d'apprentissage et de test.
- Si vous utilisez une variable pour définir les échantillons d'apprentissage et de test, les observations ayant la valeur 1 pour la variable sont attribuées à l'échantillon d'apprentissage et toutes les autres observations sont attribuées à l'échantillon de test. Il ne peut pas s'agir d'une variable dépendante, de pondération, d'influence ou d'une variable indépendante forcée.
- Vous pouvez afficher les résultats pour l'échantillon d'apprentissage et pour l'échantillon de test, ou uniquement pour l'échantillon de test.
- La validation par scission d'échantillon doit être utilisée avec précaution sur les petits fichiers de données (les fichiers de données comportant un petit nombre d'observations). Des échantillons d'apprentissage de petite taille risquent de former des modèles erronés, puisque certaines catégories peuvent ne pas comporter suffisamment d'observations pour construire correctement l'arbre.

Validation d'un arbre de décisions

1. Dans la boîte de dialogue Arbres de décisions principale, cliquez sur **Validation**.
2. Sélectionnez **Validation croisée** ou **Validation par scission d'échantillon**.

Remarque : Ces deux méthodes de validation attribuent les observations aux groupes d'échantillons de manière aléatoire. Si vous voulez reproduire les mêmes résultats dans une autre analyse, vous devez définir une valeur de départ de nombre aléatoire (menu Transformer, Générateurs de nombres aléatoires) avant de lancer l'analyse pour la première fois, puis rétablir ce nombre aléatoire pour l'autre analyse en question.

Critères de croissance de l'arbre

Les critères de croissance disponibles peuvent dépendre de la méthode de croissance, du niveau de mesure de la variable dépendante ou de la combinaison des deux.

Limites de croissance

La boîte de dialogue Limites de croissance vous permet de limiter le nombre de niveaux de l'arbre et de contrôler le nombre minimal d'observations des noeuds parent et enfant.

Profondeur maximale de l'arbre
Contrôle le nombre maximal de niveaux de croissance en dessous du noeud racine. Le paramètre **Automatique** limite l'arbre à trois niveaux en dessous du noeud racine pour les méthodes CHAID et CHAID exhaustif, et à cinq niveaux pour les méthodes CRT et QUEST.

Nombre minimal d'observations
Contrôle le nombre minimum d'observations des noeuds. Les noeuds ne respectant pas ces critères ne sont pas scindés.

- Si vous augmentez les valeurs minimum, les arbres construits ont tendance à comporter moins de noeuds.
- Si vous diminuez les valeurs minimum, les arbres construits ont plus de noeuds.

Pour les fichiers de données comportant un petit nombre d'observations, les valeurs par défaut définissant 100 observations pour les noeuds parent et 50 pour les noeuds enfant peuvent créer des arbres sans noeud en dessous du noeud racine ; dans ce cas, vous obtiendrez des résultats plus utiles en abaissant les valeurs minimales.

Spécification de limites de croissance

1. Dans la boîte de dialogue Arbre de décisions principale, cliquez sur **Limites de croissance**.

Critères CHAID

Pour les méthodes CHAID et CHAID exhaustif, vous pouvez contrôler les éléments suivants :

Niveau de signification pour
Vous pouvez contrôler la valeur de signification pour scinder des noeuds et fusionner des catégories. Pour ces deux critères, le niveau de signification par défaut est 0,05.

Scission de noeuds
Cette valeur doit être supérieure à 0 et inférieure à 1. Les valeurs faibles produisent des arbres avec moins de noeuds.

Fusion des catégories
Cette valeur doit être supérieure à 0 et inférieure ou égale à 1. Pour empêcher la fusion de catégories, spécifiez la valeur 1. Pour une variable d'échelle indépendante, ceci signifie que le nombre de catégories pour la variable dans l'arbre final correspond au nombre d'intervalles spécifié (10 par défaut). Pour plus d'informations, voir «Intervalles d'échelle pour l'analyse CHAID», à la page 7.

Statistiques du Khi-carré
Pour les variables dépendantes ordinales, le khi-carré déterminant la scission des noeuds et la fusion des catégories est calculé via la méthode du rapport de vraisemblance. Pour les variables dépendantes nominales, vous avez le choix entre plusieurs méthodes :

Pearson
Cette méthode fournit des calculs plus rapides mais doit être utilisée avec précaution sur les petits échantillons. Il s'agit de la méthode par défaut.

Rapport de vraisemblance
Cette méthode est plus fiable que Pearson mais son temps de calcul est plus long. C'est la méthode la plus adaptée aux petits échantillons.

Estimation du modèle

Pour les variables dépendantes nominales ou ordinales, vous pouvez indiquer :

Nombre maximal d'itérations

La valeur par défaut est 100. Si l'arbre cesse de croître parce que le nombre maximum d'itérations a été atteint, vous pouvez augmenter ce maximum ou modifier d'autres critères contrôlant la croissance de l'arbre.

Changement minimum dans les fréquences théoriques de cellule

Cette valeur doit être supérieure à 0 et inférieure à 1. La valeur par défaut est 0,05. Les valeurs faibles génèrent des arbres comportant moins de noeuds.

Ajustement des valeurs de signification à l'aide de la méthode Bonferroni

Pour les comparaisons multiples, les valeurs de signification des critères de fusion et de scission sont ajustées à l'aide de la méthode Bonferroni. Il s'agit de la valeur par défaut.

Autoriser la scission des catégories fusionnées à l'intérieur d'un noeud

A moins que vous n'empêchiez explicitement la fusion des catégories, la procédure tente de fusionner les catégories des variables indépendantes (prédicteur) pour produire l'arbre décrivant le modèle le plus simple. Cette option autorise la procédure à scinder des catégories fusionnées pour améliorer la solution obtenue.

Spécification de critères CHAID

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez la méthode de croissance **CHAID** ou **CHAID exhaustif**.
2. Cliquez sur **CHAID**.

