

IBM SPSS Decision Trees 20



Remarque : Avant d'utiliser ces informations et le produit qu'elles concernent, lisez les informations générales sous Remarques sur p. 114.

Cette version s'applique à IBM® SPSS® Statistics 20 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 1989, 2011.

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® Statistics est un système complet d'analyse de données. Le module complémentaire facultatif Decision Trees fournit les techniques d'analyse supplémentaires décrites dans ce manuel. Le module complémentaire Decision Trees doit être utilisé avec le système central SPSS Statistics auquel il est entièrement intégré.

A propos de IBM Business Analytics

Le logiciel IBM Business Analytics offre des informations complètes, cohérentes et précises permettant aux preneurs de décision d'améliorer leurs performances professionnelles. Un portefeuille complet de solutions de [business intelligence](#), [d'analyses prédictives](#), [de performance financière et de gestion de la stratégie](#), et [d'applications analytiques](#) permet une connaissance claire et immédiate et offre des possibilités d'actions sur les performances actuelles et la capacité de prédire les résultats futurs. En combinant des solutions du secteur, des pratiques prouvées et des services professionnels, les entreprises de toute taille peuvent générer la plus grande productivité, automatiser les décisions en toute confiance et apporter de meilleurs résultats.

Dans le cadre de ce portefeuille, le logiciel IBM SPSS Predictive Analytics aide les entreprises à prédire des événements futurs et à agir de manière proactive en fonction de ces prédictions pour apporter de meilleurs résultats. Des clients dans les domaines commerciaux, gouvernementaux et académiques se servent de la technologie IBM SPSS comme d'un avantage concurrentiel pour attirer ou retenir des clients, tout en réduisant les risques liés à l'incertitude et à la fraude. En intégrant le logiciel IBM SPSS à leurs opérations quotidiennes, les entreprises peuvent effectuer des prévisions, et sont capables de diriger et d'automatiser leurs décisions afin d'atteindre leurs objectifs commerciaux et d'obtenir des avantages concurrentiels mesurables. Pour plus d'informations ou pour contacter un représentant, visitez le site <http://www.ibm.com/spss>.

Support technique

Un support technique est disponible pour les clients du service de maintenance. 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 contacter l'assistance technique, visitez le site IBM Corp. à l'adresse <http://www.ibm.com/support>. Votre nom, celui de votre société, ainsi que votre contrat d'assistance vous seront demandés.

Support technique pour les étudiants

Si vous êtes un étudiant qui utilise la version pour étudiant, personnel de l'éducation ou diplômé d'un produit logiciel IBM SPSS, veuillez consulter les pages [Solutions pour l'éducation](#) (<http://www.ibm.com/spss/rd/students/>) consacrées aux étudiants. Si vous êtes un étudiant utilisant une copie du logiciel IBM SPSS fournie par votre université, veuillez contacter le coordinateur des produits IBM SPSS de votre université.

Service clients

Si vous avez des questions concernant votre livraison ou votre compte, contactez votre bureau local. Veuillez préparer et conserver votre numéro de série à portée de main pour l'identification.

Séminaires de formation

IBM Corp. propose des séminaires de formation, publics et sur site. Tous les séminaires font appel à des ateliers de travaux pratiques. Ces séminaires seront proposés régulièrement dans les grandes villes. Pour plus d'informations sur ces séminaires, accédez au site <http://www.ibm.com/software/analytics/spss/training>.

Documents supplémentaires

Les ouvrages *SPSS Statistics : Guide to Data Analysis*, *SPSS Statistics : Statistical Procedures Companion*, et *SPSS Statistics : Advanced Statistical Procedures Companion*, écrits par Marija Norušis et publiés par Prentice Hall, sont suggérés comme documentation supplémentaire. Ces publications présentent les procédures statistiques des modules SPSS Statistics Base, Advanced Statistics et Regression. Que vous soyez novice dans les analyses de données ou prêt à utiliser des applications plus avancées, ces ouvrages vous aideront à exploiter au mieux les fonctionnalités offertes par IBM® SPSS® Statistics. Pour obtenir des informations supplémentaires y compris le contenu des publications et des extraits de chapitres, visitez le site web de l'auteur : <http://www.norusis.com>

Contenu

Partie I: Guide de l'utilisateur

1 Création d'arbres décision 1

Sélection de modalités	6
Validation	8
Critères de croissance de l'arbre	9
Limites de croissance	9
Critères CHAID	10
Critères CRT	13
Critères QUEST	14
Elagage des arbres	15
Valeurs de substitution	16
Options.	16
Coûts de classification erronée	17
Bénéfices	18
Probabilités a priori	20
Scores.	21
Valeurs manquantes	23
Enregistrement des informations du modèle	24
Résultats	25
Affichage des arbres	26
Statistiques	28
Diagrammes	32
Règles de sélection et d'analyse	38

2 Editeur d'arbre 41

Manipulation de grands arbres	42
Carte d'arbre	43
Mise à l'échelle de l'affichage de l'arbre	44
Fenêtre Récapitulatif des noeuds.	44
Contrôle des informations affichées dans l'arbre	45
Modification des couleurs et des polices de caractères du texte des arbres.	46
Règles de sélection et d'analyse des observations	49
Filtrage des observations.	49
Enregistrement des règles de sélection et d'analyse	49

Partie II: Exemples

3 Hypothèses et exigences concernant les données 53

Effets du niveau de mesure sur les modèles d'arbre	53
Affectation permanente du niveau de mesure	56
Variables avec niveau de mesure inconnu	57
Effets des étiquettes de valeur sur les modèles d'arbre	57
Affectation d'étiquettes de valeur à toutes les valeurs	59

4 Utilisation des arbres de décision pour évaluer le risque de crédit 61

Création du modèle	61
Construction du modèle d'arbre CHAID	61
Sélection des modalités cible	62
Spécification des critères de croissance de l'arbre	63
Sélection de types de sortie supplémentaires	64
Enregistrement de prévisions	66
Evaluation du modèle	67
Tableau récapitulatif des modèles	68
Diagramme de l'arbre	69
Tableau de l'arbre	70
Gains pour les noeuds	72
Diagramme des gains	73
Diagramme des index	73
Estimation du risque et classification	74
Prévisions	75
Amélioration du modèle	76
Sélection d'observations dans les noeuds	76
Examen des observations sélectionnées	77
Affectation de coûts aux résultats	79
Récapitulatif	83

5 Construction d'un modèle d'analyse 84

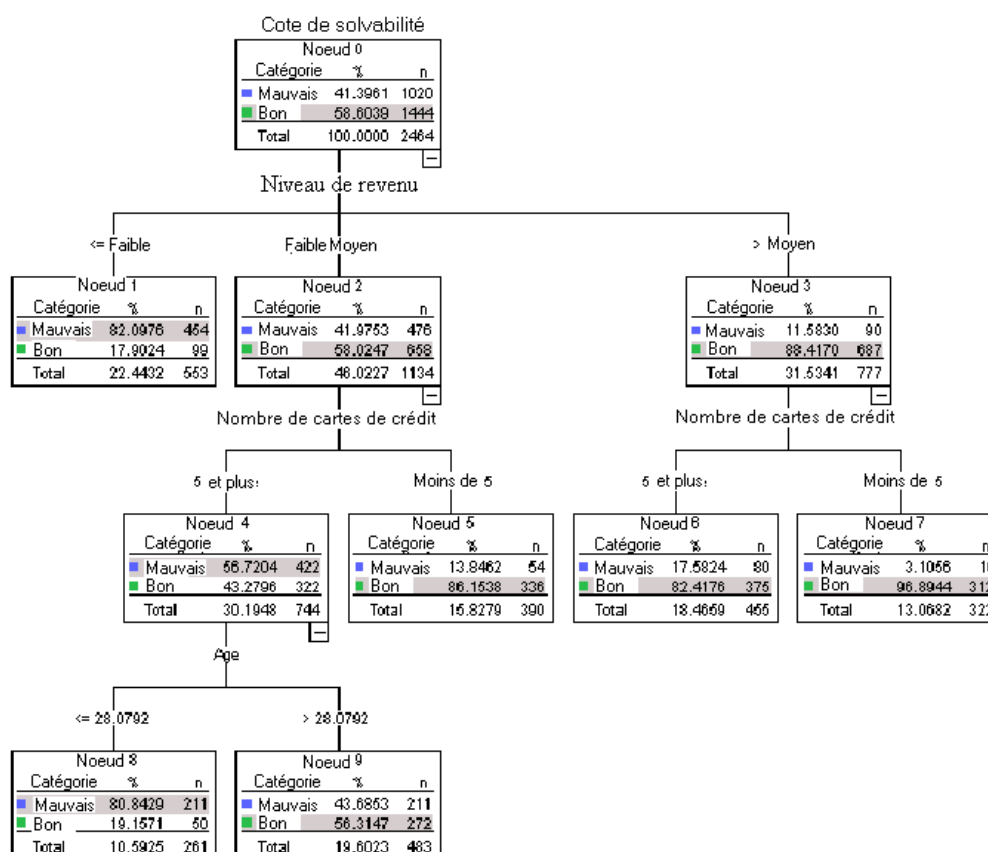
Construction du modèle	84
----------------------------------	----

Evaluation du modèle	86
Récapitulatif des modèles	86
Diagramme de modèle d'arbre	87
Estimation du risque	88
Application du modèle à un autre fichier de données	89
Récapitulatif	92
6 Valeurs manquantes dans les modèles d'arbre	93
Valeurs manquantes avec CHAID	94
Résultats CHAID	96
Valeurs manquantes avec CRT	97
Résultats CRT	100
Récapitulatif	102
Annexes	
A Fichiers d'exemple	103
B Remarques	114
Index	117

Partie I: Guide de l'utilisateur

Création d'arbres décision

Figure 1-1
Arbre décision



La procédure Arbre de décision 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 (prédites) 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 une catégorie.

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

Prédiction. 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 des données et analyse des variables. Sélectionne à partir d'un ensemble étendu de variables un sous-ensemble exploitable de variables explicatives 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 modalités et discrétisation des variables continues. Etablit un nouveau code de regroupement des modalités de variable explicative 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 la cote 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 avantages intéressants :

- 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.

Analyse des 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 modalités 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 modalités 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 modalités 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 des effectifs 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 une 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 étiquettes de valeur définies.

- **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.



Echelle



Nominales



Ordinales

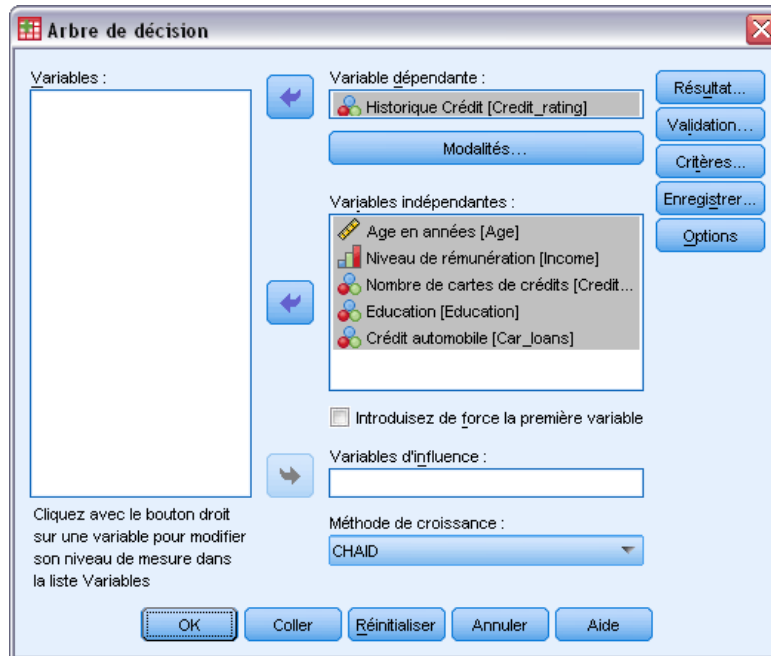
Pour modifier de manière temporaire le niveau de mesure d'une variable, cliquez sur la variable dans la liste des variables source avec le bouton droit de la souris et sélectionnez un niveau de mesure dans le menu contextuel.

- **Les étiquettes de valeurs.** L'interface de la boîte de dialogue de cette procédure considère soit que toutes les valeurs non manquantes d'une variable dépendante qualitative (nominale, ordinale) ont des étiquettes de valeur définies, soit qu'aucune d'entre elles n'en dispose. Certaines fonctions ne sont disponibles que si deux valeurs non manquantes au moins de la variable dépendante qualitative disposent d'étiquettes de valeur. Si au moins deux valeurs non manquantes disposent d'étiquettes de valeur définies, toutes les observations contenant d'autres valeurs ne disposant pas d'étiquettes de valeur seront exclues de l'analyse.

Pour obtenir des arbres de décision

- ▶ A partir des menus, sélectionnez :
Analyse > Classification > Arbre...

Figure 1-2
Boîte de dialogue Arbre de décision



- ▶ Sélectionnez une variable dépendante.
- ▶ Sélectionnez une ou plusieurs variables indépendantes.
- ▶ Sélectionnez une méthode de croissance.

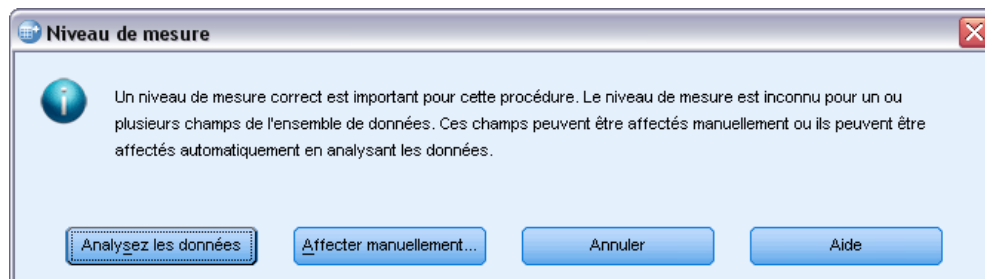
Sinon, vous pouvez :

- 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 apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble 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.

Figure 1-3
Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Modification des niveaux de mesure

- ▶ Cliquez avec le bouton droit sur la variable dans la liste source.
- ▶ Dans le menu contextuel, sélectionnez un niveau de mesure.

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

Méthodes de croissance

Les méthodes de croissance disponibles sont :

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

Exhaustive CHAID. Une version modifiée de CHAID qui examine toutes les scissions possibles pour chaque valeur prédite.

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 variables prédites avec de nombreuses modalités 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 :

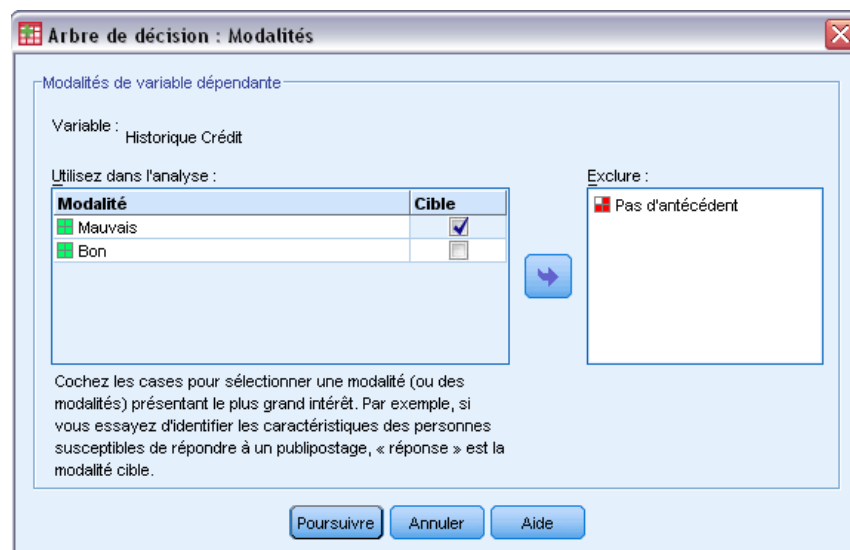
	CHAID*	CRT	QUEST
Calculé à partir du khi-deux**	X		
Variables (prédites) indépendantes de substitution		X	X
Elagage des arbres		X	X
Scission de noeud multiple	X		
Scission de noeud binaire		X	X
Variables d'influence	X	X	
Probabilités a priori		X	X
Coûts de classification erronée	X	X	X
Calcul rapide	X		X

*Inclut Exhaustive CHAID.

**QUEST utilise également une mesure du Khi-deux pour les variables indépendantes nominales.

Sélection de modalités

Figure 1-4
Boîte de dialogue Modalités



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

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

Inclure/Exclure des modalités

Vous pouvez limiter l'analyse à certaines modalités 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 modalités manquantes spécifiées par l'utilisateur dans l'analyse. (Par défaut, les modalités manquantes spécifiées par l'utilisateur s'affichent dans la liste Exclure.)

Modalités cible

Les modalités sélectionnées (qui sont cochées) sont traitées comme les modalités 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 modalité qui vous intéresse le plus, sélectionnez la modalité « mauvaise » cote de solvabilité en tant que modalité cible.

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

Modalités et étiquettes de valeurs

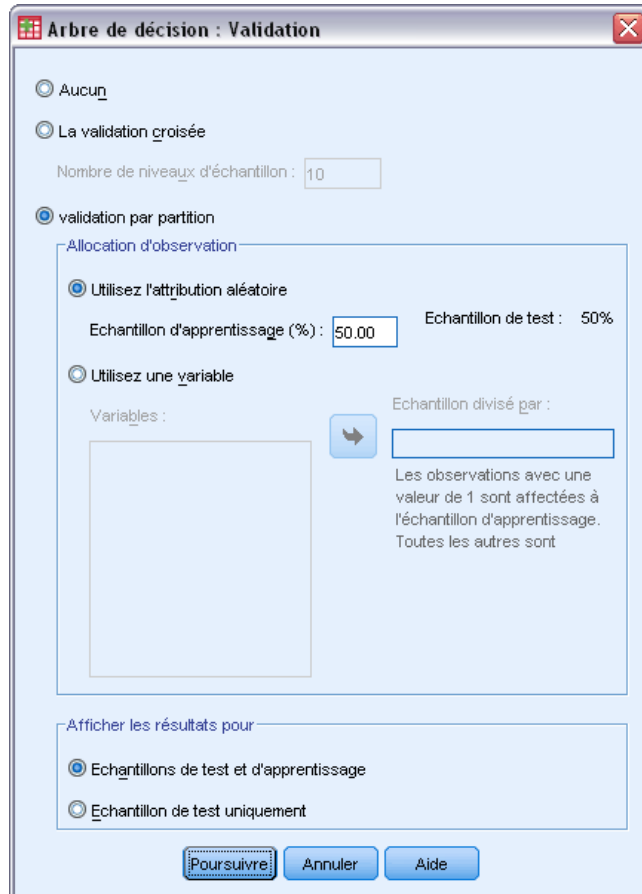
Cette boîte de dialogue requiert des étiquettes de valeur définies pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante qualitative disposent d'étiquettes de valeur définies.

Pour inclure/exclure des modalités et sélectionner des modalités cible

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante qualitative (nominale, ordinale) disposant d'au moins deux étiquettes de valeur définies.
- ▶ Cliquez sur Modalités.

Validation

Figure 1-5
Boîte de dialogue 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 partition.

La 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.

- 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

Pour la validation par partition, 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 partition 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 modalités peuvent ne pas comporter suffisamment d'observations pour construire correctement l'arbre.

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

Figure 1-6
Boîte de dialogue Critères, onglet Limites de croissance

The screenshot shows a dialog box titled 'Arbre de décision : Critères' with a close button (X) in the top right corner. It has three tabs: 'Limites de croissance' (selected), 'CHAID', and 'Intervalles'. The 'Limites de croissance' tab is divided into two main sections. The left section is titled 'Profondeur maximale de l'arborescence' and contains two radio buttons: 'Automatique' (selected) and 'Personnalisé'. Below 'Automatique' is the text 'Le nombre maximal de niveaux est de 3 pour CHAID ; 5 pour CRT et QUEST.' Below 'Personnalisé' is a text input field labeled 'Valeur :'. The right section is titled 'Nombre minimal d'observations.' and contains two text input fields: 'Noeud parent : 400' and 'Noeud enfant : 200'. At the bottom of the dialog box are three buttons: 'Poursuivre', 'Annuler', and 'Aide'.

L'onglet 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 maximum de l'arborescence : 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 Exhaustive CHAID, et à cinq niveaux pour les méthodes CRT et QUEST.

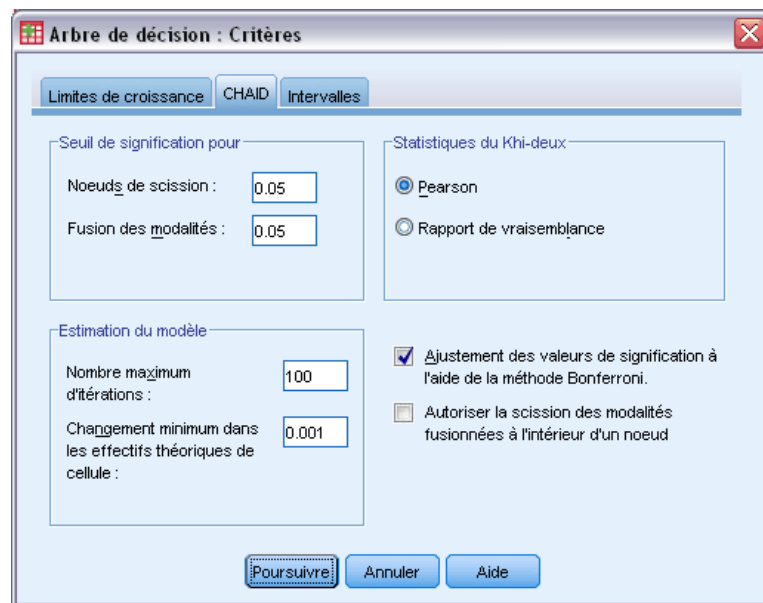
Nombre minimum d'observations. Contrôle le nombre minimum d'observations des noeuds. Les noeuds ne respectant pas ces critères ne sont pas scindés.