Intervalles d'échelle pour l'analyse CHAID : Dans l'analyse CHAID, les variables indépendantes (prédicteur) d'échelle sont toujours regroupées en catégories indépendantes (par exemple, de 0 à 10, de 11 à 20, de 21 à 30, etc.) avant d'être analysées. Vous pouvez contrôler le nombre initial/maximum de groupes (même si la procédure peut fusionner des groupes contigus après la scission initiale) :

Nombre fixe

Toutes les variables d'échelle indépendantes sont groupées à l'origine dans le même nombre de groupes. La valeur par défaut est 10.

Personnalisé

Chaque variable d'échelle indépendante est répartie à l'origine dans le nombre de groupes déterminé pour cette variable.

Spécification d'intervalles pour les variables d'échelle indépendantes

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une ou plusieurs variables d'échelle dépendantes.
2. Pour **Méthode de croissance**, sélectionnez **CHAID** ou **Exhaustive CHAID**.
3. Cliquez sur **Intervalles**.

Dans les analyses CRT et QUEST, toutes les scissions sont binaires et les variables d'échelle indépendantes ou ordinales sont traitées de la même manière ; par conséquent, vous ne pouvez pas indiquer un nombre d'intervalles pour les variables d'échelle indépendantes.

Critères CRT

La méthode de croissance CRT tente d'optimiser l'homogénéité des noeuds. La limite à laquelle un noeud ne représente pas un sous-ensemble homogène d'observations est un indicateur d'**impureté**. Par exemple, un noeud terminal dans lequel toutes les observations ont la même valeur pour la variable dépendante est un noeud homogène qui n'a pas besoin d'être scindé davantage car il est « pur ».

Vous pouvez sélectionner la méthode utilisée pour mesurer l'impureté et la diminution minimum de l'impureté pour scinder les noeuds.

Mesure d'impureté

Pour les variables d'échelle dépendantes, c'est la mesure d'impureté des moindres carrés des écarts (LSD) qui est utilisée. Elle est calculée en tant que variance intra-noeud, ajustée selon les pondérations de fréquence ou les valeurs d'influence. Pour les variables dépendantes (nominales, ordinales) catégorielles, vous pouvez sélectionner la mesure d'impureté parmi les suivantes :

Gini Des scissions sont effectuées pour optimiser l'homogénéité des noeuds enfant par rapport à la valeur de la variable dépendante. La méthode Gini est basée sur les carrés des probabilités d'appartenance à chaque catégorie de la variable dépendante. Elle atteint son minimum (zéro) lorsque toutes les observations du noeud entrent dans une seule catégorie. Il s'agit de la mesure par défaut.

Twoing

Les catégories de la variable dépendante sont regroupées en deux sous-classes. Des scissions améliorant la séparation des deux groupes sont réalisées.

Twoing ordonné

Identique au twoing, avec la contrainte supplémentaire que seules les catégories adjacentes peuvent être regroupées. Cette mesure est uniquement disponible pour les variables dépendantes ordinales.

Changement minimum de l'amélioration

Il s'agit de la diminution minimum de l'impureté requise pour scinder un noeud. La valeur par défaut est 0.0001. Les valeurs élevées génèrent des arbres comportant moins de noeuds.

Spécification de critères CRT

1. Pour **Méthode de croissance**, sélectionnez **CRT**.
2. Cliquez sur **CRT**.

Critères QUEST

Pour la méthode QUEST, vous pouvez déterminer le niveau de signification pour scinder les noeuds. Une variable indépendante ne peut pas être utilisée pour scinder des noeuds à moins que le niveau de signification ne soit inférieur ou égal à la valeur indiquée. Cette valeur doit être supérieure à 0 et inférieure à 1. La valeur par défaut est 0,05. Les valeurs faibles auront tendance à exclure plus de variables indépendantes du modèle final.

Spécification de critères QUEST

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante nominale.
2. Pour **Méthode de croissance**, sélectionnez **QUEST**.
3. Cliquez sur **QUEST**.

Elagage des arbres

Avec les méthodes CRT et QUEST, vous pouvez faire en sorte que le modèle ne soit pas trop rempli en **élaguant** l'arbre : l'arbre croît jusqu'à atteindre les critères d'arrêt ; il est ensuite automatiquement tronqué jusqu'au sous-arbre le plus petit, selon la différence maximum de risque indiquée. La valeur de risque est exprimée en erreurs standard. La valeur par défaut est 1. Elle ne doit pas être négative. Pour obtenir un sous-arbre qui possède le risque minimum, indiquez 0.

Important : La validation croisée n'est pas disponible pour les méthodes CRT et Quest si l'élagage est sélectionné.

Elagage d'un arbre

1. Dans la boîte de dialogue Arbre de décisions principale, pour **Méthode de croissance**, sélectionnez **CRT** ou **QUEST**.
2. Cliquez sur **Elagage**.

Elagage et masquage des noeuds

Lorsque vous créez un arbre élagué, tous les noeuds ayant été élagués de l'arbre ne sont pas disponibles dans l'arbre final. Vous pouvez masquer et afficher de manière interactive les noeuds enfant sélectionnés dans l'arbre final, mais vous ne pouvez pas afficher les noeuds élagués lors du processus de création de l'arbre.

Valeurs de substitution

Les méthodes CRT et QUEST peuvent utiliser des **valeurs de substitution** pour les variables indépendantes (prédicteur). Pour les observations dans lesquelles la valeur de cette variable est manquante, d'autres variables indépendantes ayant un fort degré d'association avec la variable d'origine sont utilisées pour la classification. Ces prédicteurs de rechange sont appelés valeurs de substitution. Vous pouvez déterminer le nombre maximum de valeurs de substitution pouvant être utilisé dans le modèle.

- Par défaut, le nombre maximum de valeurs de substitution correspond à une unité de moins que le nombre de variables prédites. Autrement dit, pour chaque variable indépendante, toutes les autres variables indépendantes peuvent être utilisées comme valeurs de substitution.
- Si vous ne souhaitez pas que le modèle utilise des valeurs de substitution, indiquez 0 comme nombre de valeurs de substitution.

Pour spécifier des valeurs de substitution

1. Dans la boîte de dialogue Arbre de décisions principale, pour **Méthode de croissance**, sélectionnez **CRT** ou **QUEST**.
2. Cliquez sur **Valeurs de substitution**.

Options

Les options disponibles dépendent de la méthode de croissance, du niveau de mesure de la variable dépendante et/ou de l'existence de libellés de valeur définis pour les valeurs de la variable dépendante.

Coûts de classification erronée

Pour les variables dépendantes catégorielles (nominales, ordinales), les coûts de classification erronée permettent d'inclure des informations sur les pénalités relatives associées aux classements incorrects de l'arbre. Par exemple :

- Le coût engendré par le refus d'un crédit à un client solvable sera vraisemblablement différent du coût engendré par la prolongation du crédit d'un client déjà en défaut de paiement.
- Le coût occasionné pas le classement incorrect d'une personne présentant un risque élevé de cardiopathie dans la catégorie de risque faible sera probablement beaucoup plus élevé que le coût occasionné par le classement erroné d'une personne à risque faible dans la catégorie de risque élevé.
- Le coût du publipostage d'une personne qui ne répondra sûrement pas est relativement faible, alors que le coût engendré par le non-publipostage d'une personne susceptible de répondre est plus élevé (en recettes perdues).

Remarque : La boîte de dialogue Coûts de classification erronée n'est disponible que si au moins deux valeurs de la variable dépendante catégorielle disposent de libellés de valeur définis.

Spécification de coûts de classification erronée

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante catégorielle (nominale, ordinale) avec deux libellés de valeur définis, ou plus.
2. Cliquez sur **Coûts de classification erronée**.
3. Cliquez sur **Personnalisé**.
4. Entrez un ou plusieurs coûts de classification erronée dans la grille Catégorie estimée. Les valeurs ne doivent pas être négatives. (les affectations correctes, représentées sur la diagonale, ont toujours la valeur 0.)

Rendre la matrice symétrique

La plupart du temps, vous voudrez que les coûts soient symétriques ; en d'autres termes, que le coût occasionné par la mauvaise réaffectation de A comme B soit identique au coût occasionné par la mauvaise réaffectation de B comme A. Les contrôles suivants vous aident à spécifier une matrice de coûts symétrique :

Dupliquer le triangle inférieur

Permet de copier les valeurs comprises dans le triangle inférieur de la matrice (situé en dessous de la diagonale) dans les cellules correspondantes du triangle supérieur.

Dupliquer le triangle supérieur

Permet de copier les valeurs comprises dans le triangle supérieur de la matrice (situé au-dessus de la diagonale) dans les cellules correspondantes du triangle inférieur.

Utiliser les moyennes de cellules

Cette option calcule la moyenne des deux valeurs de cellule situées chacune dans une moitié différente (l'une dans le triangle inférieur et l'autre dans le triangle supérieur) et remplace ces deux valeurs par la moyenne ainsi obtenue. Par exemple, si le coût occasionné par la mauvaise réaffectation de A comme B est 1, et le coût occasionné par la mauvaise réaffectation de B comme A est 3, ces deux valeurs sont alors remplacées par leur moyenne : (1+3)/2 = 2.

Bénéfices

Pour les variables dépendantes catégorielles, vous pouvez attribuer des valeurs de recette et de dépense aux niveaux de la variable dépendante.

- Les bénéfices sont obtenus avec le calcul suivant : recettes moins dépenses.
- Les valeurs de bénéfice ont un effet sur les valeurs de la moyenne des bénéfices et du ROI (retour sur investissement) dans les tableaux de gains. Elles n'ont pas d'effet sur la structure de base du modèle d'arbre.
- Les valeurs des recettes et des dépenses doivent être numériques et propres à toutes les catégories de la variable dépendante affichée dans la grille.

Remarque : Cette boîte de dialogue requiert que soient définis des libellés de valeurs pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante catégorielle disposent de libellés de valeur définis.

Spécification de bénéfices

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante catégorielle (nominale, ordinale) avec deux libellés de valeur définis, ou plus.
2. Cliquez sur **Bénéfices**.
3. Cliquez sur **Personnalisé**.
4. Saisissez les valeurs de recette et de dépense de toutes les catégories de variable dépendante répertoriées dans la grille.

Probabilités a priori

Pour les arbres CRT et QUEST comportant des variables dépendantes catégorielles, vous pouvez déterminer des probabilités a priori pour les groupes d'affectation. Les **probabilités a priori** sont des estimations de la fréquence relative globale de chaque catégorie de la variable dépendante, effectuées avant la prise de connaissance des valeurs des variables indépendantes (prédicteur). Les probabilités a priori aident à corriger les croissances d'arbre générées par les données de l'échantillon non représentatif de l'intégralité de la population.

Obtenue à partir d'échantillons d'apprentissage (probabilités a priori empiriques)

Utilisez ce paramètre si l'affectation des valeurs de la variable dépendante dans le fichier de

données est représentative de la distribution de la population. Si vous utilisez la validation par partition d'échantillon, c'est la distribution des observations dans l'échantillon d'apprentissage qui est utilisée.

Remarque : Etant donné que les observations sont affectées de manière aléatoire à l'échantillon d'apprentissage dans la validation par partition d'échantillon, la distribution réelle des observations dans l'échantillon d'apprentissage ne peut être connue d'avance. Pour plus d'informations, voir «Validation», à la page 5.