- 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.

Critères CHAID

Figure 1-7
Boîte de dialogue Critères, onglet CHAID



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

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

- Pour scinder des noeuds, cette valeur doit être supérieure à 0 et inférieure à 1. Les valeurs les plus basses produisent des arbres avec moins de noeuds.
- Pour la fusion des modalités, cette valeur doit être supérieure à 0 et inférieure ou égale à 1. Pour que les modalités ne fusionnent pas, indiquez la valeur 1. Pour une variable d'échelle indépendante, cela signifie que le nombre de modalités de la variable dans l'arbre final correspond au nombre d'intervalles indiqué (leur nombre par défaut est 10). [Pour plus d'informations, reportez-vous à la section Intervalles d'échelle pour l'analyse CHAID sur p. 12.](#)

Statistique du Khi-deux. Pour les variables dépendantes ordinales, le Khi-deux déterminant la scission des noeuds et la fusion des modalités 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 :

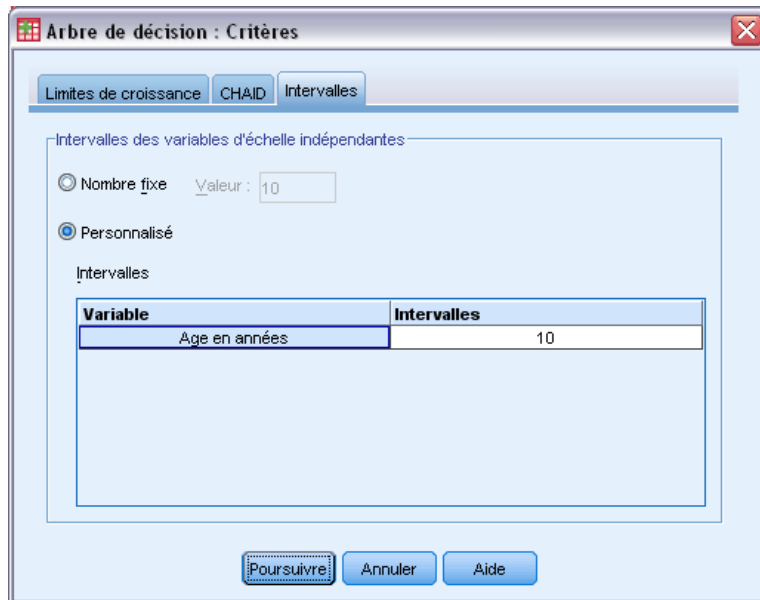
- **le nombre maximum des itérations.** La valeur par défaut est 100. Si l'arbre cesse de croître parce que le nombre maximum d'itérations a été atteint, vous pouvez augmenter ce maximum ou modifier d'autres critères contrôlant la croissance de l'arbre.
- **Changement minimum dans les effectifs théoriques 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 modalités fusionnées à l'intérieur d'un noeud. A moins que vous n'empêchiez explicitement la fusion des modalités, la procédure tente de fusionner les modalités des variables indépendantes (prédites) pour produire l'arbre décrivant le modèle le plus simple. Cette option autorise la procédure à scinder des modalités fusionnées pour améliorer la solution obtenue.

Intervalles d'échelle pour l'analyse CHAID

Figure 1-8
Boîte de dialogue Critères, onglet Intervalles



Dans l'analyse CHAID, les variables indépendantes (prédites) d'échelle sont toujours regroupées en modalités 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.
- **Personnalisée.** Chaque variable d'échelle indépendante est répartie à l'origine dans le nombre de groupes déterminé pour cette variable.

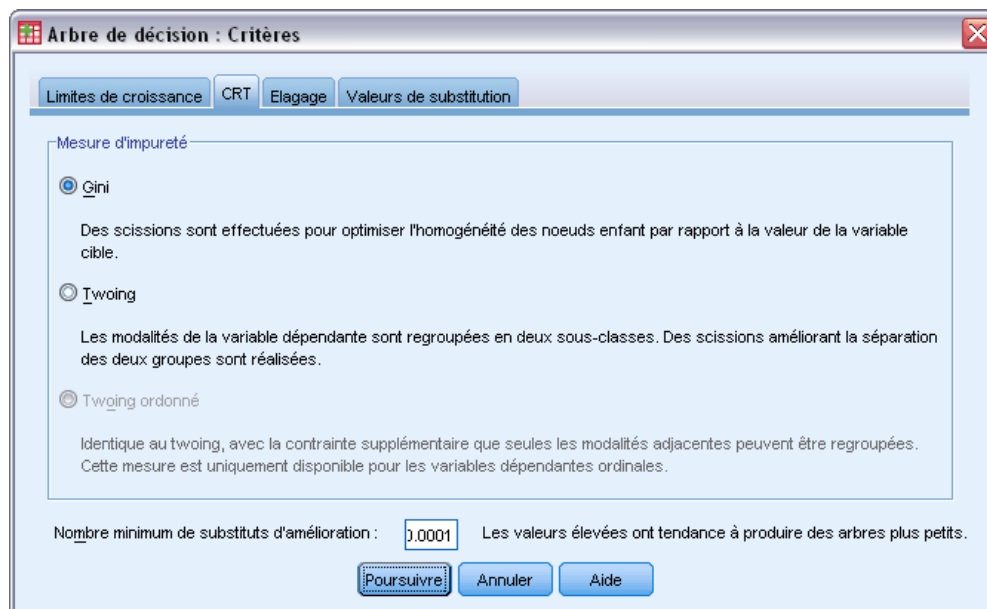
Pour déterminer les intervalles des variables d'échelle indépendantes

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez des variables d'échelle indépendantes.
- ▶ Pour la méthode de croissance, sélectionnez CHAID ou Exhaustive CHAID.
- ▶ Cliquez sur Critères.
- ▶ Cliquez sur l'onglet 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

Figure 1-9
Boîte de dialogue Critères, onglet 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 d'effectif ou les valeurs d'influence.

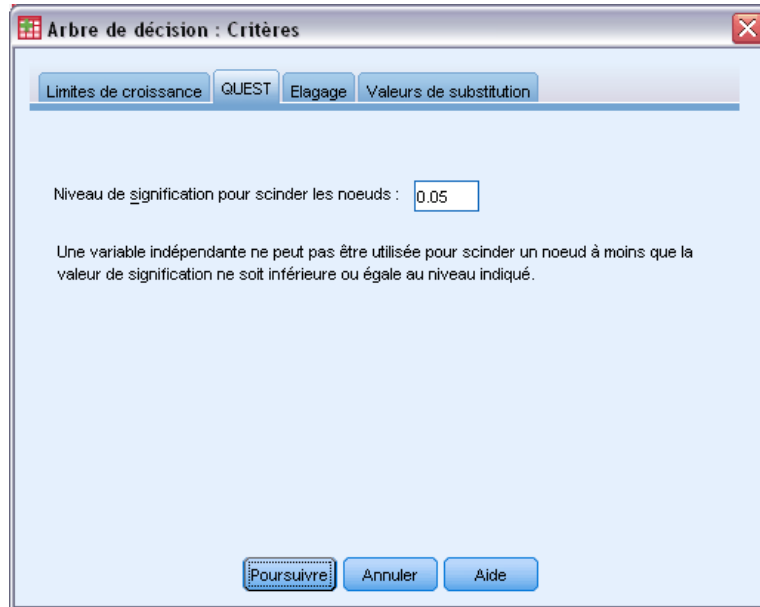
Pour les variables dépendantes (nominales, ordinales) qualitatives, vous pouvez sélectionner la mesure d'impureté parmi les suivantes :

- **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 modalité de la variable dépendante. Elle atteint son minimum (zéro) lorsque toutes les observations du noeud entrent dans une seule modalité. Il s'agit de la mesure par défaut.
- **Twoing.** Les modalités 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 modalités adjacentes peuvent être regroupées. Cette mesure est uniquement disponible pour les variables dépendantes ordinales.

Nombre minimum de substituts d'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.

Critères QUEST

Figure 1-10
Boîte de dialogue Critères, onglet QUEST



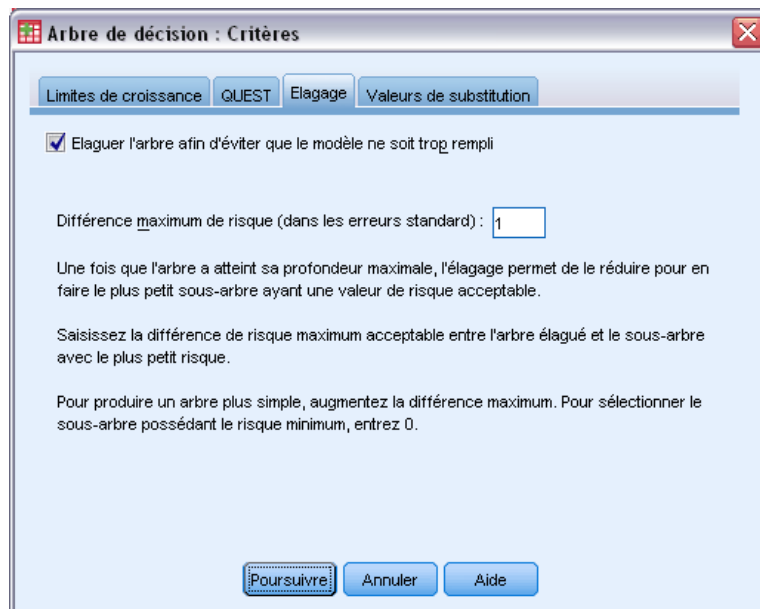
Pour la méthode QUEST, vous pouvez déterminer le niveau de signification pour scinder les noeuds. Une variable indépendante ne peut pas être utilisée pour scinder des noeuds à moins que le niveau de signification ne soit inférieur ou égal à la valeur indiquée. Cette valeur doit être supérieure à 0 et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

Pour déterminer les critères QUEST

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante nominale.
- ▶ Pour la méthode de croissance, sélectionnez QUEST.
- ▶ Cliquez sur Critères.
- ▶ Cliquez sur l'onglet QUEST.

Elagage des arbres

Figure 1-11
Boîte de dialogue Critères, onglet Elagage



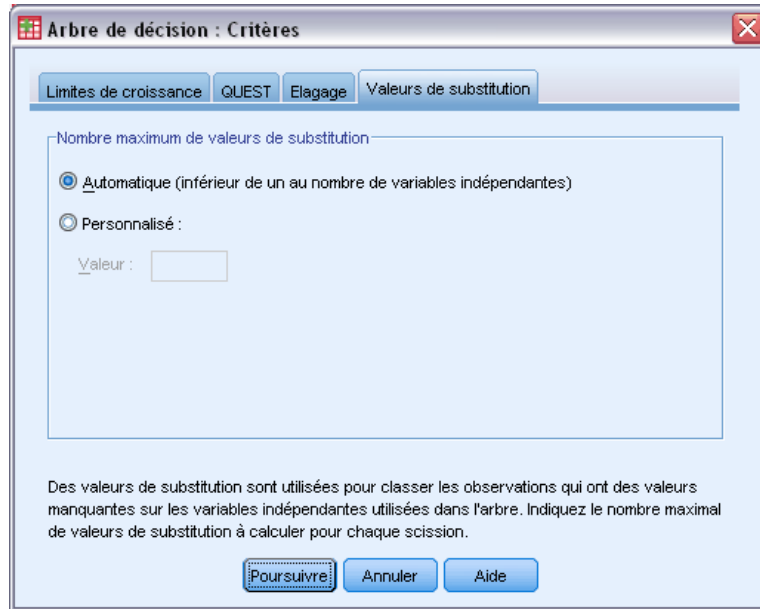
Avec les méthodes CRT et QUEST, vous pouvez faire en sorte que le modèle ne soit pas trop rempli en **élaguant** l'arbre : l'arbre croît jusqu'à atteindre les critères d'arrêt ; il est ensuite automatiquement taillé jusqu'au sous-arbre le plus petit, selon la différence maximum de risque indiquée. La 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.

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. [Pour plus d'informations, reportez-vous à la section Editeur d'arbre dans le chapitre 2 sur p. 41.](#)

Valeurs de substitution

Figure 1-12
Boîte de dialogue Critères, onglet Valeurs de substitution



Les méthodes CRT et QUEST peuvent utiliser des **valeurs de substitution** pour les variables indépendantes (prédites). 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 variables prédites de rechange sont appelées 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.

Options

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

Coûts de classification erronée

Figure 1-13
Boîte de dialogue Options, onglet Coûts de classification erronée

Arbre de décision : Options

Valeurs manquantes Coûts de classification erronée Bénéfices

Egale pour toutes les modalités

Personnalisé

Modalité estimée :

	Mauvais	Bon
Réel Mauvais	0	2
Modalité : Bon	1	0

Rendre la matrice symétrique

Dupliquer le triangle inférieur Dupliquer le triangle supérieur Utiliser les moyennes de cellules

Poursuivre Annuler Aide

Pour les variables dépendantes qualitatives (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 par le classement incorrect d'une personne présentant un risque élevé de cardiopathie dans la modalité de risque faible sera probablement beaucoup plus élevé que le coût occasionné par le classement erroné d'une personne à risque faible dans la modalité 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).

Coûts de classification erronée et étiquettes de valeur

Cette boîte de dialogue n'est disponible que si au moins deux valeurs de la variable dépendante qualitative disposent d'étiquettes de valeur définies.

Pour déterminer les coûts de classification erronée

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante qualitative (nominale, ordinale) disposant d'au moins deux étiquettes de valeur définies.
- ▶ Cliquez sur Options.

- ▶ Cliquez sur l'onglet Coûts de classification erronée.
- ▶ Cliquez sur Personnalisé.
- ▶ Saisissez des coûts de classification erronée dans la grille. 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 commandes suivantes vous aident à spécifier une matrice de coûts symétrique :

- **Copier moitié inférieure.** 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.
- **Copier moitié supérieure.** 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

Figure 1-14
Boîte de dialogue Options, onglet Bénéfices

Arbre de décision : Options

Valeurs manquantes Coûts de classification erronée Bénéfices

Aucun

Personnalisé

Valeurs de recette et de dépense :

	Recette	Dépense	Bénéfice
Mauvais	10	12	-2.0
Bon	100	5	95.0

Saisissez les valeurs de recette et de dépense pour chaque modalité. Les bénéfices sont calculés automatiquement

Poursuivre Annuler Aide

Pour les variables dépendantes qualitatives, 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 modalités de la variable dépendante affichée dans la grille.

Bénéfices et étiquettes de valeur

Cette boîte de dialogue requiert des étiquettes de valeur définies pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante qualitative disposent d'étiquettes de valeur définies.

Pour déterminer des bénéfices

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante qualitative (nominale, ordinale) disposant d'au moins deux étiquettes de valeur définies.
- ▶ Cliquez sur Options.
- ▶ Cliquez sur l'onglet Bénéfices.
- ▶ Cliquez sur Personnalisé.
- ▶ Saisissez les valeurs de recette et de dépense de toutes les modalités de variable dépendante répertoriées dans la grille.

Probabilités a priori

Figure 1-15
Boîte de dialogue Options, onglet Probabilités a priori

Arbre de décision : Options

Valeurs manquantes Coûts de classification erronée Bénéfices Probabilités a priori

Obtenue à partir d'échantillons d'apprentissage (probabilités a priori empiriques).
 Egale pour toutes les modalités
 Personnalisé

A priori :

	Valeur
Mauvais	25
Bon	75

Somme des valeurs : 100 Les valeurs sont normalisées automatiquement

Ajuster les probabilités a priori en utilisant les coûts de mauvaise réaffectation

Poursuivre Annuler Aide

Pour les arbres CRT et QUEST comportant des variables dépendantes qualitatives, 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 modalité de la variable dépendante, effectuées avant la prise de connaissance des valeurs des variables indépendantes (prédites). 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, c'est la distribution des observations dans l'échantillon d'apprentissage qui est utilisée.

Remarque : Etant donné que, pour la validation par partition, les observations sont attribuées de manière aléatoire à l'échantillon d'apprentissage, vous ne connaîtrez pas à l'avance la distribution réelle des observations à l'intérieur de l'échantillon d'apprentissage. [Pour plus d'informations, reportez-vous à la section Validation sur p. 8.](#)

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

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

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, reportez-vous à la section Coûts de classification erronée sur p. 17.](#)

Bénéfices et étiquettes de valeur

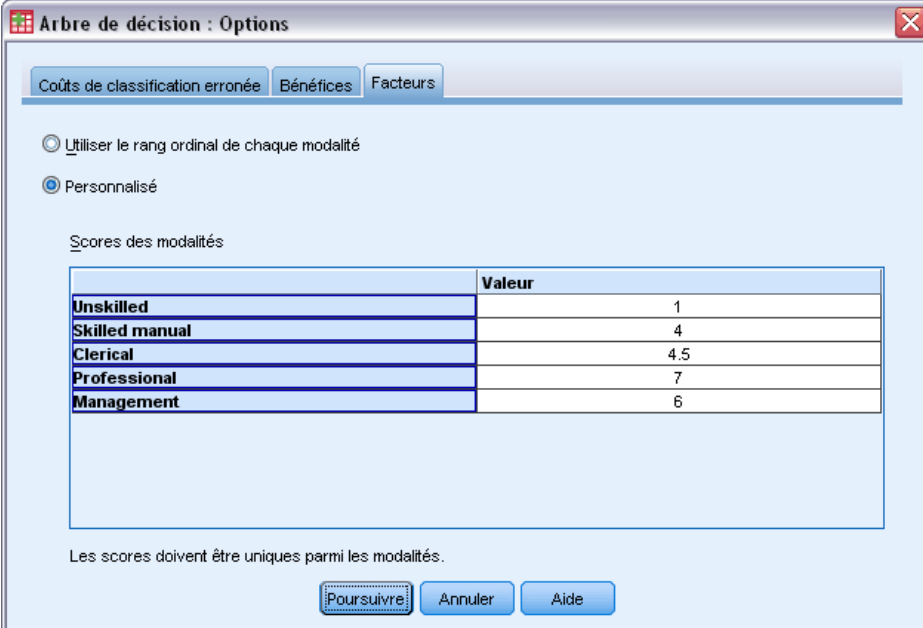
Cette boîte de dialogue requiert des étiquettes de valeur définies pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante qualitative disposent d'étiquettes de valeur définies.

Pour déterminer des probabilités a priori

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante qualitative (nominale, ordinale) disposant d'au moins deux étiquettes de valeur définies.
- ▶ Pour la méthode de croissance, sélectionnez CRT ou QUEST.
- ▶ Cliquez sur Options.
- ▶ Cliquez sur l'onglet Probabilités a priori.

Scores

Figure 1-16
Boîte de dialogue Options, onglet Scores



Arbre de décision : Options

Coûts de classification erronée Bénéfices Facteurs

Utiliser le rang ordinal de chaque modalité

Personnalisé

Scores des modalités

	Valeur
Unskilled	1
Skilled manual	4
Clerical	4,5
Professional	7
Management	6

Les scores doivent être uniques parmi les modalités.

Poursuivre Annuler Aide

Dans CHAID et Exhaustive CHAID avec une variable dépendante ordinale, vous pouvez attribuer des scores personnalisés à chaque modalité de la variable dépendante. Les scores définissent la distance entre les modalités de la variable dépendante ainsi que l'ordre de ces modalités. 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 modalité.** Le score de 1 est attribué à la modalité la plus basse de la variable dépendante, le score de 2 est attribué à la modalité supérieure suivante, etc. Il s'agit de la valeur par défaut.
- **Personnalisée.** Saisissez une valeur de score numérique pour chacune des modalités de la variable dépendante répertoriées dans la grille.

Exemple

Étiquette 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*.

Scores et étiquettes de valeur

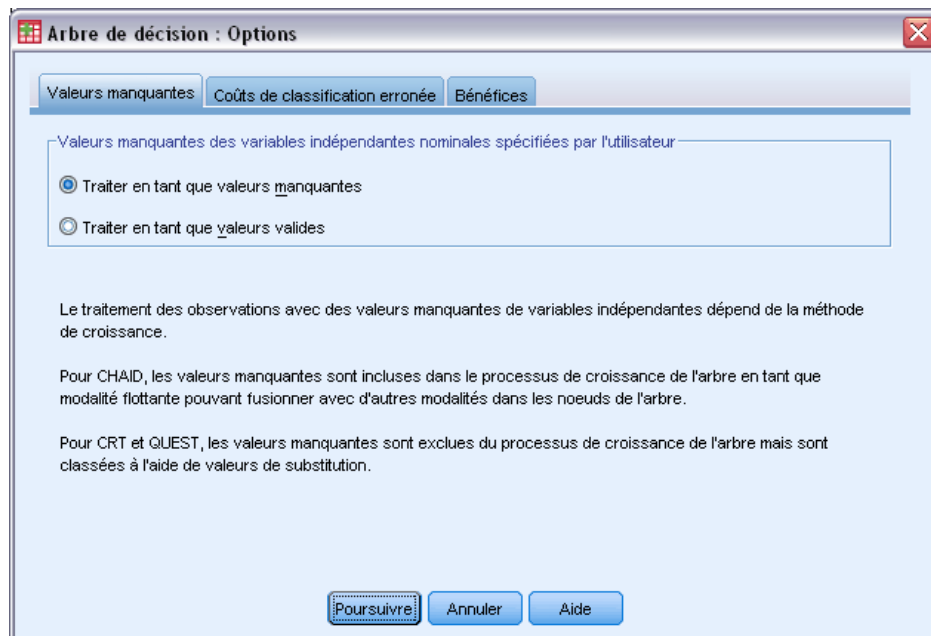
Cette boîte de dialogue requiert des étiquettes de valeur définies pour la variable dépendante. Elle n'est disponible que si au moins deux valeurs de la variable dépendante qualitative disposent d'étiquettes de valeur définies.