Egale pour toutes les catégories
Utilisez ce paramètre si les catégories de la variable dépendante sont distribuées dans des proportions égales entre toutes les catégories de population. Par exemple, s'il existe quatre catégories, environ 25 % des observations doivent se trouver dans chaque catégorie.

Personnalisée
Saisissez une valeur non négative pour chacune des catégories de la variable dépendante répertoriées dans la grille. Ces valeurs peuvent être des proportions, des pourcentages, des effectif de fréquences ou toute autre valeur représentant la distribution de valeurs entre les catégories.

Ajuster les probabilités a priori en utilisant les coûts de mauvaise réaffectation
Si vous définissez des coûts de mauvaise réaffectation, vous pouvez ajuster les probabilités a priori en fonction de ces coûts. Pour plus d'informations, voir «Coûts de classification erronée», à la page 9.

Remarque : Cette boîte de dialogue requiert que soient définis des libellés de valeurs pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante catégorielle disposent de libellés de valeur définis.

Spécification de probabilités a priori

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante catégorielle (nominale, ordinale) avec deux libellés de valeur définis, ou plus.
2. Pour **Méthode de croissance**, sélectionnez **CRT** ou **QUEST**.
3. Cliquez sur **Probabilités a priori**.

Scores

Dans CHAID et CHAID exhaustif avec une variable dépendante ordinale, vous pouvez attribuer des scores personnalisés à chaque catégorie de la variable dépendante. Les scores définissent la distance entre les catégories de la variable dépendante ainsi que l'ordre de ces catégories. Les scores peuvent être utilisés pour augmenter ou réduire la distance relative entre des valeurs ordinales ou pour changer l'ordre de ces valeurs.

Utiliser le rang ordinal de chaque catégorie
Le score de 1 est attribué à la catégorie la plus basse de la variable dépendante, le score de 2 est attribué à la catégorie supérieure suivante, etc. Il s'agit de la valeur par défaut.

Personnalisé
Saisissez une valeur de score numérique pour chacune des catégories de la variable dépendante répertoriées dans la grille.

Remarque : Cette boîte de dialogue requiert des libellés de valeur définis pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante catégorielle disposent de libellés de valeur définis.

Exemple

Tableau 2. Valeurs des scores personnalisés

Libellé de valeur	Valeur d'origine	Score
Ouvrier spécialisé	1	1
Ouvrier qualifié	2	4
Employé de bureau	3	4,5
Professionnels	4	7
Direction	5	6

- Les scores augmentent la distance relative entre les *ouvriers spécialisés* et les *ouvriers qualifiés* et réduit la distance relative entre les *ouvriers qualifiés* et les *employés de bureau*.
- Les scores inversent l'ordre de la *direction* et des *professionnels*.

Spécification de scores

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez une variable dépendante ordinale avec deux libellés de valeur définis, ou plus.
2. Pour **Méthode de croissance**, sélectionnez **CHAID** ou **Exhaustive CHAID**.
3. Cliquez sur **Scores**.

Valeurs manquantes

La boîte de dialogue Valeurs manquantes contrôle le traitement des valeurs nominales, des valeurs manquantes spécifiées par l'utilisateur et des valeurs de variable indépendante (prédicteur).

- La gestion des valeurs de variable indépendante manquantes spécifiées par l'utilisateur, d'échelle et ordinales, varie en fonction de la méthode de croissance.
- Le traitement des variables nominales dépendantes est spécifié dans la boîte de dialogue Catégories. Pour plus d'informations, voir «Sélection de catégories», à la page 4.
- Pour les variables d'échelle dépendantes et ordinales, les observations comportant des valeurs de variable dépendante système manquantes ou spécifiées par l'utilisateur sont toujours exclues.

Valeurs manquantes de l'utilisateur pour les variables indépendantes nominales

Traiter en tant que valeurs manquantes
Les valeurs manquantes de l'utilisateur sont traitées comme des valeurs système manquantes. La gestion des valeurs système manquantes varie selon les méthodes de croissance.

Traiter en tant que valeurs valides
Les valeurs manquantes de l'utilisateur des variables indépendantes nominales sont traitées comme des valeurs classiques pour la construction de l'arbre et la classification.

Règles dépendantes de la méthode

Si certaines valeurs de variable indépendante, mais pas toutes, sont manquantes par défaut ou spécifiées par l'utilisateur :

- Pour CHAID et CHAID exhaustif, les valeurs de variable indépendante manquantes par défaut ou spécifiées par l'utilisateur sont incluses dans l'analyse en tant que catégorie unique combinée. Pour les variables d'échelle indépendantes ou ordinales, les algorithmes génèrent d'abord les catégories en utilisant des valeurs valides, puis choisissent de fusionner la catégorie manquante avec la catégorie (valide) la plus ressemblante ou de la conserver à part.
- Pour CRT et QUEST, les observations comportant des valeurs de variable indépendante manquantes sont exclues du processus de construction de l'arbre mais sont classées à l'aide de valeurs de substitution, si la méthode inclut les valeurs de substitution. Si les valeurs manquantes nominales de

l'utilisateur sont traitées comme manquantes, elles seront également gérées comme telles. Pour plus d'informations, voir «Valeurs de substitution», à la page 9.

Détermination du traitement des valeurs manquantes indépendantes nominales spécifiées par l'utilisateur

1. Dans la boîte de dialogue Arbre de décisions principale, sélectionnez au moins une variable indépendante nominale.
2. Cliquez sur **Valeurs manquantes**.

Enregistrement des informations du modèle

Vous pouvez enregistrer les informations du modèle sous forme de variables dans le fichier de travail et enregistrer également l'intégralité du modèle au format XML (PMML) vers un fichier externe.

Variables enregistrées

Nombre de noeuds terminaux
Noeud terminal auquel chaque observation est affectée. La valeur est le nombre de noeuds de l'arbre.

Valeur prédite
Classe (groupe) ou valeur de la variable dépendante prévue par le modèle.