Pour déterminer des scores

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez une variable dépendante ordinale disposant d'au moins deux étiquettes de valeur définies.
- ▶ Pour la méthode de croissance, sélectionnez CHAID ou Exhaustive CHAID.
- ▶ Cliquez sur Options.
- ▶ Cliquez sur l'onglet Scores.

Valeurs manquantes

Figure 1-17
Boîte de dialogue Options, onglet Valeurs manquantes



L'onglet Valeurs manquantes commande la gestion des valeurs nominales, des valeurs manquantes spécifiées par l'utilisateur et des valeurs de variable indépendante (prédite).

- La gestion des valeurs de variable indépendante manquantes spécifiées par l'utilisateur, d'échelle et ordinales, varie en fonction de la méthode de croissance.
- La gestion des variables dépendantes nominales est indiquée dans la boîte de dialogue Modalités. [Pour plus d'informations, reportez-vous à la section Sélection de modalités sur p. 6.](#)
- Pour les variables d'échelle dépendantes et ordinales, les observations comportant des valeurs de variable dépendante manquantes par défaut ou spécifiées par l'utilisateur sont toujours exclues.

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

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

Règles dépendant de la méthode

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

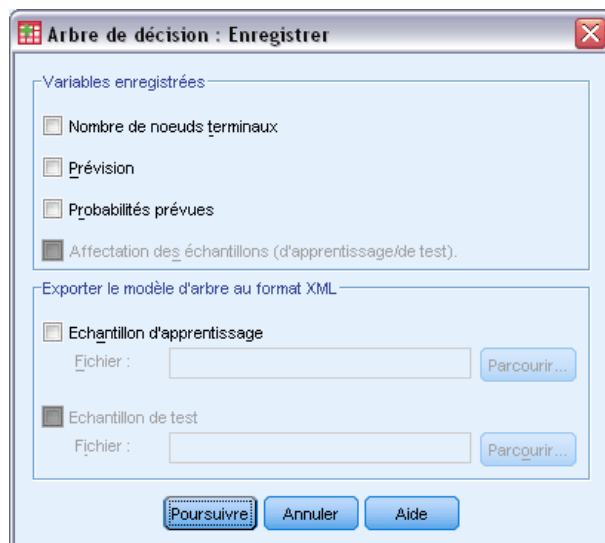
- Pour CHAID et Exhaustive CHAID, les valeurs de variable indépendante manquantes par défaut ou spécifiées par l'utilisateur sont incluses dans l'analyse en tant que modalité unique combinée. Pour les variables d'échelle indépendantes ou ordinales, les algorithmes génèrent d'abord les modalités en utilisant des valeurs valides, puis choisissent de fusionner la modalité manquante avec la modalité (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 spécifiées par l'utilisateur sont traitées comme manquantes, elles seront également gérées comme telles. [Pour plus d'informations, reportez-vous à la section Valeurs de substitution sur p. 16.](#)

Pour déterminer le traitement des valeurs manquantes indépendantes nominales spécifiées par l'utilisateur

- ▶ Dans la boîte de dialogue principale Arbre de décision, sélectionnez au moins une variable indépendante nominale.
- ▶ Cliquez sur Options.
- ▶ Cliquez sur l'onglet Valeurs manquantes.

Enregistrement des informations du modèle

Figure 1-18
Boîte de dialogue Enregistrer



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.

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

Probabilités prévues. Probabilité associée aux prévisions du modèle. Une variable est enregistrée pour chaque modalité de la variable dépendante. N'est pas disponible pour les variables d'échelle dépendantes.

Affectation des échantillons (de formation/de test). Pour la validation par partition, 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. [Pour plus d'informations, reportez-vous à la section Validation sur p. 8.](#)

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

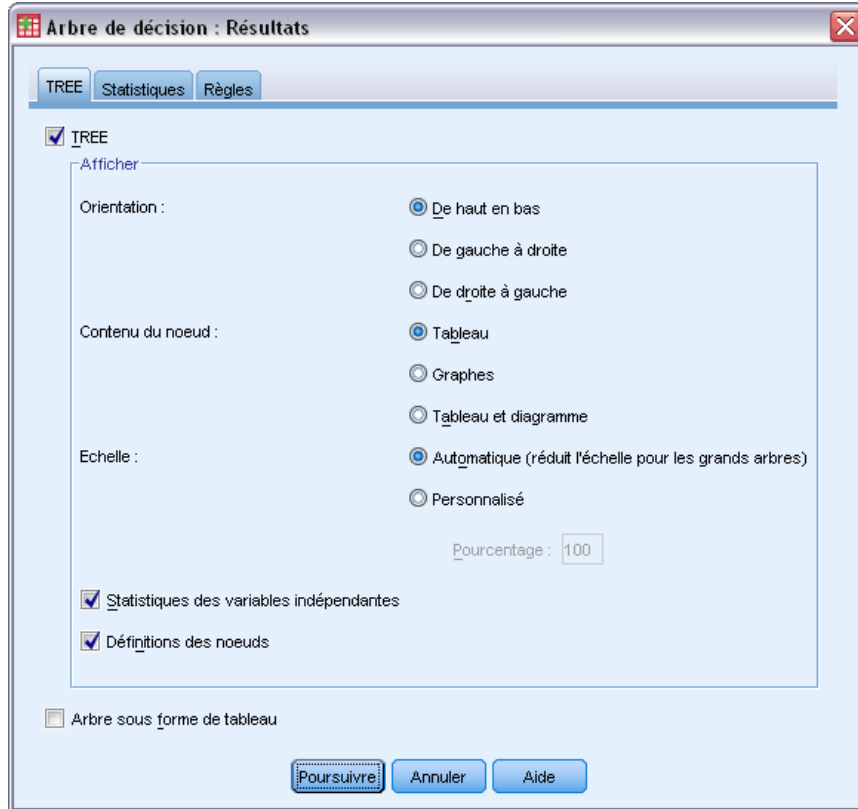
Echantillon de test. Ecrit le modèle de l'échantillon de test sur le fichier indiqué. N'est pas disponible sauf si vous avez sélectionné la validation par partition.

Résultats

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

Affichage des arbres

Figure 1-19
Boîte de dialogue Résultats, onglet Arbre



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

Arbre. Par défaut, le diagramme d'arbre est inclus dans les résultats affichés dans le Viewer. Désélectionnez cette option (supprimez la coche) pour exclure le diagramme d'arbre des résultats.

Afficher : Ces options contrôlent l'apparence initiale du diagramme d'arbre dans le Viewer. 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 des noeuds.** Les noeuds peuvent afficher des tableaux, des graphiques ou les deux. Pour les variables dépendantes qualitatives, les tableaux affichent les effectifs et les pourcentages, et les graphiques sont des diagrammes en bâtons. 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 Exhaustive CHAID, les statistiques comprennent la valeur F (pour les variables d'échelle dépendantes) ou la valeur Khi-deux (pour les variables dépendantes qualitatives) 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-deux, 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 qualitatives.

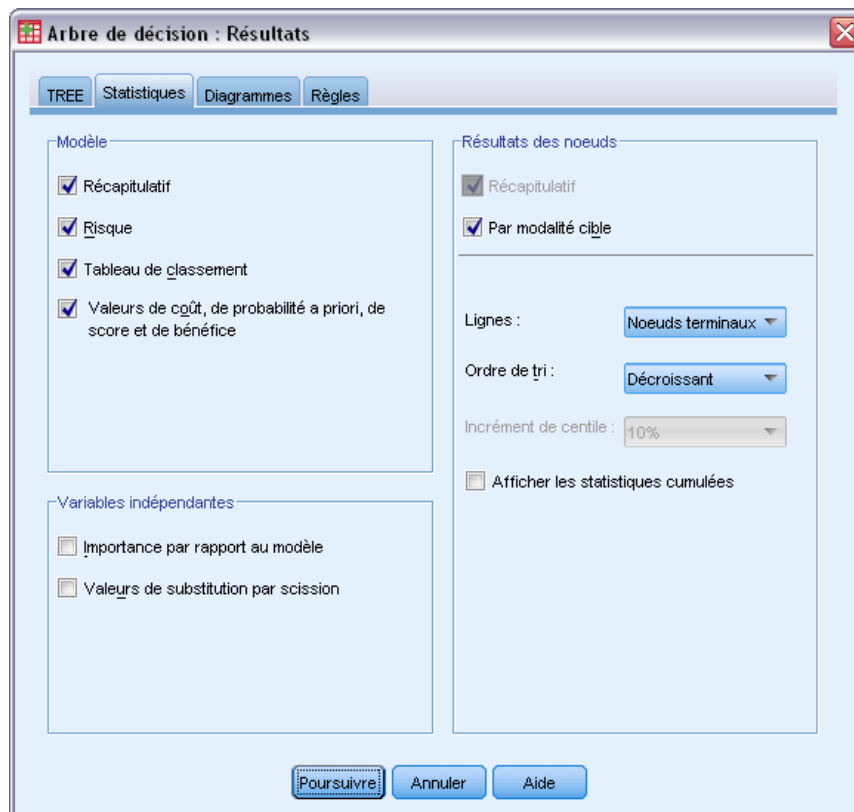
Figure 1-20

Arbre sous forme de tableau

Noeud	Mauvais		Bon		Total		Modalité estimée	Noeud parent	Principal indépendant variable				
	N	Pourcentage	N	Pourcentage	N	Pourcentage			Variable	Sig.	Chi-Square	df	Split Valeur
0	1020	41,4%	1444	58,6%	2464	100,0%	Bon						
1	454	82,1%	99	17,9%	553	22,4%	Mauvais	0	Income level	,000	662,457	2	<=Faible
2	476	42,0%	658	58,0%	1134	46,0%	Bon	0	Income level	,000	662,457	2	Faible Moyen
3	90	11,6%	687	88,4%	777	31,5%	Bon	0	Income level	,000	662,457	2	>Moyen
4	422	56,7%	322	43,3%	744	30,2%	Mauvais	2	Number of credit cards	,000	193,113	1	5 et plus
5	54	13,8%	336	86,2%	390	15,8%	Bon	2	Number of credit cards	,000	193,113	1	Moins de 5
6	80	17,6%	375	82,4%	455	18,5%	Bon	3	Number of credit cards	,000	38,587	1	5 et plus
7	10	3,1%	312	96,9%	322	13,1%	Bon	3	Number of credit cards	,000	38,587	1	Moins de 5
8	211	80,8%	50	19,2%	261	10,6%	Mauvais	4	Age	,000	95,299	1	<=28,07
9	211	43,7%	272	56,3%	483	19,6%	Bon	4	Age	,000	95,299	1	>28,07

Statistiques

Figure 1-21
Boîte de dialogue Résultat, onglet 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.

Figure 1-22
Tableau récapitulatif des modèles

Spécifications	Méthode de développement	CHAID	
	Variable dépendante :	Notation Crédit	
	Variables indépendantes	Age, Niveau de revenu, Nombre de cartes de crédit, Education, Crédit automobile	
	Validation	NONE	
	Profondeur maximum de l'arbre		3
	Nombre minimum d'observations d'un noeud parent		400
Résultats	Nombre minimum d'observations d'un noeud enfant		200
	Variables indépendantes incluses	Niveau de revenu, Nombre de cartes de crédit, Age	
	Nombre de noeuds		10
	Nombre de noeuds terminaux		6
	Profondeur		3

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

- Pour les variables dépendantes qualitatives, 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.

Tableau de classement : Pour les variables dépendantes qualitatives (nominales, ordinales), ce tableau comporte le nombre d'observations classées correctement et incorrectement pour chaque modalité de la variable dépendante. N'est pas disponible pour les variables d'échelle dépendantes.

Figure 1-23
Tableaux de risque et de classement

Risque

Echantillon	Estimation	Erreur std.
Formation	,245	,012

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

Classification

Echantillon	Observations	Prévisions		
		Mauvais	Bon	Pourcentage correct
Formation	Mauvais	366	153	70,5%
	Bon	148	561	79,1%
	Pourcentage global	41,9%	58,1%	75,5%

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

Valeurs de coût, de probabilité a priori, de score et de bénéfice. Pour les variables dépendantes qualitatives, 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 (prédite) selon son importance dans le modèle. N'est pas disponible pour les méthodes QUEST ou CHAID.

Valeurs de substitution par partition. 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, reportez-vous à la section Valeurs de substitution sur p. 16.](#)

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 qualitatives 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 qualitatives dont les bénéfices ne sont pas définis. [Pour plus d'informations, reportez-vous à la section Bénéfices sur p. 18.](#)

Figure 1-24

Tableaux récapitulatifs des gains pour les noeuds et les centiles

Récapitulatif des gains pour les noeuds

Noeud	N	Pourcentage	Bénéfice	ROI
7	322	13,1%	77.826	377,4%
5	390	15,8%	70.308	308,8%
6	455	18,5%	67.692	287,9%
9	483	19,6%	49.420	172,0%
8	261	10,6%	23.410	64,7%
1	553	22,4%	22.532	61,9%

Récapitulatif des gains pour les centiles

Percentile	Noeud	N	Bénéfice	ROI
10	7	246	77.826	377,4%
20	7 ; 5	493	75.218	352,0%
30	5 ; 6	739	73.488	336,2%
40	6	966	72.036	323,4%
50	6 ; 9	1232	70.205	307,9%
60	9	1478	66.745	280,6%
70	9 ; 8	1725	63.134	254,4%
80	8 ; 1	1971	58.149	221,6%
90	1	2218	54.183	197,9%
100	1	2464	51.023	180,4%

Par modalité cible. Pour les variables dépendantes qualitatives dont les modalités cible sont définies, le tableau comporte le pourcentage de gains, le pourcentage de réponses et le pourcentage d'index par noeud ou groupe de centiles. Un tableau distinct est produit pour chaque modalité

cible. N'est pas disponible pour les variables d'échelle dépendantes ou qualitatives dont les modalités cible ne sont pas définies. [Pour plus d'informations, reportez-vous à la section Sélection de modalités sur p. 6.](#)

Figure 1-25
Gains des modalités cible pour les noeuds et les centiles

Modalité cible : mauvaise

Gains pour les noeuds

Noeud	Noeud		Gains		Réponse	Index
	N	Pourcentage	N	Pourcentage		
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Gains pour les centiles

Centile	Noeud	N	Gains		Réponse	Index
			N	Pourcentage		
10	1	246	202	19,8%	82,1%	198,3%
20	1	493	405	39,7%	82,1%	198,3%
30	1 ; 8	739	604	59,3%	81,8%	197,6%
40	8 ; 9	986	740	72,6%	75,1%	181,3%
50	9	1232	848	83,1%	68,8%	166,2%
60	9 ; 6	1478	908	89,0%	61,4%	148,4%
70	6	1725	951	93,3%	55,1%	133,2%
80	6 ; 5	1971	986	96,7%	50,0%	120,9%
90	5 ; 7	2218	1012	99,3%	45,6%	110,3%
100	7	2464	1020	100,0%	41,4%	100,0%

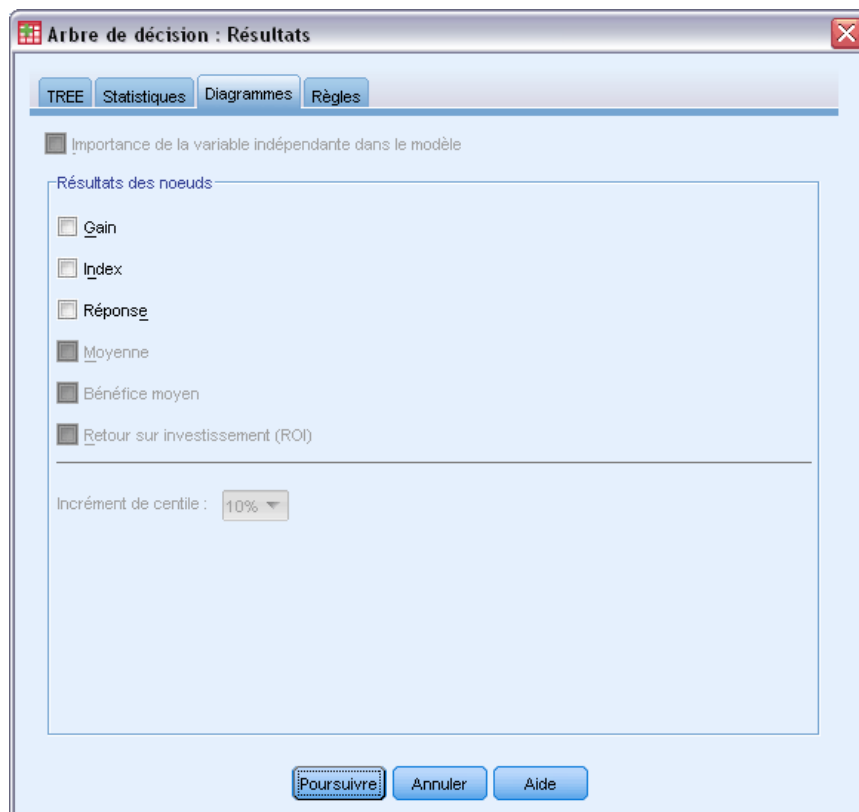
Lignes. Les tableaux de résultats des noeuds peuvent afficher les résultats par noeuds terminaux, par centiles ou les deux. Si vous sélectionnez les deux, vous obtenez deux tableaux, un pour chaque modalité cible. Les tableaux utilisant des centiles comportent des valeurs cumulatives pour chaque centile,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

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

Afficher les statistiques cumulées. Pour les tableaux utilisant des noeuds terminaux, ajoutez une colonne comportant les résultats cumulés.

Diagrammes

Figure 1-26
Boîte de dialogue *Résultat*, onglet *Diagrammes*



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

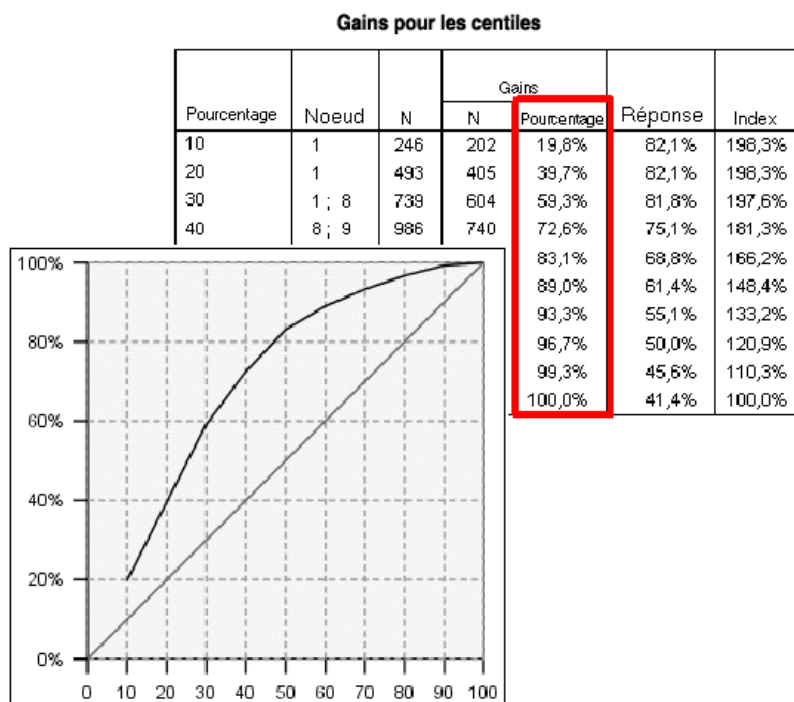
Importance de la variable indépendante dans le modèle. Diagramme en bâtons représentant l'importance dans le modèle de chaque variable indépendante (prédite). Valable uniquement pour la méthode de croissance CRT.

Résultats des noeuds

Gain. Le gain est le pourcentage d'observations totales de la modalité cible dans chaque noeud, calculé de la manière suivante : $(\text{cibles des noeuds } n / \text{nombre total de cibles } n) \times 100$. Le diagramme des gains est un diagramme curviligne représentant les gains cumulés en centiles, calculé de la manière suivante : $(\text{cibles des centiles cumulés } n / \text{nombre total de cibles } n) \times 100$. Un diagramme curviligne distinct est créé pour chaque modalité cible. Est uniquement disponible pour les variables dépendantes qualitatives dont les modalités cible sont définies. [Pour plus d'informations, reportez-vous à la section Sélection de modalités sur p. 6.](#)

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

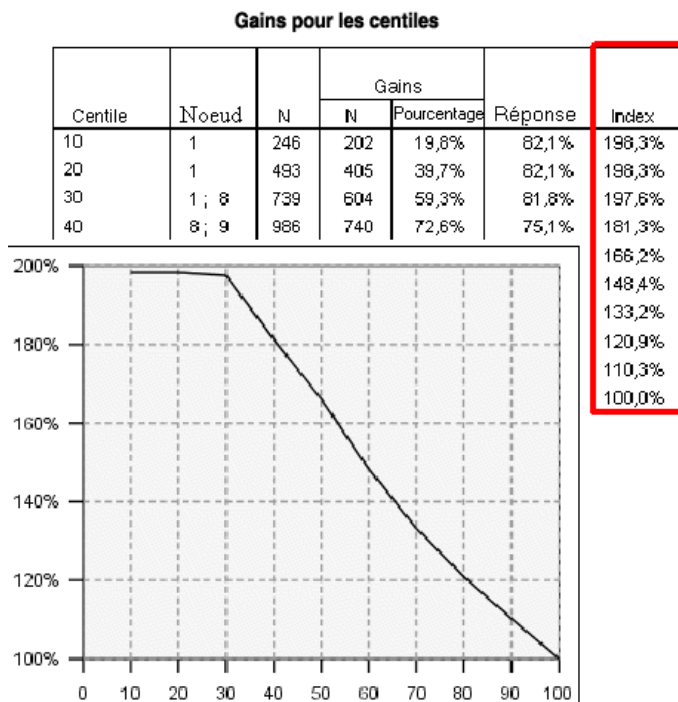
Figure 1-27
Tableau Gains pour les centiles et diagramme des gains



Index. L'index correspond au ratio du pourcentage de réponses du nœud pour la catégorie cible comparé au pourcentage de réponses global pour la catégorie cible de l'ensemble de l'échantillon. Le diagramme des index est un diagramme curviligne représentant les valeurs de l'index des centiles cumulés. Est uniquement disponible pour les variables dépendantes qualitatives. L'index des centiles cumulés est calculé de la manière suivante : (pourcentage de réponse des centiles cumulés/pourcentage total de réponses) x 100. Un diagramme distinct est créé pour chaque modalité cible, et les modalités cible doivent être définies.