Probabilités prédites
Probabilité associée à la prévision du modèle. Une variable est enregistrée pour chaque catégorie de la variable dépendante. N'est pas disponible pour les variables d'échelle dépendantes.

Affectation des échantillons (d'apprentissage/de test)
Pour la validation par partition d'échantillon, cette variable indique si l'observation a été utilisée dans l'échantillon d'apprentissage ou l'échantillon de test. Sa valeur est 1 pour l'échantillon d'apprentissage et 0 pour l'échantillon de test. N'est pas disponible sauf si vous avez sélectionné la validation par partition d'échantillon. Pour plus d'informations, voir «Validation», à la page 5.

Exporter le modèle d'arbre au format XML
Vous pouvez enregistrer l'intégralité du modèle d'arbre au format XML (PMML). Vous pouvez utiliser ce fichier de modèle pour appliquer les informations du modèle aux autres fichiers de données à des fins d'évaluation.

Echantillon d'apprentissage
Ecrit le modèle sur le fichier indiqué. Pour les arbres validés par partition, il s'agit du modèle de l'échantillon d'apprentissage.

Echantillon de test
Ecrit le modèle de l'échantillon de test sur le fichier indiqué. N'est pas disponible sauf si vous avez sélectionné la validation par partition d'échantillon.

Sortie

Les options des sorties disponibles dépendent de la méthode de croissance, du niveau de mesure de la variable dépendante et d'autres paramètres.

Affichage des arbres

Vous pouvez contrôler l'apparence initiale de l'arbre ou supprimer complètement l'affichage de l'arbre.

Arbre Par défaut, le diagramme d'arbre est inclus dans la sortie affichée dans l'onglet Sortie. Décochez cette option pour l'exclure de la sortie.

Affichage
Ces options contrôlent l'apparence initiale du diagramme d'arbre dans l'onglet Sortie. Vous pouvez également modifier tous ces attributs en modifiant l'arbre créé.

Orientation
Vous pouvez afficher l'arbre de haut en bas avec le noeud racine en haut, de gauche à droite ou de droite à gauche.

Contenu du noeud
Les noeuds peuvent afficher des tableaux, des graphiques ou les deux. Pour les variables dépendantes catégorielles, les tableaux affichent les effectif de fréquences et les pourcentages, et les graphiques sont des graphiques à barres. Pour les variables d'échelle dépendantes, les tableaux affichent les moyennes, les écarts types, le nombre d'observations et les prévisions. Les graphiques sont des histogrammes.

Echelle
Par défaut, les arbres volumineux sont automatiquement réduits avec conservation des proportions pour que l'arbre tienne dans la page. Vous pouvez indiquer un pourcentage d'échelle personnalisé allant jusqu'à 200 %.

Statistiques des variables indépendantes
Pour CHAID et CHAID exhaustif, les statistiques comprennent la valeur *F* (pour les variables d'échelle dépendantes) ou la valeur khi-carré (pour les variables dépendantes catégorielles) ainsi que la valeur de signification et les degrés de liberté. Pour CRT, la valeur d'amélioration est affichée. Pour QUEST, la valeur *F*, la valeur de signification et les degrés de liberté sont affichés pour les variables indépendantes ordinales et d'échelle ; pour les variables indépendantes nominales, la valeur khi-carré, la valeur de signification et les degrés de liberté sont affichés.

Définitions des noeuds
Les définitions de noeud affichent les valeurs de la variable indépendante utilisée à chaque scission des noeuds.

Arbre sous forme de tableau
Informations récapitulatives de chaque noeud de l'arbre, dont le nombre de noeuds parent, les statistiques de variable indépendante, les valeurs de variable indépendante pour le noeud, la moyenne et l'écart type pour les variables d'échelle dépendantes, ou les effectifs et les pourcentages pour les variables dépendantes catégorielles.

Contrôle de l'affichage initial de l'arbre

1. Dans la boîte de dialogue Arbre de décisions principale, cliquez sur **Arbre**.

Statistiques

Les tableaux de statistiques disponibles dépendent du niveau de mesure de la variable dépendante, de la méthode de croissance et d'autres paramètres.

Modèle

Récapitulatif
Le récapitulatif comprend la méthode utilisée, les variables incluses dans le modèle et les variables indiquées mais non incluses dans le modèle.

Risque
Estimation du risque et de l'erreur standard. Mesure de l'exactitude des prévisions de l'arbre.

- Pour les variables dépendantes catégorielles, l'estimation du risque correspond à la proportion d'observations mal classées après ajustement aux probabilités a priori et aux coûts de mauvaise réaffectation.
- Pour les variables d'échelle dépendantes, l'estimation du risque correspond à la variance intra-noeud.

Table de classification

Pour les variables dépendantes catégorielles (nominales, ordinales), cette table comporte le nombre d'observations classées correctement et incorrectement pour chaque catégorie de la variable dépendante. N'est pas disponible pour les variables d'échelle dépendantes.

Valeurs de coût, de probabilité a priori, de score et de bénéfice

Pour les variables dépendantes catégorielles, ce tableau comporte les valeurs de coût, de probabilité a priori, de score et de bénéfice utilisées pour l'analyse. N'est pas disponible pour les variables d'échelle dépendantes.

Variables indépendantes

Importance par rapport au modèle

Pour la méthode de croissance CRT, classe chaque variable indépendante (prédicteur) selon son importance dans le modèle. N'est pas disponible pour les méthodes QUEST ou CHAID.

Valeurs de substitution par scission

Pour les méthodes de croissance CRT et QUEST, si le modèle inclut les valeurs de substitution, répertorie les valeurs de substitution de chaque partition de l'arbre. N'est pas disponible pour les méthodes CHAID. Pour plus d'informations, voir «Valeurs de substitution», à la page 9.

Résultats des noeuds

Récapitulatif

Pour les variables d'échelle dépendantes, le tableau comporte le nombre de noeuds, le nombre d'observations et la valeur moyenne de la variable dépendante. Pour les variables dépendantes catégorielles dont les bénéfices sont définis, le tableau comporte le nombre de noeuds, le nombre d'observations, la moyenne des bénéfices et les valeurs du ROI (retour sur investissement). N'est pas disponible pour les variables dépendantes catégorielles dont les bénéfices ne sont pas définis. Pour plus d'informations, voir «Bénéfices», à la page 10.

Par catégorie cible

Pour les variables dépendantes catégorielles dont les catégories cible sont définies, le tableau comporte le pourcentage de gains, le pourcentage de réponses et le pourcentage d'index (lift) par noeud ou groupe de percentiles. Un tableau distinct est produit pour chaque catégorie cible. N'est pas disponible pour les variables d'échelle dépendantes ou catégorielles dont les catégories cible ne sont pas définies. Pour plus d'informations, voir «Sélection de catégories», à la page 4.

Lignes

Les tableaux de résultats des noeuds peuvent afficher les résultats par noeuds terminaux, par percentiles ou les deux. Si vous sélectionnez les deux, vous obtenez deux tableaux, un pour chaque catégorie cible. Les tableaux utilisant des percentiles comportent des valeurs cumulatives pour chaque percentile, dans l'ordre du tri.

Ordre de tri

La valeur varie en fonction du niveau de mesure de la variable dépendante, et elle est différente pour le récapitulatif des gains et le tableau des gains.

Incrément de percentile

Pour les tableaux utilisant des percentiles, vous pouvez sélectionner l'incrément de percentiles suivant : 1, 2, 5, 10, 20 ou 25.

Afficher les statistiques cumulées.

Pour les tableaux utilisant des noeuds terminaux, ajoute une colonne comportant les résultats cumulés.

Sélection de la sortie des statistiques

1. Dans la boîte de dialogue Arbre de décisions principale, cliquez sur **Statistiques**.

Graphiques

Les graphiques disponibles dépendent du niveau de mesure de la variable dépendante, de la méthode de croissance et d'autres paramètres.

Importance de la variable indépendante dans le modèle
Graphique à barres représentant l'importance dans le modèle de chaque variable indépendante (prédicteur). Valable uniquement pour la méthode de croissance CRT.

Résultats des noeuds

Gain Le gain est le pourcentage d'observations totales de la catégorie cible dans chaque noeud, calculé de la manière suivante : (cibles des noeuds *n* / nombre total de cibles *n*) x 100. Le graphique des gains est un graphique curviligne représentant les gains cumulés en percentiles, calculé de la manière suivante : (cibles des percentiles cumulés *n* / nombre total de cibles *n*) x 100. Un graphique curviligne distinct est produit pour chaque catégorie cible. Est uniquement disponible pour les variables dépendantes catégorielles dont les catégories cible sont définies. Pour plus d'informations, voir «Sélection de catégories», à la page 4.

Le tracé des gains trace point par point les valeurs de la colonne *Pourcentage de gain* du tableau Gains pour les percentiles, qui comporte également les valeurs cumulées.

Index L'index correspond au rapport du pourcentage de réponses du noeud pour la catégorie cible comparé au pourcentage de réponses global pour la catégorie cible de l'ensemble de l'échantillon. Le graphique des index est un graphique curviligne représentant les valeurs de l'index des percentiles cumulés. Est uniquement disponible pour les variables dépendantes catégorielles. L'index des percentiles cumulés est calculé de la manière suivante : (pourcentage de réponse des percentiles cumulés / pourcentage total de réponses) x 100. Un graphique distinct est produit pour chaque catégorie cible et les catégories cible doivent être définies.

Le graphique d'index trace point par point les valeurs de la colonne *Index* du tableau Gains pour les percentiles.

Réponse
Pourcentage d'observations dans le noeud de la catégorie cible spécifiée. Le graphique de réponse est un graphique curviligne représentant les réponses des percentiles cumulées, calculé de la manière suivante : (cibles de percentiles cumulés *n* / nombre total de percentiles cumulés *n*) x 100. Est uniquement disponible pour les variables dépendantes catégorielles dont les catégories cible sont définies.

Le graphique de réponse trace point par point les valeurs de la colonne *Réponse* du tableau Gains pour les percentiles.

Moyenne
Graphique curviligne représentant les valeurs moyennes des percentiles cumulés pour la variable dépendante. Est uniquement disponible pour les variables d'échelle dépendantes.

Bénéfice moyen
Graphique curviligne représentant les profits moyens cumulés. Disponible uniquement pour les variables dépendantes catégorielles dont les bénéfices sont définis. Pour plus d'informations, voir «Bénéfices», à la page 10.

Le graphique des profits moyens trace point par point les valeurs de la colonne *Bénéfices* de la table récapitulative des gains pour les percentiles.

Retour sur investissement (ROI)
Graphique curviligne du ROI (retour sur investissement) cumulé. ROI est le rapport recettes/dépenses. Disponible uniquement pour les variables dépendantes catégorielles dont les bénéfices sont définis.

Le graphique du ROI trace point par point les valeurs de la colonne *ROI* de la table récapitulative des gains pour les percentiles.

Incrément de percentile
Pour tous les graphiques utilisant des percentiles, ce paramètre contrôle l'affichage des incréments des percentiles sur le graphique : 1, 2, 5, 10, 20 ou 25.

Sélection de la sortie des graphiques

1. Dans la boîte de dialogue Arbre de décisions principale, cliquez sur **Tracés**.

Règles de sélection et d'évaluation

La boîte de dialogue Règles permet de générer des règles de sélection ou de classification/prévision sous la forme de syntaxe de commande, au format SQL ou sous forme de texte simple (standard). Vous pouvez afficher ces règles dans l'onglet Sortie et/ou les enregistrer dans un fichier externe.