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

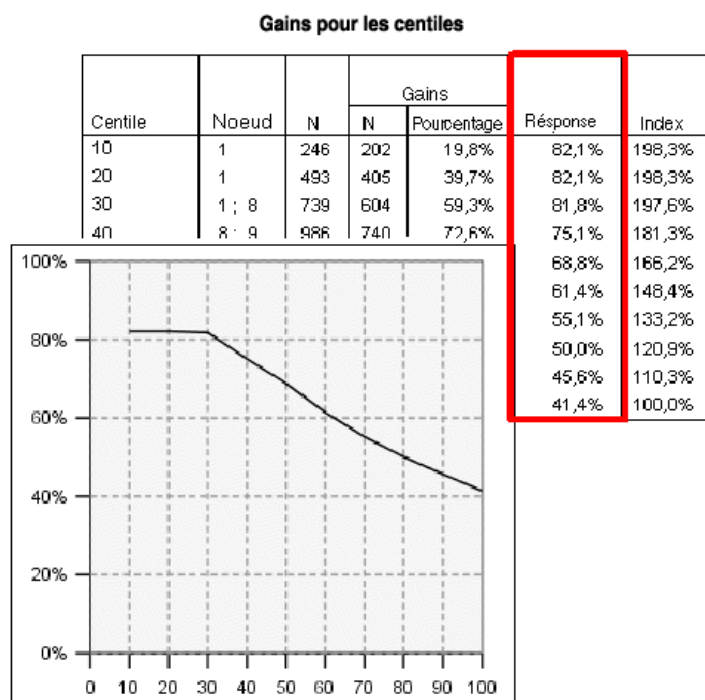
Figure 1-28
Tableau Gains pour les centiles et diagramme d'index



Réponse. Le pourcentage d'observations dans le noeud dans la modalité cible spécifiée; Le diagramme de réponse est un diagramme curviligne représentant les réponses des centiles cumulés, calculé de la manière suivante : (cibles des centiles cumulés n /nombre total de centiles cumulés n) x 100. Est uniquement disponible pour les variables dépendantes qualitatives dont les modalités cible sont définies.

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

Figure 1-29
Tableau Gains pour les centiles et diagramme de réponse



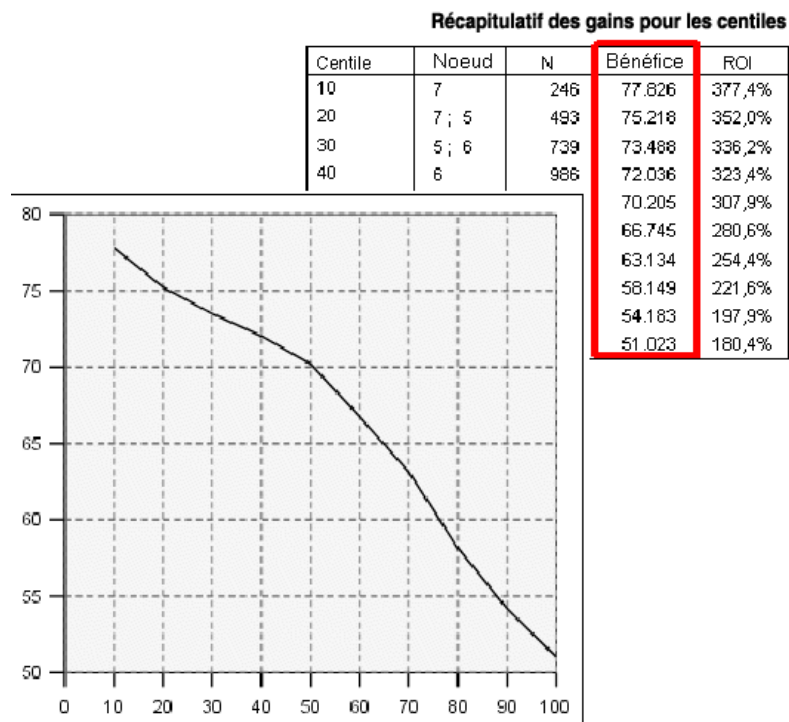
Moyenne. Diagramme curviligne représentant les valeurs moyennes des centiles cumulés pour la variable dépendante. Est uniquement disponible pour les variables d'échelle dépendantes.

Bénéfice moyen. Diagramme curviligne représentant les profits moyens cumulés. Disponible uniquement pour les variables dépendantes qualitatives dont les bénéfices sont définis. [Pour plus d'informations, reportez-vous à la section Bénéfices sur p. 18.](#)

Le diagramme des profits moyens trace point par point les valeurs de la colonne *Bénéfices* du tableau Récapitulatif des gains pour les centiles.

Figure 1-30

Tableau récapitulatif des gains pour les centiles et profit moyen

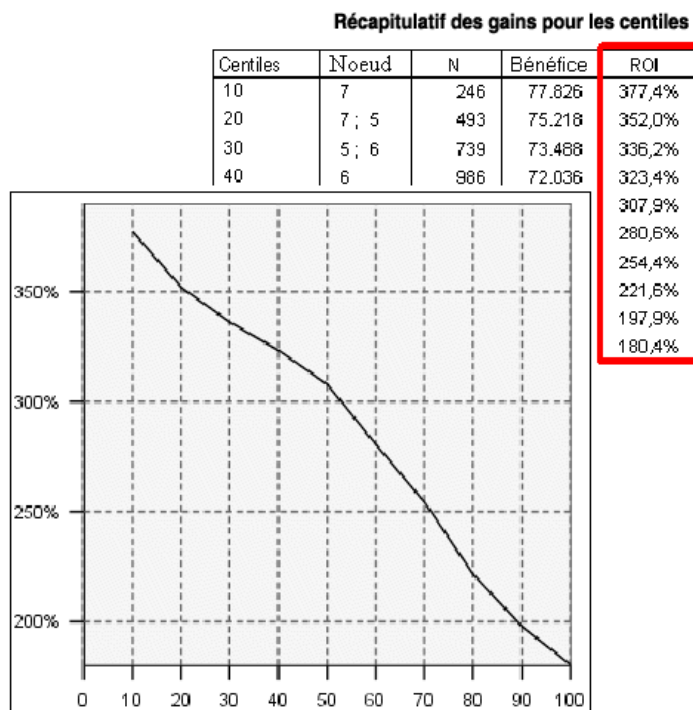


Retour sur investissement (ROI). Diagramme curviligne du ROI (retour sur investissement) cumulé. ROI est le ratio recettes/dépenses. Disponible uniquement pour les variables dépendantes qualitatives dont les bénéfices sont définis.

Le diagramme du ROI trace point par point les valeurs de la colonne *ROI* du tableau Récapitulatif des gains pour les centiles.

Figure 1-31

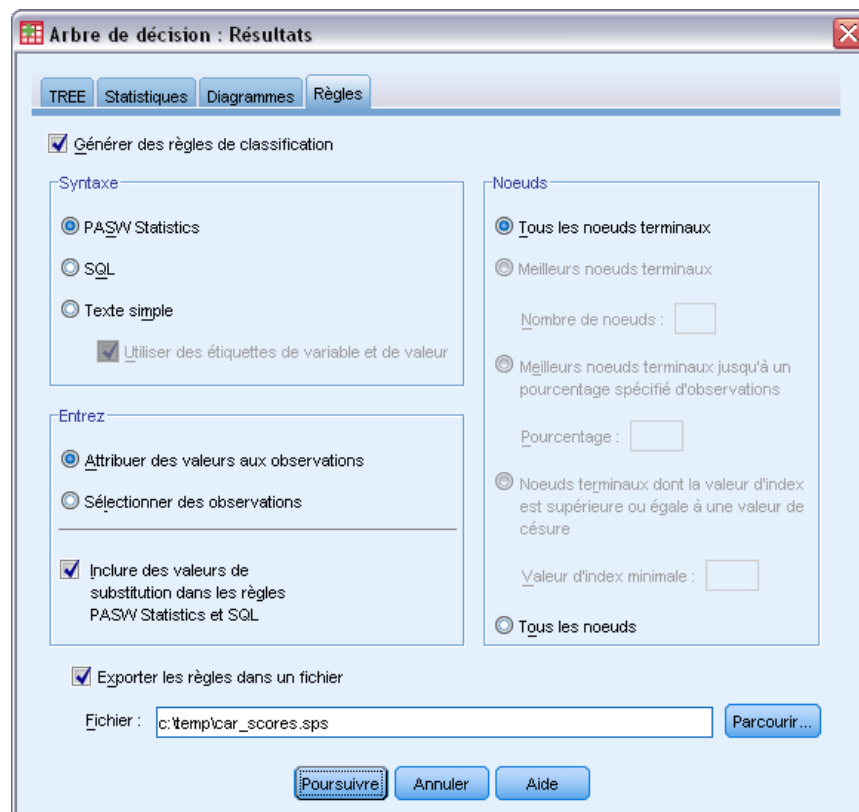
Tableau récapitulatif des gains pour les centiles et diagramme du ROI



Incrément de centile. Pour tous les diagrammes utilisant des centiles, ce paramètre contrôle l'affichage des incréments des centiles sur le diagramme : 1, 2, 5, 10, 20 ou 25.

Règles de sélection et d'analyse

Figure 1-32
Boîte de dialogue Résultat, onglet Règles



L'onglet 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 le Viewer et/ou les enregistrer dans un fichier externe.

Syntaxe. Contrôle la forme des règles de sélection des résultats affichés dans le Viewer et des règles de sélection enregistrées dans un fichier externe.

- **Langage de syntaxe de commande IBM® SPSS® Statistics.** 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 étiquettes de valeur ou de variable définies, ou des noms de variables et des valeurs de données.

Type. Pour SPSS Statistics et les règles SQL, commande le type de règles affiché : règles de sélection ou d'analyse.

- **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 SPSS Statistics et SQL, une règle unique est créée pour sélectionner toutes les observations respectant les critères de sélection.

Inclure des valeurs de substitution dans SPSS Statistics et les règles SQL. Pour CRT et QUEST, vous pouvez inclure des variables prédites 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 (prédite) incomplètes et que vous souhaitez que les règles reproduisent votre arbre, incluez les valeurs de substitution. [Pour plus d'informations, reportez-vous à la section Valeurs de substitution sur p. 16.](#)

Noeuds. Commande 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. (Voir la remarque ci-après.)
- **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. (Voir la remarque ci-après.)
- **Noeuds terminaux dont la valeur d'index est égale ou supérieure 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 modalité cible de ce noeud dépasse le pourcentage du noeud racine. (Voir la remarque ci-après.)
- **Tous les noeuds.** Génère des règles pour tous les noeuds.

Remarque 1 : La sélection des noeuds basée sur les valeurs d'index est uniquement disponible pour les variables dépendantes qualitatives comportant des modalités cible définies. Si vous avez indiqué plusieurs modalités cible, un jeu de règles distinct est créé pour chaque modalité cible.

Remarque 2 : Pour SPSS Statistics et les règles SQL de sélection des observations (et non les règles d'affectation des valeurs), Tous les noeuds et Tous les noeuds terminaux génèrent efficacement 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'analyse de manière interactive, en fonction des noeuds sélectionnés dans le modèle d'arbre final. [Pour plus d'informations, reportez-vous à la section Règles de sélection et d'analyse des observations dans le chapitre 2 sur p. 49.](#)

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).

Editeur d'arbre

Avec l'éditeur d'arbre, vous pouvez :

- Masquer et afficher des branches d'arbre sélectionnées.
- Contrôler l'affichage du contenu des noeuds, des statistiques à l'endroit de la scission des noeuds, ainsi que d'autres informations.
- Modifier les noeuds, les arrière-plans, les bordures, les diagrammes et les couleurs de police.
- Modifier le style et la taille de police.
- Modifier l'alignement des arbres.
- Sélectionner des sous-ensembles d'observations pour une analyse plus approfondie basée sur les noeuds sélectionnés.
- Créer et enregistrer des règles de sélection ou d'analyse des observations basées sur les noeuds sélectionnés.

Pour modifier un modèle d'arbre :

- ▶ Double-cliquez sur le modèle d'arbre dans la fenêtre du Viewer.

ou

- ▶ Dans le menu Edition ou le menu contextuel, choisissez :
Modifier le contenu > Dans une fenêtre distincte

Affichage/Masquage des noeuds

Pour masquer (réduire) tous les noeuds enfant dans une branche située en dessous d'un noeud parent :

- ▶ Cliquez sur le signe moins (-) dans la petite case située sous le coin inférieur droit du noeud parent.

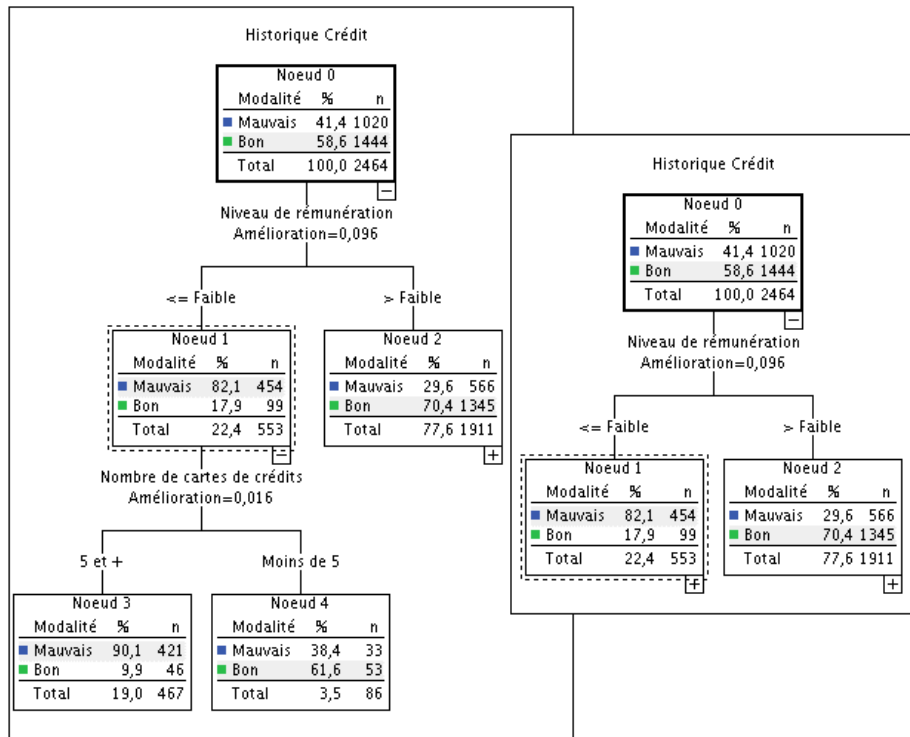
Tous les noeuds en dessous du noeud parent de cette branche seront masqués.

Pour afficher (développer) tous les noeuds enfant dans une branche située en dessous d'un noeud parent :

- ▶ Cliquez sur le signe plus (+) dans la petite case située sous le coin inférieur droit du noeud parent.

Remarque : Masquer les noeuds enfant d'une branche ne revient pas à élaguer un arbre. Si vous souhaitez élaguer votre arbre, vous devez demander un élagage avant de créer l'arbre ; ainsi, les branches élaguées ne sont pas incluses dans l'arbre final. [Pour plus d'informations, reportez-vous à la section Elagage des arbres dans le chapitre 1 sur p. 15.](#)

Figure 2-1
Arbre développé et réduit



Sélection de plusieurs noeuds

Vous pouvez sélectionner des observations, générer des règles d'analyse et de sélection, et réaliser d'autres actions basées sur les noeuds sélectionnés. Pour sélectionner plusieurs noeuds :

- ▶ Cliquez sur le noeud que vous voulez sélectionner.
- ▶ Cliquez sur les autres noeuds que vous voulez sélectionner en maintenant la touche Ctrl enfoncée.

Vous pouvez sélectionner des noeuds enfant et/ou des noeuds parent dans une branche et des noeuds enfant dans une autre branche. Cependant, il est impossible d'utiliser la sélection multiple sur un noeud parent et un noeud enfant de la même branche.

Manipulation de grands arbres

Il peut arriver que les modèles d'arbre contiennent tellement de noeuds et de branches qu'il est difficile, voire impossible d'afficher l'intégralité de l'arbre en taille normale. Les fonctions suivantes peuvent vous être utiles lorsque vous manipulez de grands arbres :

- **Carte d'arbre.** Vous pouvez utiliser la carte d'arbre, une version beaucoup plus petite et simplifiée de l'arbre, pour vous déplacer dans l'arbre et sélectionner des noeuds. [Pour plus d'informations, reportez-vous à la section Carte d'arbre sur p. 43.](#)

- **Echelle.** Vous pouvez effectuer des zooms arrière et avant en modifiant le pourcentage d'échelle utilisé pour l'affichage de l'arbre. [Pour plus d'informations, reportez-vous à la section Mise à l'échelle de l'affichage de l'arbre sur p. 44.](#)
- **Affichage des noeuds et des branches.** Vous pouvez rendre l'arbre plus compact en affichant uniquement les tableaux ou uniquement les diagrammes dans les noeuds, et/ou en supprimant l'affichage des étiquettes de noeud ou des informations sur les variables indépendantes. [Pour plus d'informations, reportez-vous à la section Contrôle des informations affichées dans l'arbre sur p. 45.](#)

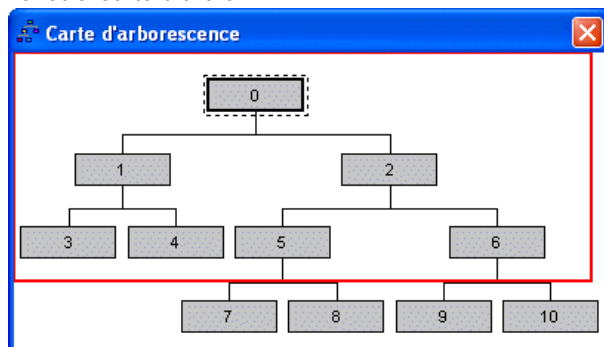
Carte d'arbre

La carte d'arbre fournit une vue compacte et simplifiée de l'arbre pouvant être utilisée pour se déplacer dans l'arbre et sélectionner des noeuds.

Pour utiliser la fenêtre de la carte d'arbre :

- ▶ A partir des menus de l'éditeur d'arbre, sélectionnez :
Affichage > Carte d'arbre

Figure 2-2
Fenêtre Carte d'arbre



- Le noeud sélectionné est mis en évidence dans l'éditeur de modèle d'arbre et dans la fenêtre de la carte d'arbre.
- La portion de l'arbre figurant actuellement dans la zone d'affichage de l'éditeur de modèle d'arbre est indiquée par un rectangle rouge dans la carte d'arbre. Cliquez avec le bouton droit et faites glisser le rectangle pour modifier la section de l'arbre affichée dans la zone d'affichage.
- Si vous sélectionnez un noeud de la carte d'arbre ne figurant pas dans la zone d'affichage de l'éditeur, l'affichage change pour inclure le noeud sélectionné.
- La sélection de plusieurs noeuds fonctionne de la même manière dans la carte d'arbre que dans l'éditeur d'arbre : Tout en maintenant la touche Ctrl enfoncée, cliquez sur les noeuds pour les sélectionner. Il est impossible d'utiliser la sélection multiple sur un noeud parent et un noeud enfant de la même branche.

Mise à l'échelle de l'affichage de l'arbre

Par défaut, l'échelle des arbres est automatiquement ajustée à la fenêtre du Viewer, ce qui risque de rendre très difficile la lecture de certains arbres. Vous pouvez sélectionner un paramètre d'échelle prédéfini ou saisir votre propre valeur personnalisée située entre 5 et 200 %.

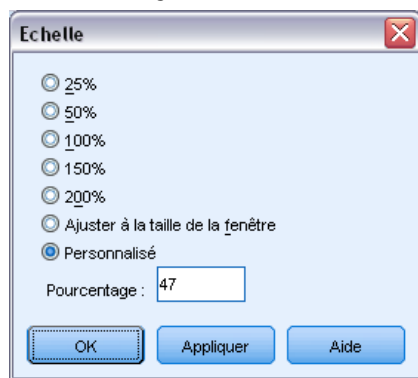
Pour modifier l'échelle de l'arbre :

- Sélectionnez un pourcentage d'échelle dans la liste déroulante de la barre d'outils ou saisissez un pourcentage personnalisé.

ou

- A partir des menus de l'éditeur d'arbre, sélectionnez :
Affichage > Echelle...

Figure 2-3
Boîte de dialogue Echelle



Vous pouvez également indiquer une valeur d'échelle avant de créer le modèle d'arbre. [Pour plus d'informations, reportez-vous à la section Résultats dans le chapitre 1 sur p. 25.](#)

Fenêtre Récapitulatif des noeuds

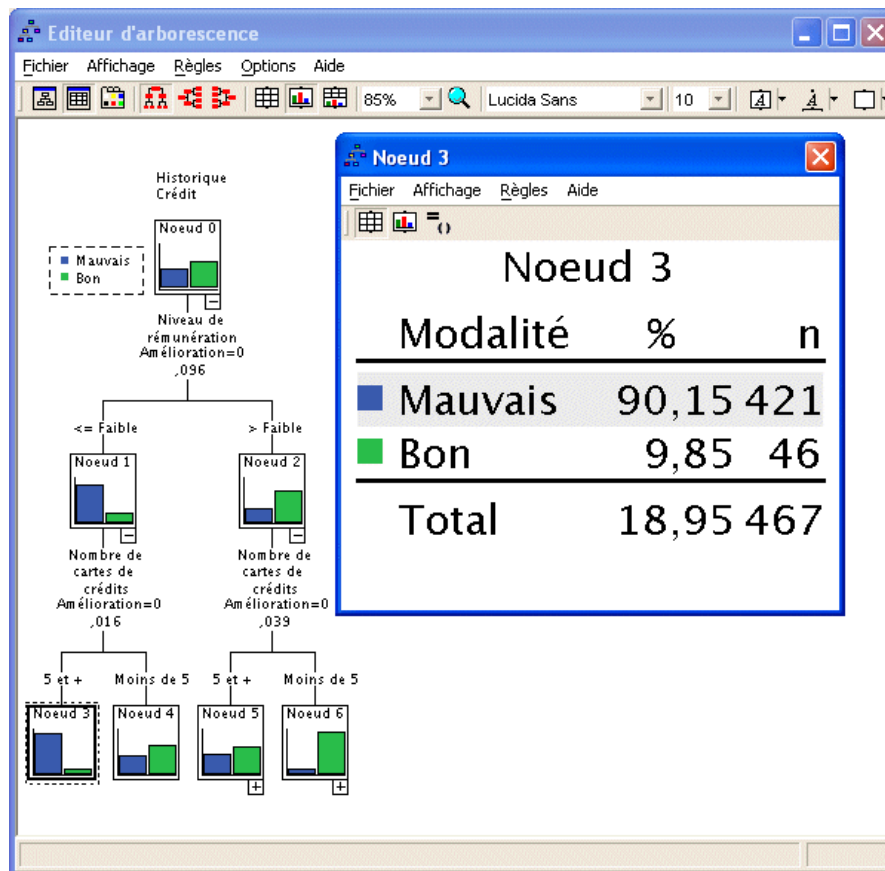
La fenêtre Récapitulatif des noeuds offre une plus grande vue des noeuds sélectionnés. Vous pouvez également utiliser la fenêtre récapitulative pour afficher, appliquer ou enregistrer des règles de sélection ou d'analyse basées sur les noeuds sélectionnés.

- Utilisez le menu Affichage de la fenêtre Récapitulatif des noeuds pour changer l'affichage d'un tableau récapitulatif, d'un diagramme ou de règles.
- Utilisez le menu Règles de la fenêtre Récapitulatif des noeuds pour sélectionner le type de règles que vous voulez afficher. [Pour plus d'informations, reportez-vous à la section Règles de sélection et d'analyse des observations sur p. 49.](#)
- Tous les affichages de la fenêtre Récapitulatif des noeuds reflètent un récapitulatif combiné de tous les noeuds sélectionnés.

Pour utiliser la fenêtre Récapitulatif des noeuds :

- ▶ Sélectionnez les noeuds dans l'éditeur d'arbre. Tout en maintenant la touche Ctrl enfoncée, cliquez sur les noeuds pour les sélectionner.
- ▶ A partir des menus, sélectionnez :
Affichage > Récapitulatif

Figure 2-4
Fenêtre récapitulative



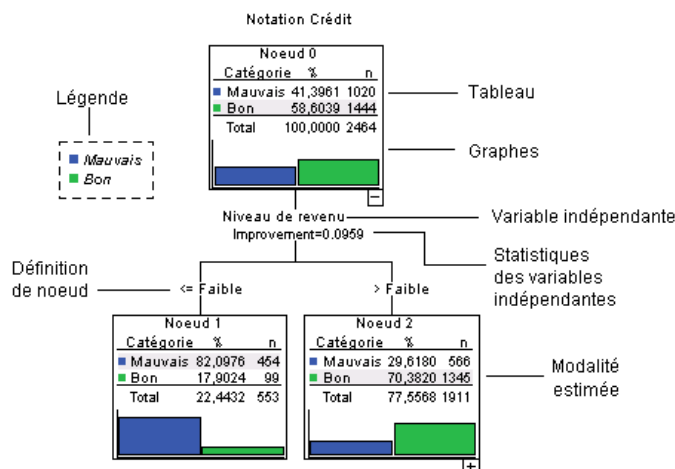
Contrôle des informations affichées dans l'arbre

Le menu Options de l'éditeur d'arbre permet de contrôler l'affichage du contenu des noeuds, des noms et des statistiques des variables indépendantes (explicatives), des définitions de noeud, etc. La majeure partie de ces paramètres peut également être contrôlée depuis la barre d'outils.

Paramètre	Sélection du menu Options
Sélectionner la modalité prévue (variable dépendante qualitative)	Sélectionner les prévisions
Tableaux et/ou diagrammes dans les noeuds	Contenu du noeud
Valeurs des tests de signification et valeurs p	Statistiques des variables indépendantes

Paramètre	Sélection du menu Options
Nom des variables indépendantes (explicatives)	Variables indépendantes
Valeurs indépendantes (explicatives) des noeuds	Définitions de noeud
Alignement (de haut en bas, de gauche à droite, de droite à gauche)	Orientation
Légende de diagramme	Légende

Figure 2-5
Éléments d'arbre



Modification des couleurs et des polices de caractères du texte des arbres

Vous pouvez modifier les couleurs de l'arbre suivantes :

- Couleur de la bordure de noeud, de l'arrière-plan et du texte
- Couleur des branches et du texte des branches
- Couleur de l'arrière-plan de l'arbre
- Couleur de mise en évidence des modalités prévues (variables dépendantes qualitatives)
- Couleurs des diagrammes de noeud

Vous pouvez également modifier le type, le style et la taille de la police pour l'intégralité des textes de l'arbre.

Remarque : Il est impossible de modifier la couleur ou les attributs de police de noeuds ou de branches individuellement. Les modifications apportées à la couleur s'appliquent à tous les éléments d'un même type et les modifications de police (à l'exception de la couleur) s'appliquent à tous les éléments du diagramme.

Pour modifier les couleurs et les attributs de police de caractère :

- ▶ Utilisez la barre d'outils pour modifier les attributs de police pour l'intégralité de l'arbre ou les couleurs des divers éléments d'arbre. (Les info-bulles décrivent chaque commande de la barre d'outils lorsque vous placez le pointeur de la souris sur la commande.)

ou

- ▶ Double-cliquez n'importe où dans l'éditeur d'arbre pour ouvrir la fenêtre Propriétés ou choisissez dans les menus :
Affichage > Propriétés
- ▶ Pour les bordures, les branches, l'arrière-plan des noeuds, les modalités prévues et l'arrière-plan de l'arbre, cliquez sur l'onglet Couleur.
- ▶ Pour les couleurs et les attributs de police, cliquez sur l'onglet Texte.
- ▶ Pour les couleurs des diagrammes de noeud, cliquez sur l'onglet Graphiques de noeud.

Figure 2-6

Fenêtre Propriétés, onglet Couleur

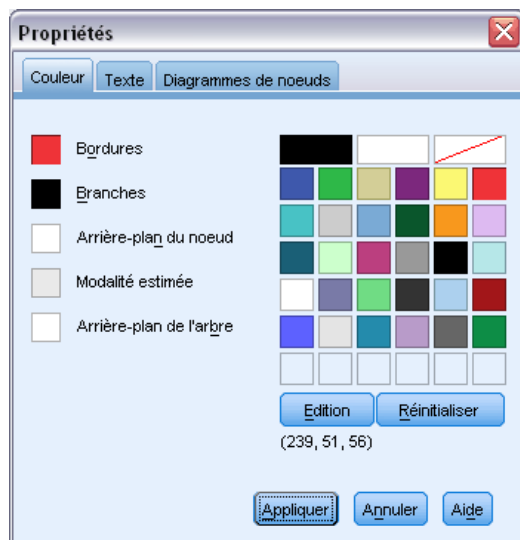


Figure 2-7
Fenêtre Propriétés, onglet Texte

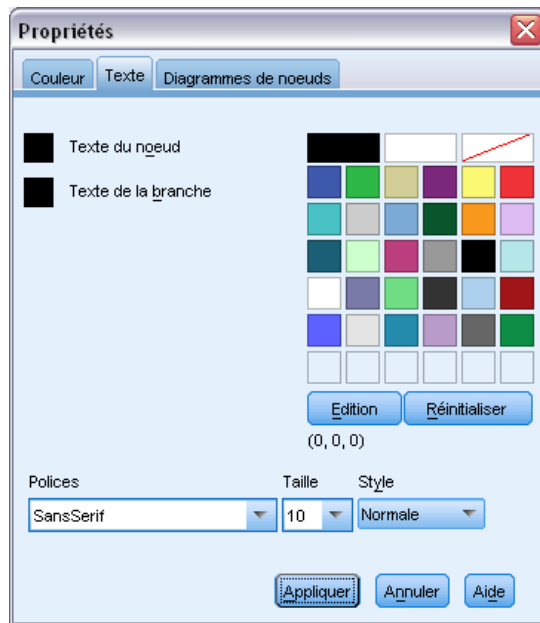
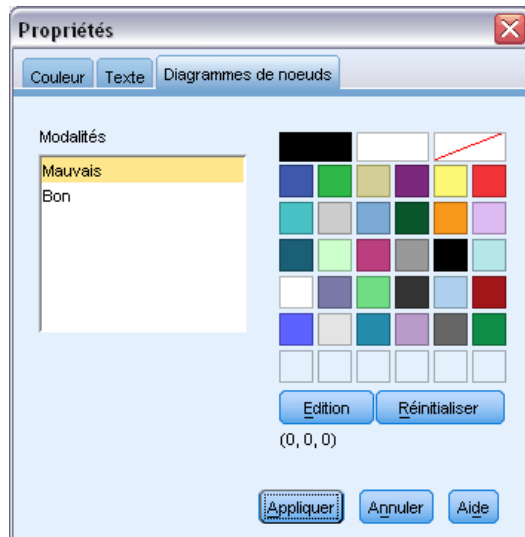


Figure 2-8
Fenêtre Propriétés, onglet Graphiques de noeud



Règles de sélection et d'analyse des observations

Vous pouvez utiliser l'éditeur d'arbre pour :

- Sélectionner des sous-ensembles d'observations basés sur les noeuds sélectionnés. [Pour plus d'informations, reportez-vous à la section Filtrage des observations sur p. 49.](#)
- Générer des règles de sélection des observations ou des règles d'analyse au format syntaxe de commande IBM® SPSS® Statistics ou au format SQL. [Pour plus d'informations, reportez-vous à la section Enregistrement des règles de sélection et d'analyse sur p. 49.](#)

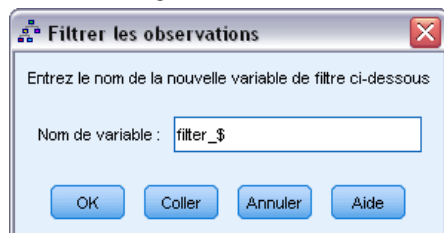
Vous pouvez également enregistrer automatiquement des règles basées sur plusieurs critères lors du lancement de la procédure Arbre de décision de création du modèle d'arbre. [Pour plus d'informations, reportez-vous à la section Règles de sélection et d'analyse dans le chapitre 1 sur p. 38.](#)

Filtrage des observations

Si vous souhaitez en savoir plus sur les observations d'un noeud ou d'un groupe de noeuds précis, vous pouvez sélectionner un sous-ensemble d'observations pour qu'il soit analysé de manière plus approfondie sur la base des noeuds sélectionnés.

- ▶ Sélectionnez les noeuds dans l'éditeur d'arbre. Tout en maintenant la touche Ctrl enfoncée, cliquez sur les noeuds pour les sélectionner.
- ▶ A partir des menus, sélectionnez :
Règles > Filtrer les observations...
- ▶ Entrez le nom d'une variable de filtre. Les observations des noeuds sélectionnés recevront la valeur 1 pour cette variable. Toutes les autres observations recevront la valeur 0 et seront exclues de l'analyse suivante jusqu'à modification de l'état du filtre.
- ▶ Cliquez sur OK.

Figure 2-9
Boîte de dialogue Filtrer les observations



Enregistrement des règles de sélection et d'analyse

Vous pouvez enregistrer les règles d'analyse et de sélection des observations dans un fichier externe, puis les appliquer à une autre source de données. Les règles sont basées sur les noeuds sélectionnés dans l'éditeur d'arbre.

Syntaxe. Contrôle la forme des règles de sélection des résultats affichés dans le Viewer et des règles de sélection enregistrées dans un fichier externe.

- **IBM® 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/extraire des enregistrements dans la base de données, 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.

Type. Vous pouvez créer des règles d'analyse ou de sélection.

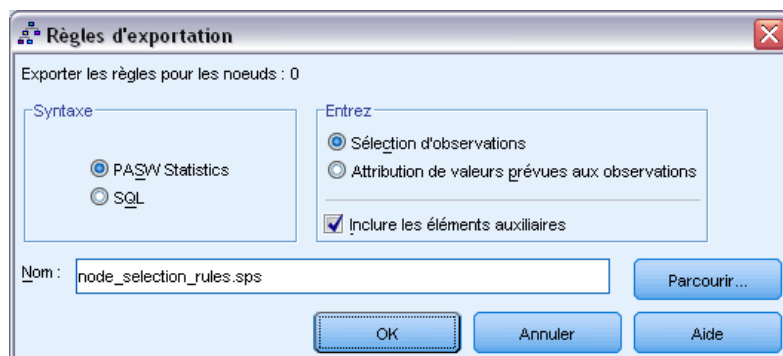
- **Sélectionner des observations.** Les règles peuvent être utilisées pour sélectionner les observations respectant les critères d'appartenance aux noeuds. Pour les règles SPSS Statistics et SQL, une règle unique est créée pour sélectionner toutes les observations respectant les critères de sélection.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

Inclure les éléments auxiliaires. Pour CRT et QUEST, vous pouvez inclure des variables indépendantes 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 (explicative) incomplètes et que vous souhaitez que les règles reproduisent votre arbre, incluez les valeurs de substitution. [Pour plus d'informations, reportez-vous à la section Valeurs de substitution dans le chapitre 1 sur p. 16.](#)

Pour enregistrer des règles d'analyse ou de sélection des observations :

- ▶ Sélectionnez les noeuds dans l'éditeur d'arbre. Tout en maintenant la touche Ctrl enfoncée, cliquez sur les noeuds pour les sélectionner.
- ▶ A partir des menus, sélectionnez :
Règles > Exporter...
- ▶ Sélectionnez le type de règles voulu et entrez un nom de fichier.

Figure 2-10
Boîte de dialogue Exporter les règles



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).

Partie II: Exemples

Hypothèses et exigences concernant les données

La procédure Arbre de décision suppose que :

- Le niveau de mesure approprié a été attribué à toutes les variables d'analyse.
- Pour les valeurs dépendantes qualitatives (**nominales** et **ordinales**), les étiquettes de valeur ont été définies pour toutes les modalités devant être incluses dans l'analyse.

Nous utiliserons le fichier *tree_textdata.sav* pour illustrer l'importance de ces deux exigences. Ce fichier de données reflète l'état par défaut des données lues ou entrées avant que des attributs, tels que le niveau de mesure ou les étiquettes de valeur, aient été définis. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans IBM SPSS Decision Trees 20.](#)

Effets du niveau de mesure sur les modèles d'arbre

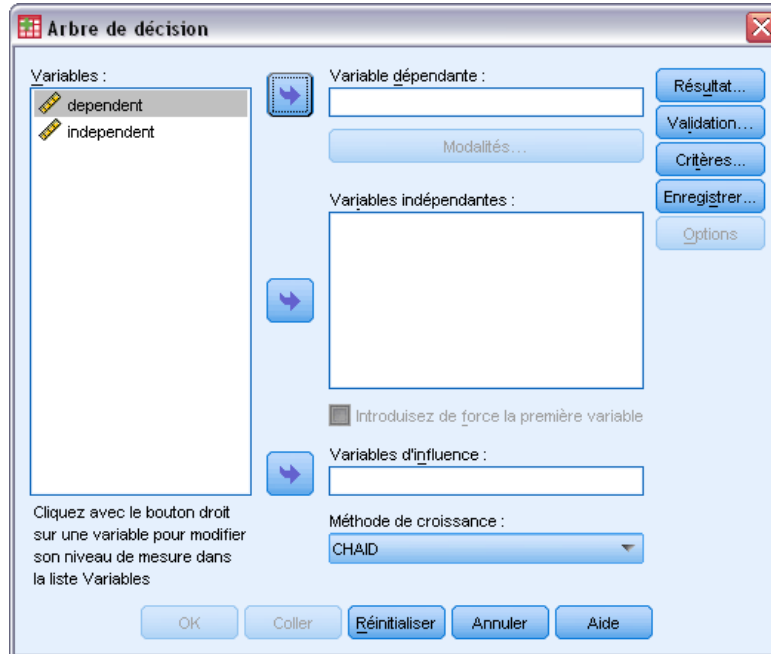
Les deux variables de ce fichier de données sont numériques et elles ont toutes deux un niveau de mesure **d'échelle**. Cependant (comme nous le verrons plus tard), ces deux variables sont véritablement des variables qualitatives reposant sur des codes numériques qui font office de valeurs de modalité.

- ▶ Pour lancer une analyse d'arbre de décision, choisissez les options suivantes dans les menus :
Analyse > Classification > Arbre...

Les icônes situées en regard des deux variables dans la liste de variables source indiquent qu'elles seront traitées comme des variables d'échelle.

Figure 3-1

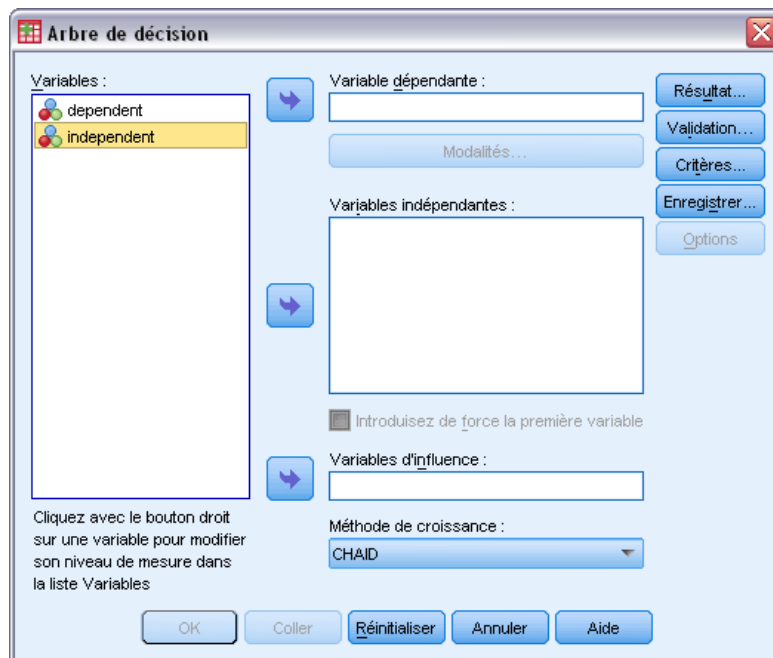
Boîte de dialogue principale Arbre de décision comportant deux variables d'échelle



- ▶ Sélectionnez la variable dépendante *dependante*.
- ▶ Sélectionnez la variable indépendante *independante*.
- ▶ Cliquez sur OK pour exécuter la procédure.
- ▶ Ouvrez à nouveau la boîte de dialogue Arbre de décision et cliquez sur Réinitialiser.
- ▶ Cliquez avec le bouton droit sur *dependante* dans la liste source et sélectionnez Nominal dans le menu contextuel.
- ▶ Procédez de la même façon pour la variable *independante* de la liste source.

Les icônes en regard de chaque variable indiquent qu'elles seront traitées comme des variables nominales.

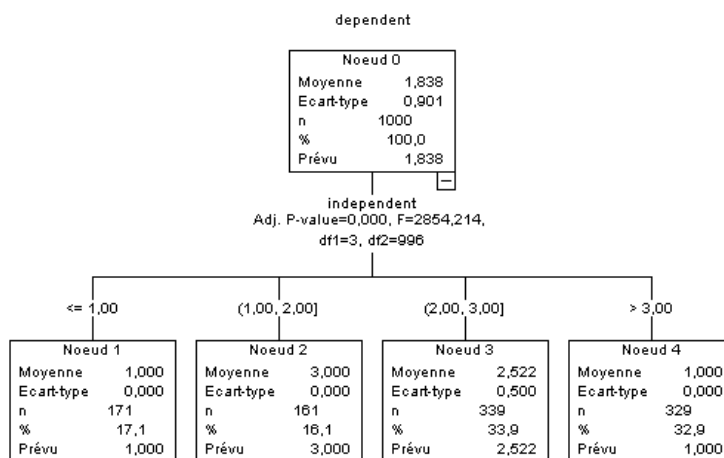
Figure 3-2
Icônes nominales de la liste source



- Sélectionnez *dépendante* pour la variable dépendante et *indépendante* pour la variable indépendante, et cliquez sur OK pour relancer la procédure.