Générer des règles de classification
Sélectionnez cette option pour activer la définition de règles de sélection et d'évaluation.

Syntaxe
Contrôle la forme des règles de sélection des sorties affichées dans l'onglet Sortie et des règles de sélection enregistrées dans un fichier externe.

SPSS Statistics
Langage de syntaxe de commande. Les règles sont exprimées sous la forme d'un ensemble de commandes définissant une condition de filtre pouvant être utilisée pour sélectionner des sous-ensembles d'observations ou sous la forme d'instructions `COMPUTE` pouvant être utilisées pour analyser les observations.

SQL Les règles SQL standard sont générées pour sélectionner des enregistrements dans la base de données, pour les extraire ou pour attribuer des valeurs à ces enregistrements. Les règles SQL générées ne comportent aucun nom de tableau ou aucune autre information de source de données.

Texte simple
Pseudo-code pour la langue standard. Les règles sont exprimées sous forme d'instructions logiques "si...alors" décrivant les classifications et les prévisions du modèle pour chaque noeud. Sous cette forme, les règles peuvent utiliser des libellés de valeur ou de variable définis, ou des noms de variables et des valeurs de données.

Type Pour IBM® SPSS Statistics et les règles SQL, contrôle le type de règles générées : règles de sélection ou d'évaluation.

Attribuer des valeurs aux observations
Les règles peuvent être utilisées pour attribuer les prévisions du modèle aux observations respectant les critères d'appartenance aux noeuds. Une règle distincte est créée pour chaque observation respectant les critères d'appartenance aux noeuds.

Sélectionner des observations
Les règles peuvent être utilisées pour sélectionner les observations respectant les critères d'appartenance aux noeuds. Pour les règles IBM SPSS Statistics et SQL, une règle unique est créée pour sélectionner toutes les observations respectant les critères de sélection.

Inclure les valeurs de substitution dans SPSS Statistics et les règles SQL
Pour CRT et QUEST, vous pouvez inclure des prédicteurs de substitution provenant du modèle dans les règles. Les règles comportant des valeurs de substitution peuvent être relativement complexes. En général, si vous souhaitez simplement dégager des

informations conceptuelles sur votre arbre, excluez les valeurs de substitution. Si certaines observations comportent des données de variable indépendante (prédicteur) incomplètes et que vous souhaitez que les règles reproduisent votre arbre, incluez les valeurs de substitution. Pour plus d'informations, voir «Valeurs de substitution», à la page 9.

Noeuds

Contrôle le champ d'application des règles créées. Une règle distincte est créée pour chaque noeud inclus dans le champ d'application.

Tous les noeuds terminaux
Génère des règles pour chaque noeud terminal.

Meilleurs noeuds terminaux
Génère des règles pour les *n* noeuds terminaux les plus hauts selon les valeurs d'index. Si le nombre dépasse le nombre de noeuds terminaux de l'arbre, les règles sont créées pour tous les noeuds terminaux.

Meilleurs noeuds terminaux jusqu'à un pourcentage spécifié d'observations.
Génère des règles pour les noeuds terminaux pour les *n* pourcentages d'observations les plus hauts selon les valeurs d'index.

Noeuds terminaux dont la valeur d'index est supérieure ou égale à une valeur de césure.
Génère des règles pour tous les noeuds terminaux dont la valeur d'index est supérieure ou égale à la valeur spécifiée. Une valeur d'index supérieure à 100 signifie que le pourcentage d'observations dans la catégorie cible de ce noeud dépasse le pourcentage du noeud racine.

Tous les noeuds
Génère des règles pour tous les noeuds.

Remarques :

- La sélection de noeuds basée sur les valeurs d'index est uniquement disponible pour les variables dépendantes catégorielles comportant des catégories cible définies. Si vous avez indiqué plusieurs catégories cible, un jeu de règles distinct est créé pour chaque catégorie cible.
- Pour IBM SPSS Statistics et les règles SQL de sélection d'observations (et non celles d'affectation de valeurs), **Tous les noeuds** et **Tous les noeuds terminaux** génèrent en fait une règle sélectionnant toutes les observations utilisées dans l'analyse.

Exporter les règles dans un fichier
Enregistre les règles dans un fichier texte externe.

Vous pouvez également générer et enregistrer les règles de sélection ou d'évaluation de manière interactive, en fonction des noeuds sélectionnés dans le modèle d'arbre final.

Remarque : Si vous appliquez des règles sous forme de syntaxe de commande à un autre fichier de données, ce fichier de données doit contenir des variables portant les mêmes noms que les variables indépendantes incluses dans le modèle final, mesurées avec la même unité, comportant les mêmes valeurs manquantes spécifiées par l'utilisateur (s'il en existe).

Spécification de règles de sélection ou d'évaluation

1. Dans la boîte de dialogue Arbre de décisions principale, cliquez sur **Règles**.

Remarques

Le présent document a été développé pour des produits et des services proposés aux Etats-Unis et peut être mis à disposition par IBM dans d'autres langues. Toutefois, il peut être nécessaire de posséder une copie du produit ou de la version du produit dans cette langue pour pouvoir y accéder.

Le présent document peut contenir des informations ou des références concernant certains produits, logiciels ou services IBM non annoncés dans ce pays. Pour plus de détails, référez-vous aux documents d'annonce disponibles dans votre pays, ou adressez-vous à votre partenaire commercial IBM. Toute référence à un produit, logiciel ou service IBM n'implique pas que seul ce produit, logiciel ou service puisse être utilisé. Tout autre élément fonctionnellement équivalent peut être utilisé, s'il n'enfreint aucun droit d'IBM. Il est de la responsabilité de l'utilisateur d'évaluer et de vérifier lui-même les installations et applications réalisées avec des produits, logiciels ou services non expressément référencés par IBM.