Comparons à présent les deux arbres obtenus. Tout d'abord, observons l'arbre dans lequel les deux variables numériques sont traitées en tant que variables d'échelle.

Figure 3-3
Arbre dont les deux variables sont traitées comme des variables d'échelle



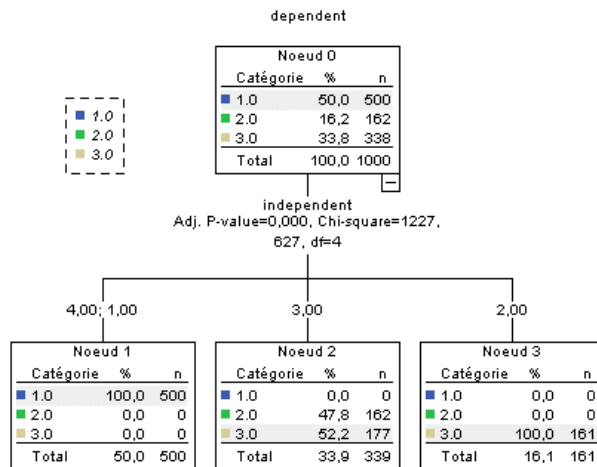
- Chaque noeud d'arbre montre la valeur « prévue », qui est la valeur moyenne de la variable dépendante de ce noeud. Pour une variable qui est réellement qualitative, la moyenne ne sera peut-être pas une statistique représentative.
- L'arbre comporte quatre noeuds enfant, un pour chaque valeur de la variable indépendante.

Les modèles d'arbre fusionnent souvent des noeuds similaires, mais pour une variable d'échelle, seules les valeurs attenantes peuvent être fusionnées. Dans cet exemple, aucune valeur attenante n'était suffisamment identique pour que des noeuds aient pu fusionner.

L'arbre dans lequel les deux variables sont traitées comme des variables nominales est légèrement différent à plusieurs égards.

Figure 3-4

Arbre dont les deux variables sont traitées comme des variables nominales



- Au lieu d'une prévision, chaque noeud contient un tableau d'effectifs indiquant le nombre d'observations (effectif et pourcentage) de chaque modalité de la variable dépendante.
- La modalité « prévue », correspondant à la modalité comportant l'effectif le plus élevé dans chaque noeud, est sélectionnée. Par exemple, la modalité prévue pour le noeud 2 est la modalité 3.
- Au lieu de quatre noeuds enfant, il n'en existe que trois, avec deux valeurs de la variable indépendante fusionnées en un seul noeud.

Les deux valeurs indépendantes fusionnées en un même noeud sont 1 et 4. Etant donné que, par définition, les valeurs nominales ne suivent aucun ordre inhérent, la fusion des valeurs non attenantes est autorisée.

Affectation permanente du niveau de mesure

Lorsque vous modifiez le niveau de mesure d'une variable dans la boîte de dialogue Arbre de décision, cette modification est temporaire et n'est pas enregistrée dans le fichier de données. De plus, vous ne connaîtrez peut-être pas toujours le niveau de mesure correct de toutes les variables.

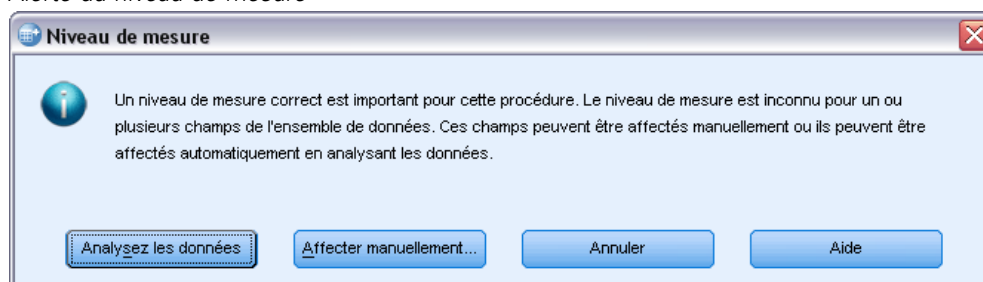
L'option Définir les propriétés de variable peut vous aider à déterminer le niveau de mesure correct de chaque variable et de modifier de manière permanente le niveau de mesure affecté. Pour utiliser l'option Définir les propriétés de variable :

- ▶ A partir des menus, sélectionnez :
Données > Définir les propriétés de variables

Variables avec niveau de mesure inconnu

L'alerte du niveau de mesure apparaît lorsque le niveau de mesure d'une ou plusieurs variables (champs) de l'ensemble 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.

Figure 3-5
Alerte du niveau de mesure



- **Analysez les données.** Lit les données dans l'ensemble de données actifs et attribue le niveau de mesure par défaut à tous les champs ayant un niveau de mesure inconnu. Si l'ensemble de données est important, cette action peut prendre un certain temps.
- **Attribuer manuellement.** Ouvre une boîte de dialogue qui répertorie tous les champs ayant un niveau de mesure inconnu. Vous pouvez utiliser cette boîte de dialogue pour attribuer un niveau de mesure à ces champs. Vous pouvez également attribuer un niveau de mesure dans l'affichage des variables de l'éditeur de données.

Le niveau de mesure étant important pour cette procédure, vous ne pouvez pas accéder à la boîte de dialogue d'exécution de cette procédure avant que tous les champs n'aient des niveaux de mesure définis.

Effets des étiquettes de valeur sur les modèles d'arbre

L'interface de la boîte de dialogue Arbre de décision suppose que, soit *toutes* les valeurs non manquantes d'une variable dépendante qualitative (nominale, ordinale) disposent d'étiquettes de valeurs définies, soit qu'*aucune* n'en dispose. Certaines fonctions ne sont disponibles que si au moins deux valeurs non manquantes de la variable dépendante qualitative disposent d'étiquettes de valeur. Si au moins deux valeurs non manquantes disposent d'étiquettes de valeur définies,

toutes les observations contenant d'autres valeurs ne disposant pas d'étiquettes de valeur seront exclues de l'analyse.

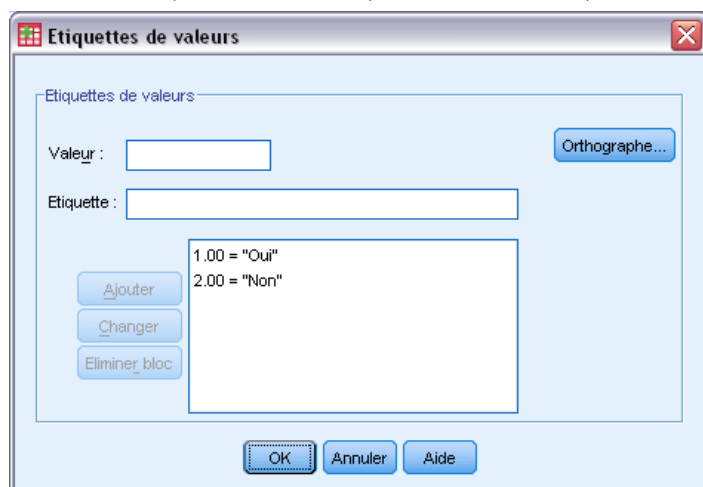
Dans cet exemple, le fichier de données d'origine ne contient aucune étiquette de valeur définie, et lorsque la variable dépendante est traitée comme une variable nominale, le modèle d'arbre utilise toutes les valeurs non manquantes dans l'analyse. Dans cet exemple, ces valeurs sont 1, 2 et 3.

Qu'arrive-t-il lorsque certaines variables dépendantes disposent d'étiquettes de valeur définies, mais pas toutes ?

- ▶ Dans la fenêtre de l'éditeur de données, cliquez sur l'onglet Affichage des variables.
- ▶ Cliquez sur la cellule Valeurs pour la variable *dépendante*.

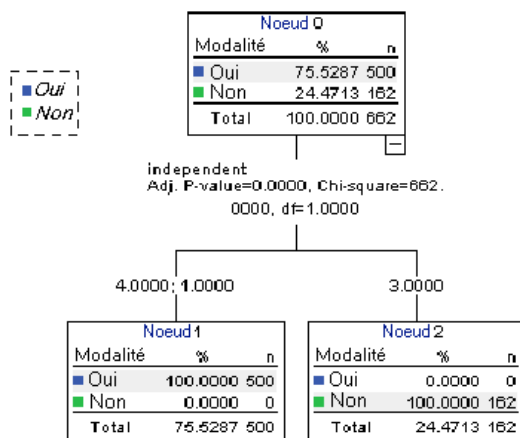
Figure 3-6

Définition d'étiquettes de valeurs pour une variable dépendante



- ▶ Tout d'abord, saisissez 1 pour Valeur et Oui pour Etiquette de valeur, puis cliquez sur Ajouter.
- ▶ Saisissez ensuite 2 pour Valeur et Non pour Etiquette de valeur, puis cliquez à nouveau sur Ajouter.
- ▶ Cliquez ensuite sur OK.
- ▶ Ouvrez à nouveau la boîte de dialogue Arbre de décision. Dans la boîte de dialogue, l'option *dépendante* doit encore être sélectionnée en tant que variable dépendante, ainsi qu'un niveau de mesure nominal.
- ▶ Cliquez sur OK pour exécuter à nouveau la procédure.

Figure 3-7
Arbre de variable dépendante nominale avec étiquettes de valeur partielles



A présent, seules les deux valeurs de variable dépendante comportant des étiquettes de valeur définies sont incluses dans le modèle d'arbre. Toutes les observations ayant la valeur 3 pour la variable dépendante ont été exclues, même si cela n'est peut être pas évident si vous ne connaissez pas bien les données.

Affectation d'étiquettes de valeur à toutes les valeurs

Pour éviter d'oublier par accident les valeurs qualitatives valides dans l'analyse, utilisez Définir les propriétés de variable pour affecter des étiquettes de valeur à toutes les valeurs de variable dépendante des données.

Lorsque les informations du dictionnaire de données sont affichées pour le *nom* de variable dans la boîte de dialogue Définir les propriétés de variable, vous pouvez voir que, même si plus de 300 observations ont la valeur 3 pour cette variable, aucune étiquette de valeur n'a été définie pour cette valeur.

Figure 3-8

Variable avec étiquettes de valeur partielles dans la boîte de dialogue Définir les propriétés de variable

Variable actuelle : Etiquette :

Niveau de mesure : Type :

Rôle : Largeur : Décimales :

Valeurs non étiquetées :

Grille d'étiquettes de valeur : Entrez ou modifiez des étiquettes dans la grille. Vous pouvez saisir des valeurs supplémentaires dans la partie inférieure.

	Modifié	Manquant	Effectif	Valeur	Etiquette
1	<input type="checkbox"/>	<input type="checkbox"/>	500	1.00	Oui
2	<input type="checkbox"/>	<input type="checkbox"/>	162	2.00	Non
3	<input type="checkbox"/>	<input type="checkbox"/>	338	3.00	
4	<input type="checkbox"/>	<input type="checkbox"/>			

Observations analysées : Valeurs non étiquetées :

Limite de la liste de valeurs :

Utilisation des arbres de décision pour évaluer le risque de crédit

Une banque tient à jour une base de données contenant des informations chronologiques sur les clients ayant emprunté de l'argent, indiquant s'ils ont remboursé la somme empruntée ou manqué à leurs engagements. Vous pouvez utiliser un modèle d'arbre pour analyser les caractéristiques de ces deux groupes de clients et pour construire des modèles afin de prédire la probabilité selon laquelle les demandeurs de prêt risquent de ne pas parvenir à rembourser leur emprunt.

Les données de crédit se trouvent dans le fichier *tree_credit.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans IBM SPSS Decision Trees 20.](#)

Création du modèle

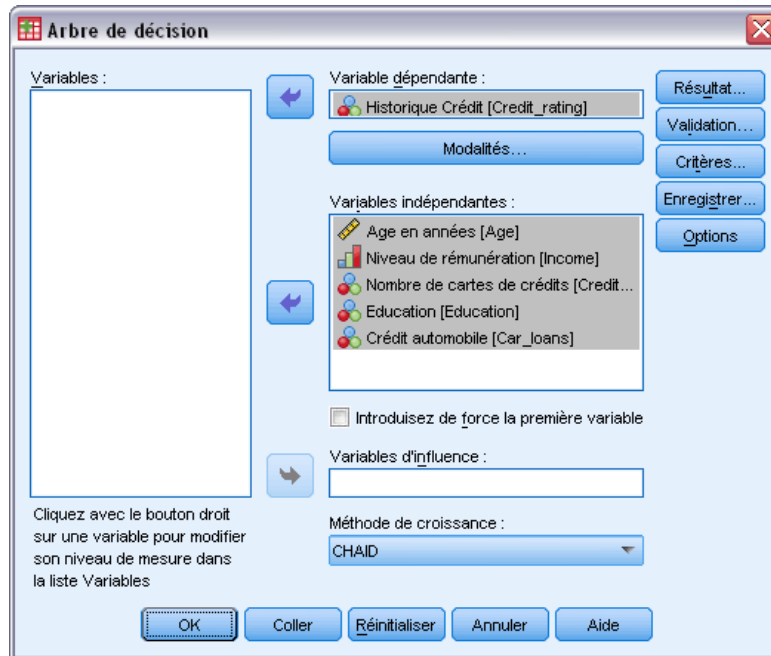
La procédure Arbre de décision propose différentes méthodes de création de modèles d'arbre. Dans cet exemple, nous utiliserons la méthode par défaut :

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

Construction du modèle d'arbre CHAID

- Pour lancer une analyse d'arbre de décision, choisissez les options suivantes dans les menus :
Analyse > Classification > Arbre...

Figure 4-1
Boîte de dialogue Arbre de décision



- ▶ Sélectionnez la variable dépendante *Cote de solvabilité*.
- ▶ Sélectionnez toutes les variables restantes en tant que variables indépendantes. (La procédure exclut automatiquement les variables qui n'apportent rien au modèle final.)

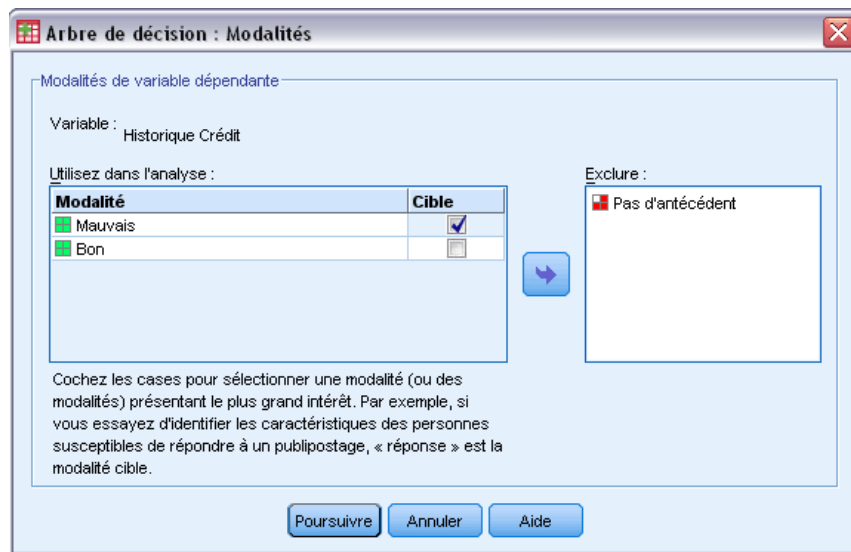
A ce stade, vous pourriez déjà exécuter la procédure et produire un modèle d'arbre de base, mais nous allons sélectionner quelques types de sortie supplémentaires et apporter de petits ajustements aux critères utilisés pour générer le modèle.

Sélection des modalités cible

- ▶ Cliquez sur le bouton Modalités qui figure sous la variable dépendante sélectionnée.

Dans la boîte de dialogue Modalités qui apparaît, vous pouvez indiquer les modalités cible de variable dépendante souhaitées. Les modalités cible n'ont pas d'impact direct sur le modèle d'arbre, mais certains types de sortie et options ne sont disponibles que si vous en avez sélectionné.

Figure 4-2
Boîte de dialogue Modalités



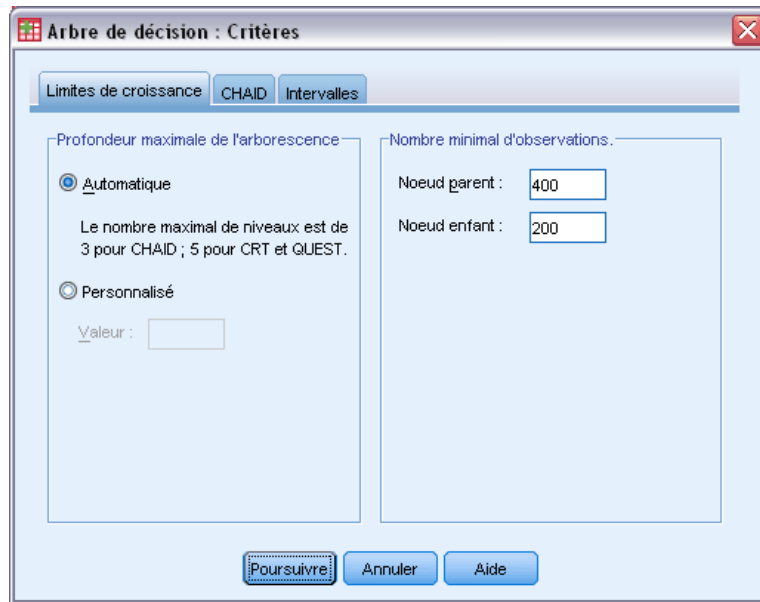
- ▶ Cochez la case Cible de la modalité *Mauvaise*. Les clients avec une mauvaise cote de solvabilité (qui ne parviennent pas à rembourser leur emprunt) sont alors considérés comme la modalité cible à étudier.
- ▶ Cliquez sur Poursuivre.

Spécification des critères de croissance de l'arbre

Pour cet exemple, nous avons voulu présenter un arbre relativement simple ; nous limiterons donc la croissance de l'arbre en augmentant le nombre minimum d'observations pour les noeuds parent et enfant.

- ▶ Dans la boîte de dialogue Arbre de décision principale, cliquez sur Critères.

Figure 4-3
Boîte de dialogue Critères, onglet Limites de croissance



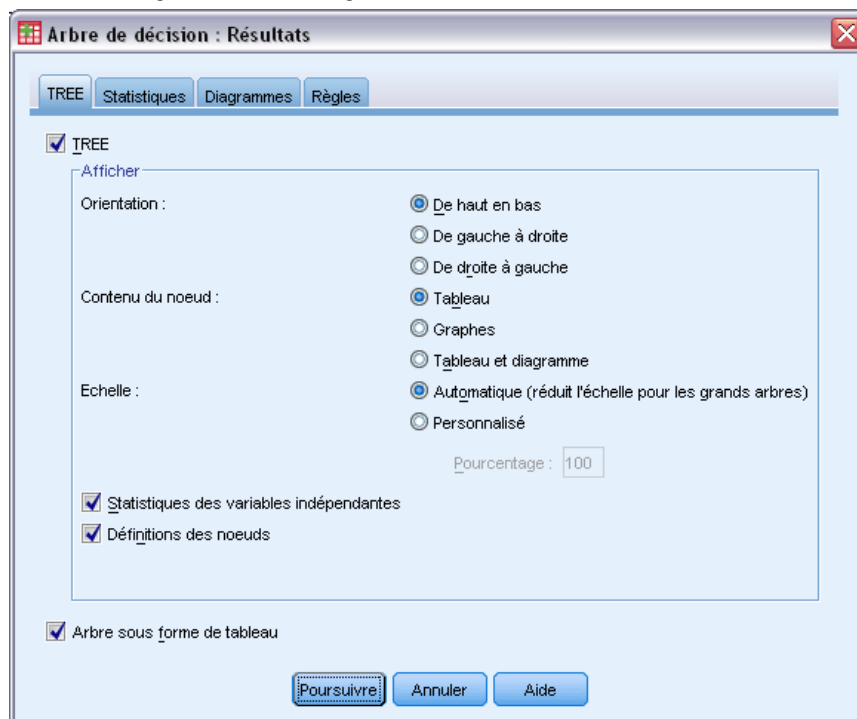
- ▶ Dans le groupe Nombre minimal d'observations, saisissez 400 pour l'option Noeud parent et 200 pour l'option Noeud enfant.
- ▶ Cliquez sur Poursuivre.

Sélection de types de sortie supplémentaires

- ▶ Dans la boîte de dialogue Arbre de décision principale, cliquez sur Résultat.

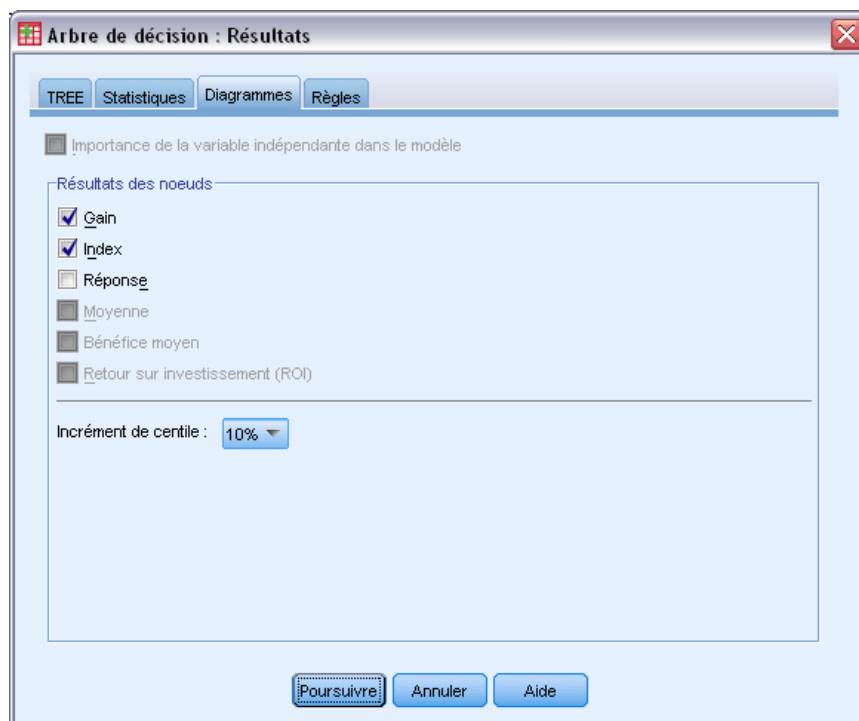
Dans la boîte de dialogue à onglets qui apparaît, vous pouvez sélectionner divers types de sortie supplémentaires.

Figure 4-4
Boîte de dialogue Résultats, onglet Arbre



- ▶ Dans l'onglet Arbre, cochez la case Arbre au format tableau.
- ▶ Cliquez ensuite sur l'onglet Diagrammes.

Figure 4-5
Boîte de dialogue Résultats, onglet Diagrammes



- Cochez Gain et Index.

Remarque : Ces diagrammes requièrent une modalité cible pour la variable dépendante. Dans cet exemple, l'onglet Diagrammes n'est accessible que lorsque vous avez indiqué une ou plusieurs modalités cible.

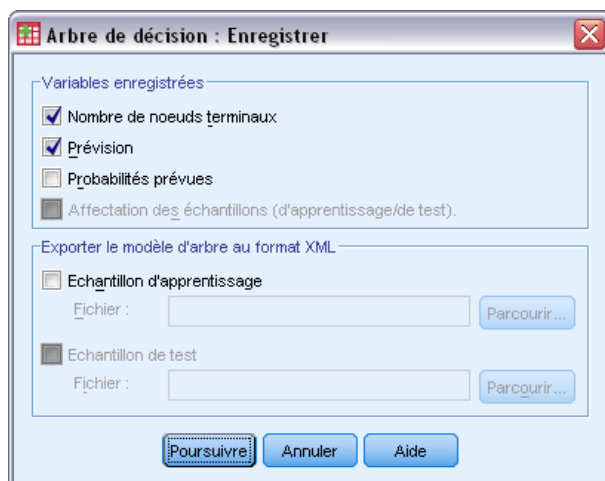
- Cliquez sur Poursuivre.

Enregistrement de prévisions

Vous pouvez enregistrer des variables contenant des informations sur les prévisions de modèle. Par exemple, vous pouvez enregistrer la cote de solvabilité prévue pour chaque observation et la comparer à la cote de solvabilité réelle.

- Dans la boîte de dialogue Arbre de décision principale, cliquez sur Enregistrer.

Figure 4-6
Enregistrer



- ▶ Cochez les cases Nombre de noeuds terminaux, Prévisions et Probabilités prévues.
- ▶ Cliquez sur Poursuivre.
- ▶ Dans la boîte de dialogue Arbre de décision principale, cliquez sur OK pour lancer la procédure.

Evaluation du modèle

Dans notre exemple, les résultats du modèle comprennent :

- Des tableaux fournissant des informations sur le modèle.
- Un diagramme d'arbre.
- Des diagrammes fournissant des indications sur les performances du modèle.
- Des variables de prévision de modèle ajoutées à l'ensemble de données actif.

Tableau récapitulatif des modèles

Figure 4-7
Récapitulatif du modèle

Spécifications	Méthode de développement	CHAID	
	Variable dépendante :	Notation Crédit	
	Variabes indépendantes	Age, Niveau de revenu, Nombre de cartes de crédit, Education, Crédit automobile	
	Validation	NONE	
	Profondeur maximum de l'arbre		3
	Nombre minimum d'observations d'un noeud parent		400
Résultats	Nombre minimum d'observations d'un noeud enfant		200
	Variabes indépendantes incluses	Niveau de revenu, Nombre de cartes de crédit, Age	
	Nombre de noeuds		10
	Nombre de noeuds terminaux		6
	Profondeur		3

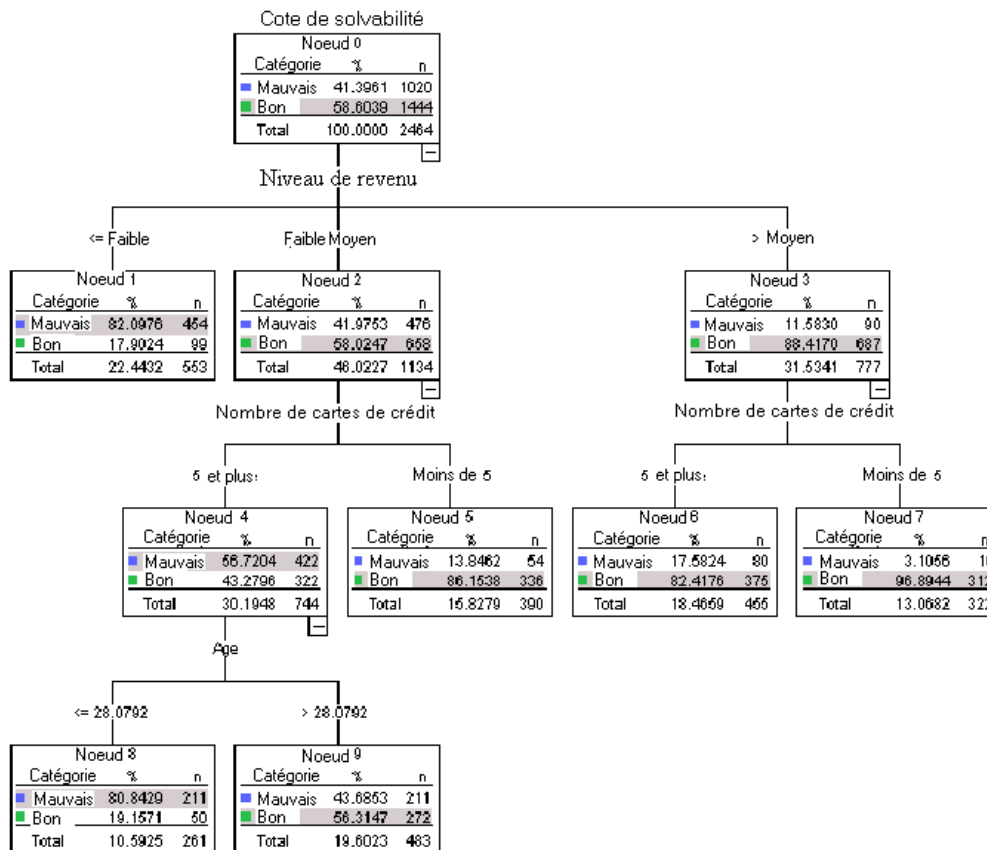
Le tableau récapitulatif des modèles fournit des informations très générales sur les spécifications utilisées pour construire le modèle et sur le modèle obtenu.

- La section Spécifications donne des informations sur les paramètres utilisés pour générer le modèle d'arbre, dont les variables utilisées lors de l'analyse.
- La section Résultats contient des informations sur le nombre total de noeuds et le nombre de noeuds terminaux, la profondeur de l'arbre (nombre de niveaux en dessous du noeud racine) et les variables indépendantes incluses dans le modèle final.

Cinq variables indépendantes ont été indiquées, mais seulement trois incluses dans le modèle final. Les variables concernant le *niveau d'études* et le nombre de *prêts auto* en cours n'apportaient rien au modèle et ont donc été supprimées du modèle final.

Diagramme de l'arbre

Figure 4-8
Diagramme d'arbre du modèle de la cote de solvabilité



Le diagramme d'arbre est une représentation graphique du modèle d'arbre. Il montre les éléments suivants :

- Dans le cadre de la méthode CHAID, le *niveau de revenu* est la meilleure variable indépendante de la *cote de solvabilité*.
- Pour la modalité des revenus faibles, le *niveau de revenu* est même la seule variable indépendante significative de la *cote de solvabilité*. Parmi les clients de la banque appartenant à cette modalité, 82 % ont manqué à leurs engagements. Etant donné qu'aucun noeud enfant ne figure sous cet élément, il est considéré comme un noeud **terminal**.
- Pour les modalités de revenus moyens et élevés, la meilleure variable indépendante suivante est le *nombre de cartes de crédit*.
- Pour les clients disposant de revenus moyens et détenteurs d'au moins cinq cartes de crédit, le modèle inclut une variable indépendante supplémentaire : l'*âge*. En effet, plus de 80 % des clients âgés de 28 ans ou moins ont une mauvaise cote de solvabilité, alors que, pour les plus de 28 ans, la mauvaise cote de solvabilité ne concerne plus qu'un peu moins de la moitié des personnes.

Vous pouvez utiliser l'éditeur d'arbre pour masquer et afficher les branches sélectionnées, modifier les couleurs et les polices, et sélectionner des sous-ensembles d'observations en fonction des noeuds sélectionnés. [Pour plus d'informations, reportez-vous à la section Sélection d'observations dans les noeuds sur p. 76.](#)

Tableau de l'arbre

Figure 4-9

Tableau d'arbre de la cote de solvabilité

Noeud	Mauvais		Bon		Total		Modalité estimée	Noeud parent
	N :	Pourcentage :	N :	Pourcentage :	N :	Pourcentage :		
0	1020	41,4%	1444	58,6%	***	100,0%	Bon	
1	454	82,1%	99	17,9%	553	22,4%	Mauvais	0
2	476	42,0%	658	58,0%	***	46,0%	Bon	0
3	90	11,6%	687	88,4%	777	31,5%	Bon	0
4	422	56,7%	322	43,3%	744	30,2%	Mauvais	2
5	54	13,8%	336	86,2%	390	15,8%	Bon	2
6	80	17,6%	375	82,4%	455	18,5%	Bon	3
7	10	3,1%	312	96,9%	322	13,1%	Bon	3
8	211	80,8%	50	19,2%	261	10,6%	Mauvais	4
9	211	43,7%	272	56,3%	483	19,6%	Bon	4

Méthode de développement: CHAID

Variable dépendante: Notation Crédit

Comme son nom l'indique, le tableau d'arbre reprend dans un tableau la plupart des informations essentielles du diagramme d'arbre. Le tableau affiche les informations suivantes pour chaque noeud :

- Nombre et pourcentage d'observations dans chaque modalité de la variable dépendante.
- Modalité prévue de la variable dépendante. Dans cet exemple, il s'agit de la modalité de la *cote de solvabilité*, avec plus de 50 % des observations dans ce noeud, les cotes de solvabilité possibles étant au nombre de deux.
- Noeud parent de chaque noeud de l'arbre. Notez que le noeud 1, —celui du niveau de revenu faible—, n'est parent d'aucun noeud. En effet, il s'agit d'un noeud terminal qui n'a donc pas de noeuds enfant.

Figure 4-10
Tableau d'arbre de la cote de solvabilité (suite)

Noeud	Variable indépendante principale				Valeurs de scission
	Variable	Sig. ^a	Khi-deux	ddl	
1	Niveau de revenu	,000	662,457	2	<= Faible
2	Niveau de revenu	,000	662,457	2	(Faible, Moyen]
3	Niveau de revenu	,000	662,457	2	> Moyen
4	Nombre de cartes de crédit	,000	193,113	1	5 et plus
5	Nombre de cartes de crédit	,000	193,113	1	Moins de 5
6	Nombre de cartes de crédit	,000	38,587	1	5 et plus
7	Nombre de cartes de crédit	,000	38,587	1	Moins de 5
8	Age	,000	95,299	1	<= 28,079205 818990676
9	Age	,000	95,299	1	> 28,079205 818990676

Méthode de développement: CHAID

Variable dépendante: Notation Crédit

- Variable indépendante utilisée pour scinder le noeud.
- Valeur Khi-deux (l'arbre ayant été généré à l'aide de la méthode CHAID), degrés de liberté (*ddl*) et seuil de signification (*Sig.*) de la scission. Dans la plupart des applications pratiques, vous ne serez certainement intéressé que par le seuil de signification, inférieur à 0,0001 pour toutes les scissions de ce modèle.
- Valeurs de la variable indépendante du noeud.

Remarque : Pour les variables indépendantes ordinales et d'échelle, les intervalles de l'arbre et du tableau d'arbre sont généralement exprimés sous la forme (*valeur1, valeur2*], ce qui signifie supérieur à valeur1 et inférieur ou égal à valeur2. Dans notre exemple, le niveau de revenu n'a que trois valeurs possibles : *faible, moyen* et *élevé*. (*faible, moyen*] signifie donc tout simplement *moyen*. De même, *>moyen* signifie *>élevé*.

Gains pour les noeuds

Figure 4-11
Gains pour les noeuds

Gains pour les noeuds						
Noeud	Noeud		Gain		Réponse	Index
	N :	Pourcentage :	N :	Pourcentage :		
1	553	22,4%	454	44,5%	82,1%	198,3%
8	261	10,6%	211	20,7%	80,8%	195,3%
9	483	19,6%	211	20,7%	43,7%	105,5%
6	455	18,5%	80	7,8%	17,6%	42,5%
5	390	15,8%	54	5,3%	13,8%	33,4%
7	322	13,1%	10	1,0%	3,1%	7,5%

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

Le tableau de gains pour les noeuds récapitule les informations ayant trait aux noeuds terminaux du modèle.

- Seuls les noeuds terminaux, —noeuds au niveau desquels l'arbre arrête de se développer—, sont répertoriés dans ce tableau. La plupart du temps, seuls ces noeuds, qui représentent les meilleures prévisions de classification du modèle, vous intéressent.
- Les valeurs de gain fournissant des informations sur les modalités cible, ce tableau n'est disponible que si vous avez indiqué une ou plusieurs modalités cible. Dans notre exemple, il n'existe qu'une modalité cible. Un seul tableau de gains pour les noeuds est donc généré.
- *Noeud N* représente le nombre d'observations dans chaque noeud terminal et *Pourcentage de noeud* le pourcentage d'observations dans chaque noeud par rapport au nombre total d'observations.
- *Gain N* représente le nombre d'observations dans chaque noeud terminal de la modalité cible et *Pourcentage de gain* le pourcentage d'observations dans la modalité cible par rapport au nombre total d'observations de cette modalité—, à savoir le nombre et le pourcentage d'observations affichant une mauvaise côte de solvabilité dans l'exemple qui nous occupe.
- Pour les variables dépendantes qualitatives, l'option *Réponse* correspond au pourcentage d'observations dans le noeud de la modalité cible spécifiée. Dans cet exemple, il s'agit des mêmes pourcentages que ceux affichés pour la modalité *Mauvaise* dans le diagramme d'arbre.
- Pour les variables dépendantes qualitatives, l'option *Index* correspond au rapport entre le pourcentage de réponses de la modalité cible et le pourcentage de réponses de l'intégralité de l'échantillon.

Valeurs d'index

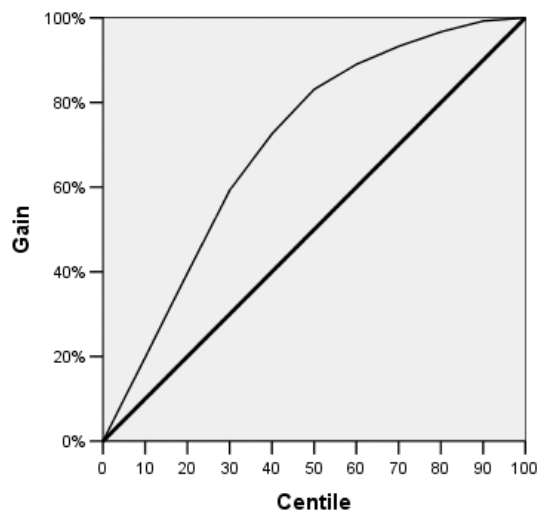
La valeur d'index indique l'importance de la différence existant entre le pourcentage de modalité cible *observé* pour le noeud et le pourcentage *attendu*. Le pourcentage de modalité cible du noeud racine représente le pourcentage attendu, avant prise en compte de l'impact des variables indépendantes.

Une valeur d'index supérieure à 100 % signifie qu'il existe plus d'observations dans la modalité cible que le pourcentage global de la modalité cible. Inversement, une valeur d'index inférieure à 100 % signifie qu'il existe moins d'observations dans la modalité cible que le pourcentage global.

Diagramme des gains

Figure 4-12

Diagramme des gains de la modalité cible de mauvaise cote de solvabilité



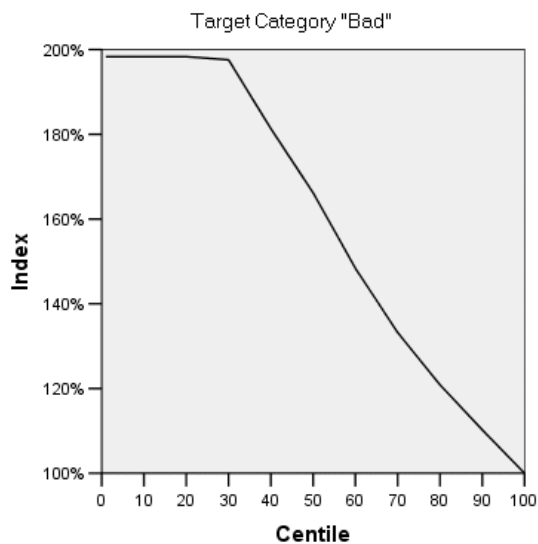
Le diagramme des gains indique que le modèle est assez bon.

Les diagrammes de gains cumulés commencent à 0 % et finissent à 100 %. Les diagrammes de gains des bons modèles présentent une hausse rapide en direction de la valeur 100 %, puis se stabilisent. Les modèles ne fournissant pas d'informations suivent la diagonale de référence.

Diagramme des index

Figure 4-13

Diagramme des index de la modalité cible de mauvaise cote de solvabilité



Le diagramme des index indique également que le modèle est bon. Les diagrammes d'index cumulés commencent généralement au-dessus de 100 % pour descendre ensuite progressivement jusqu'à 100 %.

Les valeurs d'index des bons modèles débutent bien au-dessus de 100 %, restent à un niveau élevé pendant un certain temps, puis diminuent rapidement en direction de la valeur 100 %. Dans les modèles ne fournissant pas d'informations, la ligne reste aux alentours de 100 % dans l'intégralité du diagramme.

Estimation du risque et classification

Figure 4-14
Tableaux de risque et de classement

Risque

Echantillon	Estimation	Erreur std.
Formation	,245	,012

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

Classification

Echantillon	Observations	Prévisions		
		Mauvais	Bon	Pourcentage correct
Formation	Mauvais	366	153	70,5%
	Bon	148	561	79,1%
	Pourcentage global	41,9%	58,1%	75,5%

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

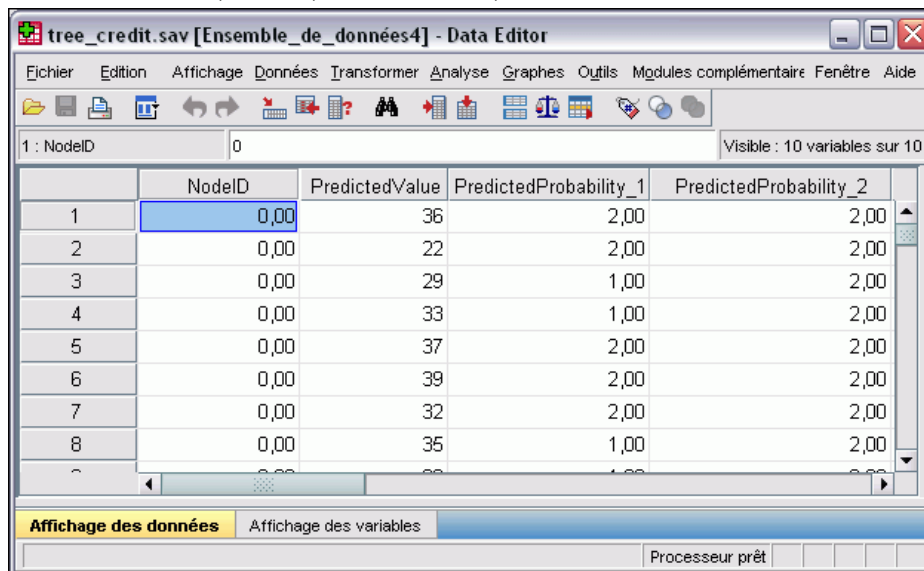
Les tableaux de risques et de classement permettent d'évaluer rapidement l'efficacité du modèle.

- L'estimation du risque (qui s'élève à 0,205) indique que la modalité prévue par le modèle (bonne ou mauvaise cote de solvabilité) est fautive dans 20,5 % des observations. Le risque de classification erronée d'un client est donc d'environ 21 %.
- Les résultats du tableau de classement confirment les informations données par l'estimation du risque. Le tableau indique que le modèle classe correctement environ 79,5 % des clients.

Le tableau de classement indique néanmoins que le modèle pose un problème potentiel : le modèle ne prédit une mauvaise cote de solvabilité que pour 41,9 % des clients réellement concernés. Autrement dit, 58,1 % de ces clients sont placés à tort dans les bons clients.