IBM peut détenir des brevets ou des demandes de brevet couvrant les produits mentionnés dans le présent document. La remise de ce document ne vous donne aucun droit de licence sur ces brevets ou demandes de brevet. Si vous désirez recevoir des informations concernant l'acquisition de licences, veuillez en faire la demande par écrit à l'adresse suivante :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A

Pour le Canada, veuillez adresser votre courrier à :

IBM Director of Commercial Relations
IBM Canada Ltd.
3600 Steeles Avenue East
Markham, Ontario
L3R 9Z7 Canada

Les informations sur les licences concernant les produits utilisant un jeu de caractères double octet peuvent être obtenues par écrit à l'adresse suivante :

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

LE PRESENT DOCUMENT EST LIVRE EN L'ETAT SANS AUCUNE GARANTIE EXPLICITE OU IMPLICITE. IBM DECLINE NOTAMMENT TOUTE RESPONSABILITE RELATIVE A CES INFORMATIONS EN CAS DE CONTREFAÇON AINSI QU'EN CAS DE DEFAUT D'APTITUDE A L'EXECUTION D'UN TRAVAIL DONNE. Certaines juridictions n'autorisent pas l'exclusion des garanties implicites, auquel cas l'exclusion ci-dessus ne vous sera pas applicable.

Le présent document peut contenir des inexactitudes ou des coquilles. Ce document est mis à jour périodiquement. Chaque nouvelle édition inclut les mises à jour. IBM peut, à tout moment et sans préavis, modifier les produits et logiciels décrits dans ce document.

Les références à des sites Web non IBM sont fournies à titre d'information uniquement et n'impliquent en aucun cas une adhésion aux données qu'ils contiennent. Les éléments figurant sur ces sites Web ne font pas partie des éléments du présent produit IBM et l'utilisation de ces sites relève de votre seule responsabilité.

IBM pourra utiliser ou diffuser, de toute manière qu'elle jugera appropriée et sans aucune obligation de sa part, tout ou partie des informations qui lui seront fournies.

Les licenciés souhaitant obtenir des informations permettant : (i) l'échange des données entre des logiciels créés de façon indépendante et d'autres logiciels (dont celui-ci), et (ii) l'utilisation mutuelle des données ainsi échangées, doivent adresser leur demande à :

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A

Ces informations peuvent être soumises à des conditions particulières, prévoyant notamment le paiement d'une redevance.

Le logiciel sous licence décrit dans ce document et tous les éléments sous licence disponibles s'y rapportant sont fournis par IBM conformément aux dispositions du Livret Contractuel IBM, des Conditions Internationales d'Utilisation de Logiciels IBM, des Conditions d'Utilisation du Code Machine ou de tout autre contrat équivalent.

Les données de performances et les exemples de clients sont fournis à titre d'exemple uniquement. Les performances réelles peuvent varier en fonction des configurations et des conditions d'exploitation spécifiques.

Les informations concernant des produits non IBM ont été obtenues auprès des fournisseurs de ces produits, par l'intermédiaire d'annonces publiques ou via d'autres sources disponibles. IBM n'a pas testé ces produits et ne peut confirmer l'exactitude de leurs performances ni leur compatibilité. Elle ne peut recevoir aucune réclamation concernant des produits non IBM. Toute question concernant les performances de produits non IBM doit être adressée aux fournisseurs de ces produits.

Toute instruction relative aux intentions d'IBM pour ses opérations à venir est susceptible d'être modifiée ou annulée sans préavis, et doit être considérée uniquement comme un objectif.

Le présent document peut contenir des exemples de données et de rapports utilisés couramment dans l'environnement professionnel. Ces exemples mentionnent des noms fictifs de personnes, de sociétés, de marques ou de produits à des fins illustratives ou explicatives uniquement. Toute ressemblance avec des noms de personnes, de sociétés ou des données réelles serait purement fortuite.

LICENCE DE COPYRIGHT :

Le présent logiciel contient des exemples de programmes d'application en langage source destinés à illustrer les techniques de programmation sur différentes plateformes d'exploitation. Vous avez le droit de copier, de modifier et de distribuer ces exemples de programmes sous quelque forme que ce soit et sans paiement d'aucune redevance à IBM, à des fins de développement, d'utilisation, de vente ou de distribution de programmes d'application conformes aux interfaces de programmation des plateformes pour lesquels ils ont été écrits ou aux interfaces de programmation IBM. Ces exemples de programmes n'ont pas été rigoureusement testés dans toutes les conditions. Par conséquent, IBM ne peut garantir expressément ou implicitement la fiabilité, la maintenabilité ou le fonctionnement de ces programmes. Les exemples de programmes sont fournis "EN L'ETAT", sans garantie d'aucune sorte. IBM ne sera en aucun cas responsable des dommages liés à l'utilisation des exemples de programmes.

Toute copie totale ou partielle de ces programmes exemples et des oeuvres qui en sont dérivées doit comprendre une notice de copyright, libellée comme suit :

Marques

IBM, le logo IBM et ibm.com sont des marques d'International Business Machines Corp. dans de nombreux pays. Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés. La liste actualisée de toutes les marques d'IBM est disponible sur la page Web "Copyright and trademark information" à l'adresse www.ibm.com/legal/copytrade.shtml.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques d'Adobe Systems Incorporated aux Etats-Unis et/ou dans certains autres pays.

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 d'Intel Corporation ou de ses filiales aux Etats-Unis et/ou dans certains autres pays.

Linux est une marque de Linus Torvalds aux Etats-Unis et/ou dans certains autres pays.

Microsoft, Windows, Windows NT et le logo Windows sont des marques de Microsoft Corporation aux Etats-Unis et/ou dans certains autres pays.

UNIX est une marque enregistrée de The Open Group aux Etats-Unis et/ou dans certains autres pays.

Java ainsi que toutes les marques et tous les logos incluant Java sont des marques d'Oracle et/ou de ses sociétés affiliées.

Index