Prévisions

Figure 4-15
Nouvelles variables pour les prévisions et les probabilités



	NodeID	PredictedValue	PredictedProbability_1	PredictedProbability_2
1	0,00	36	2,00	2,00
2	0,00	22	2,00	2,00
3	0,00	29	1,00	2,00
4	0,00	33	1,00	2,00
5	0,00	37	2,00	2,00
6	0,00	39	2,00	2,00
7	0,00	32	2,00	2,00
8	0,00	35	1,00	2,00
9	0,00	38	1,00	2,00
10	0,00	34	2,00	2,00

Quatre nouvelles variables ont été créées dans l'ensemble de données actif :

IDNoeud. Nombre de noeuds terminaux pour chaque observation.

ValeurPrévue. Prédiction de la variable dépendante pour chaque observation. La variable dépendante est codée de la manière suivante : 0 = *Mauvais* et 1 = *Bon*. Une prévision de 0 indique donc que l'observation obtiendra une mauvaise cote de solvabilité.

ProbabilitéPrévue. Probabilité selon laquelle l'observation appartient à chaque modalité de la variable dépendante. Étant donné que la variable dépendante ne peut recevoir que deux valeurs, deux variables sont créées :

- **ProbabilitéPrévue_1.** Probabilité selon laquelle l'observation appartient à la modalité de mauvaise cote de solvabilité.
- **ProbabilitéPrévue_2.** Probabilité selon laquelle l'observation appartient à la modalité de bonne cote de solvabilité.

La probabilité prévue correspond simplement à la proportion d'observations dans chaque modalité de la variable dépendante, pour le noeud terminal contenant chaque observation. Par exemple, dans le noeud 1, 82 % des observations appartiennent à la modalité de mauvaise cote de solvabilité et 18 % à celle de bonne cote de solvabilité, d'où des probabilités prévues de 0,82 et de 0,18, respectivement.

Dans les variables dépendantes qualitatives, la prévision est la modalité correspondant à la proportion la plus élevée d'observations dans le noeud terminal de chaque observation. Par exemple, la prévision de la première observation est de 1 (bonne cote de solvabilité), car environ 56 % des observations de son noeud terminal ont une bonne cote de solvabilité. Inversement, la prévision de la seconde observation est de 0 (mauvaise cote de solvabilité), car environ 81 % des observations de son noeud terminal ont une mauvaise cote de solvabilité.

Toutefois, si vous définissez des coûts, la relation entre la modalité prévue et les probabilités prévues n'est pas toujours aussi évidente. [Pour plus d'informations, reportez-vous à la section Affectation de coûts aux résultats sur p. 79.](#)

Amélioration du modèle

Globalement, le modèle présente un taux de classification correcte légèrement inférieur à 80 %. Cette constatation se reflète dans la plupart des noeuds terminaux, où la modalité estimée (la modalité sélectionnée dans le noeud) est identique à la modalité réelle pour au moins 80 % des observations.

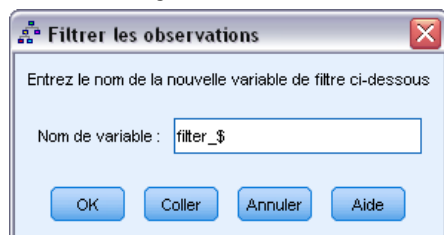
Cependant, un noeud terminal présente des observations réparties équitablement entre bonne et mauvaise cote de solvabilité. La cote de solvabilité prévue pour le noeud 9 est de type « bonne ». Pourtant, seulement 56 % des observations du noeud ont réellement une bonne cote de solvabilité. Autrement dit, presque la moitié des observations de ce noeud (44 %) ont une modalité prévue incorrecte. Or, si l'objectif principal est d'identifier les risques de mauvaise cote de solvabilité, ce noeud n'est pas très efficace.

Sélection d'observations dans les noeuds

Étudions les observations du noeud 9 pour voir si les données nous apportent des informations supplémentaires utiles.

- ▶ Dans le Viewer, double-cliquez sur l'arbre pour ouvrir l'éditeur d'arbre.
- ▶ Cliquez sur le noeud 9 pour le sélectionner. (Pour sélectionner plusieurs noeuds, appuyez sur la touche Ctrl tout en cliquant sur les noeuds souhaités.)
- ▶ A partir des menus de l'Éditeur d'arbre, sélectionnez : Règles > Filtrer les observations...

Figure 4-16
Boîte de dialogue Filtrer les observations



La boîte de dialogue Filtrer les observations crée une variable de filtre et applique un paramètre de filtre basé sur les valeurs de cette variable. Le nom par défaut de la variable de filtre est *filter_\$*.

- Les observations des noeuds sélectionnés reçoivent une valeur de 1 pour la variable de filtre.
- Toutes les autres observations recevront la valeur 0 et seront exclues des analyses suivantes jusqu'à modification de l'état du filtre.

Dans notre exemple, les observations n'appartenant pas au noeud 9 seront donc éliminées pour l'instant (mais pas supprimées).

- Cliquez sur OK pour créer la variable de filtre et appliquer la condition correspondante.

Figure 4-17

Observations filtrées dans l'éditeur de données

	Credit_rating	Age	Income	Credit_cards	Education
1	0,00	36	2,00	2,00	2,00
2	0,00	22	2,00	2,00	2,00
3	0,00	29	1,00	2,00	1,00
4	0,00	33	1,00	2,00	2,00
5	0,00	37	2,00	2,00	2,00
6	0,00	39	2,00	2,00	2,00
7	0,00	32	2,00	2,00	2,00
8	0,00	35	1,00	2,00	1,00
9	0,00	32	1,00	2,00	1,00
10	0,00	25	2,00	2,00	2,00

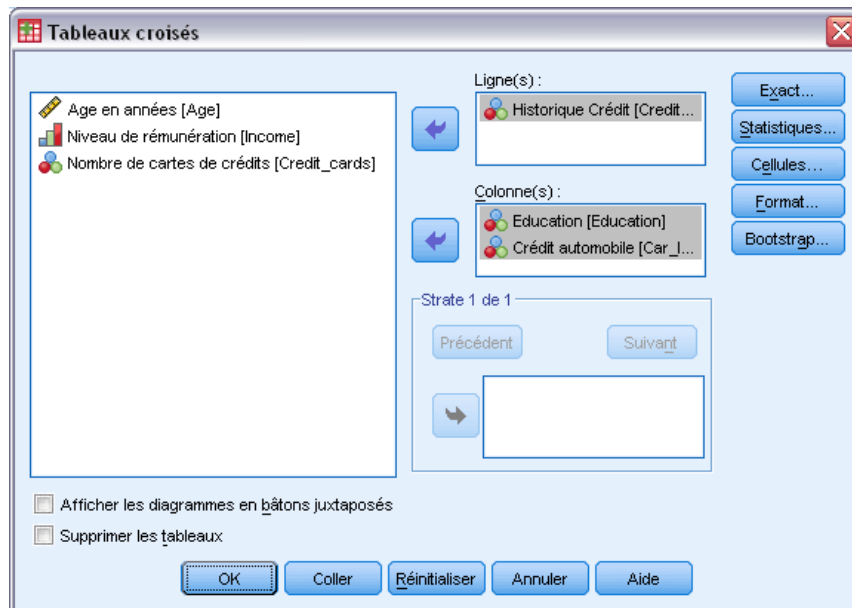
Dans l'éditeur de données, les observations éliminées sont signalées par un trait en diagonale barrant le numéro de la ligne. Les observations n'appartenant pas au noeud 9 sont éliminées. En revanche, les observations du noeud 9 ne sont pas exclues. Les analyses effectuées par la suite n'intégreront donc que ces observations.

Examen des observations sélectionnées

Pour commencer l'examen des observations du noeud 9, vous pouvez étudier les variables non utilisées par le modèle. Dans cet exemple, toutes les variables du fichier de données ont été incluses dans l'analyse, mais deux d'entre elles n'ont pas été intégrées au modèle final : le *niveau d'études* et le nombre de *prêts auto*. Si la procédure les a omises du modèle final, c'est certainement qu'elles ne sont pas très significatives, mais jetons-y tout de même un oeil.

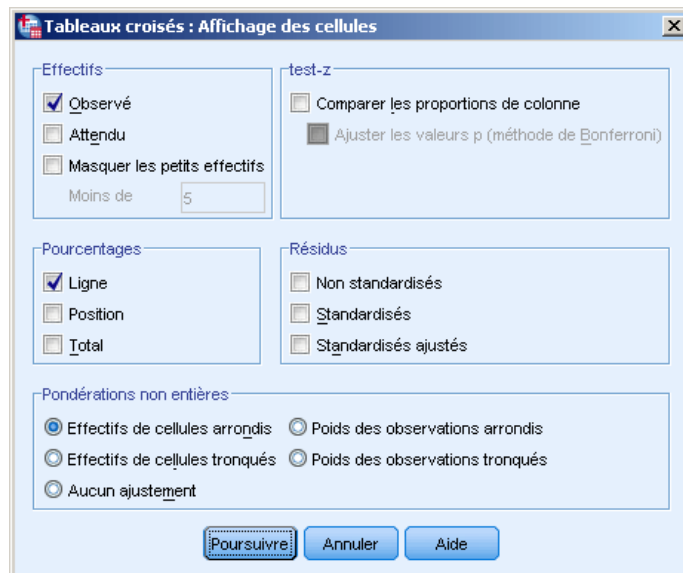
- A partir des menus, sélectionnez :
Analyse > Statistiques descriptives > Tableaux croisés

Figure 4-18
Boîte de dialogue Tableaux croisés



- ▶ Sélectionnez la variable de ligne *Cote de solvabilité*.
- ▶ Sélectionnez les variables de colonne *Années d'étude* et *Prêts auto*.
- ▶ Cliquez sur *Cells*.

Figure 4-19
Boîte de dialogue Tableaux croisés : Contenu des cases (cellules)



- ▶ Dans le groupe *Pourcentages*, cochez la case *Ligne*.

- Cliquez ensuite sur Continuer, puis, dans la boîte de dialogue Tableaux croisés principale, cliquez sur OK pour lancer la procédure.

En examinant les tableaux croisés, vous constatez que, pour les deux variables exclues du modèle, les observations des modalités de bonne et de mauvaise cote de solvabilité diffèrent peu.

Figure 4-20

Tableaux croisés des observations du noeud sélectionné

Tableau croisé Notation Crédit * Education

			Education		Total
			Niveau bac et bac	Etudes supérieures	
Notation Crédit	Mauvais	Effectif	110	101	211
		% dans Notation Crédit	50,3%	49,7%	100,0%
	Bon	Effectif	128	144	272
		% dans Notation Crédit	49,7%	50,3%	100,0%
Total	Effectif		238	245	483
	% dans Notation Crédit		49,9%	50,1%	100,0%

Tableau croisé Notation Crédit * Crédit automobile

			Crédit automobile		Total
			Aucun ou 1	2 et plus	
Notation Crédit	Mauvais	Effectif	18	193	211
		% dans Notation Crédit	17,5%	82,5%	100,0%
	Bon	Effectif	39	233	272
		% dans Notation Crédit	49,5%	50,5%	100,0%
Total	Effectif		57	426	483
	% dans Notation Crédit		11,8%	88,2%	100,0%

- En ce qui concerne le *niveau d'études*, un peu plus de la moitié des observations dénotant une mauvaise cote de solvabilité correspondent à des personnes ayant seulement un niveau bac, tandis qu'un peu plus de la moitié des observations dénotant une bonne cote de solvabilité correspondent à des personnes ayant poursuivi des études supérieures. Cette différence n'est pas significative sur le plan statistique.
- En ce qui concerne les *prêts auto*, le pourcentage d'observations dotées d'une bonne cote de solvabilité et correspondant à des personnes n'ayant contracté aucun prêt-auto ou un seul est supérieur au pourcentage d'observations dotées d'une mauvaise cote de solvabilité. Toutefois, la grande majorité des observations des deux groupes correspond à des personnes ayant contracté plusieurs prêts auto.

Nous savons maintenant pourquoi ces variables n'ont pas été incluses dans le modèle final, mais nous n'avons toujours pas trouvé le moyen d'améliorer les prévisions du noeud 9. Si d'autres variables n'ont pas été retenues pour l'analyse, vous pouvez les examiner avant de poursuivre.

Affectation de coûts aux résultats

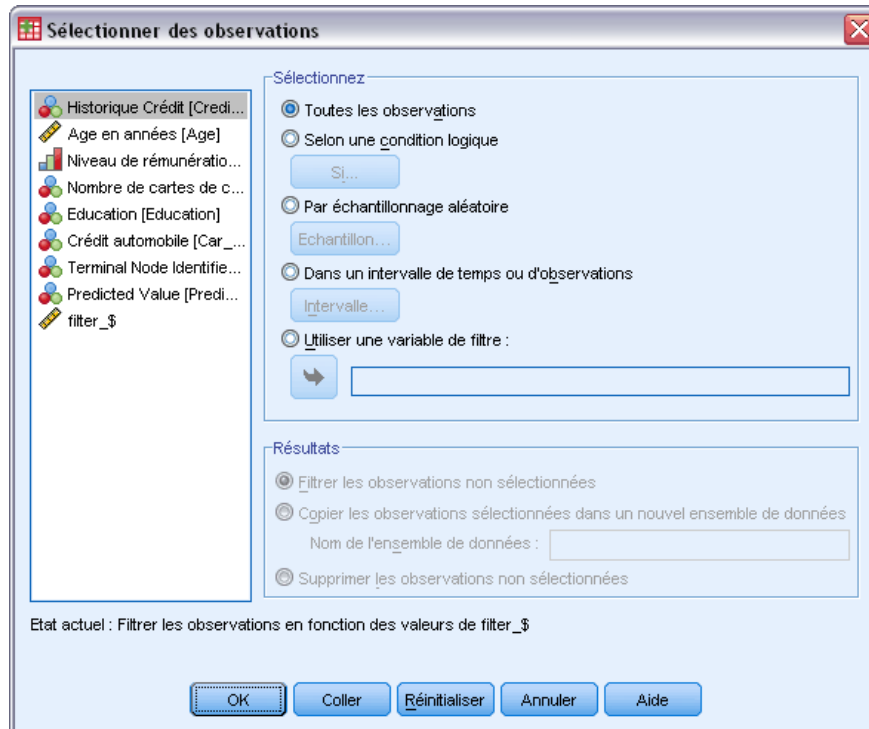
Comme nous l'avons constaté précédemment, les observations du noeud 9 se répartissent équitablement dans les deux modalités de cote de solvabilité. En outre, la modalité prévue est de type « bonne » ce qui est problématique si votre objectif principal est de construire un modèle identifiant correctement les risques de mauvaise cote de solvabilité. Bien qu'il soit

peut-être impossible d'améliorer les performances du nœud 9, vous pouvez affiner le modèle afin d'accroître le taux de classification correcte des observations dénotant une mauvaise cote de solvabilité— (même si cette opération entraînera un taux de classification erronée plus élevé pour les observations dénotant une bonne cote de solvabilité).

En premier lieu, désactivez le filtrage de sorte à utiliser toutes les observations dans l'analyse.

- ▶ A partir des menus, sélectionnez :
Données > Sélectionner des observations
- ▶ Dans la boîte de dialogue Sélectionner des observations, sélectionnez Toutes les observations, puis cliquez sur OK.

Figure 4-21
Boîte de dialogue Sélectionner des observations



- ▶ Ouvrez de nouveau la boîte de dialogue Arbre de décision et cliquez sur Options.

- Cliquez sur l'onglet Coûts de classification erronée.

Figure 4-22

Boîte de dialogue Options, onglet Coûts de classification erronée

Arbre de décision : Options

Valeurs manquantes Coûts de classification erronée Bénéfices

Egale pour toutes les modalités
 Personnalisé

Modalité estimée :

	Mauvais	Bon
Réal Modalité : Mauvais	0	2
Bon	1	0

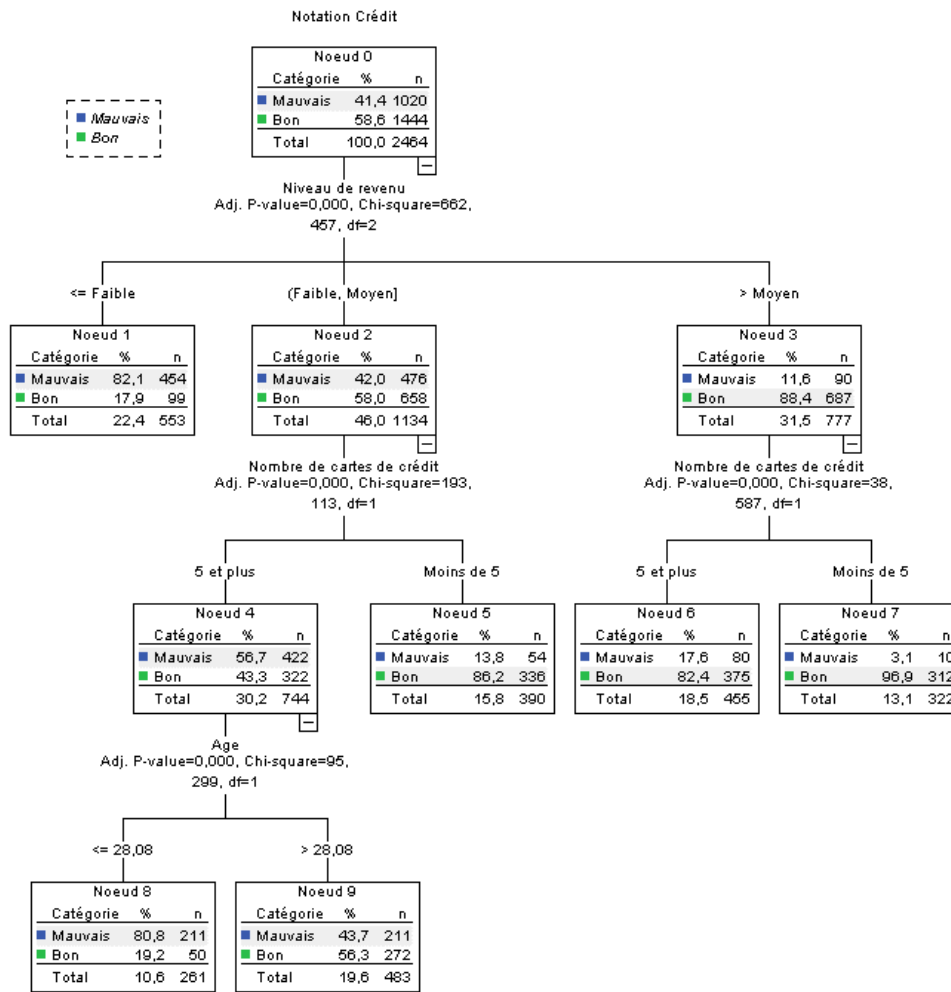
Rendre la matrice symétrique

- Sélectionnez Personnalisé, et saisissez la valeur 2 pour la modalité réelle *Mauvaise* / modalité prévue *Bonne*.

Vous indiquez ainsi à la procédure que le coût de la classification erronée d'une mauvaise cote de solvabilité potentielle dans le type bonne cote de solvabilité est deux fois plus élevé que celui de la classification erronée d'une bonne cote de solvabilité potentielle dans le type mauvaise cote de solvabilité.

- Cliquez sur Poursuivre, puis sur OK dans la boîte de dialogue principale pour exécuter la procédure.

Figure 4-23
Modèle d'arbre avec valeurs de coût ajustées



Au premier abord, l'arbre généré par la procédure ressemble à l'arbre initial. En examinant plus attentivement l'arbre, vous constatez toutefois que certaines modalités prévues ont changé, même si la distribution des observations dans chaque noeud est identique.

La modalité prévue reste identique dans l'ensemble des noeuds terminaux à l'exception d'un seul : le noeud 9. La modalité prévue est désormais de type *mauvaise*, même si un peu plus de la moitié des observations appartiennent à la modalité *bonne*.

Etant donné que nous avons indiqué à la procédure que le coût de la classification erronée d'une mauvaise cote de solvabilité potentielle dans le type bonne cote de solvabilité était plus élevé, tous les noeuds où les observations étaient réparties équitablement entre les deux modalités présentent désormais une modalité prévue de type *mauvaise*, même si une petite majorité des observations appartient à la modalité de type *bonne*.

Ce changement au niveau de la modalité prévue est reflété par le tableau de classement.

Figure 4-24

Tableaux de risques et de classement basés sur les coûts ajustés

Risque	
Estimation	Erreur std.
,288	,011

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

Classification			
Observations	Prévisions		
	Mauvais	Bon	Pourcentage correct
Mauvais	876	144	85,9%
Bon	421	1023	70,8%
Pourcentage global	52,6%	47,4%	77,1%

Méthode de développement: CHAID
Variable dépendante: Notation Crédit

- Près de 86 % des risques de mauvaise cote de solvabilité sont désormais correctement classés, contre seulement 65 % auparavant.
- En revanche, le taux de classification correcte des bonnes cotes de solvabilité potentielles est passé de 90 % à 71 %, et le taux de classification correcte global de 79,5 % à 77,1 %.

Notez également que l'estimation du risque et le taux de classification correcte global ne vont plus dans le même sens. Avec un taux de classification correcte global de 77,1 %, vous attendez une estimation du risque de 0,229. Or, l'augmentation du coût de la classification erronée des mauvaises cotes de solvabilité potentielles a, dans cet exemple, accru la valeur du risque, ce qui rend l'interprétation moins évidente.

Récapitulatif

Les modèles d'arbre permettent de classer les observations dans des groupes identifiés par des caractéristiques spécifiques, comme celles associées aux clients ayant des antécédents de bonne ou de mauvaise cote de solvabilité auprès de la banque. Si un résultat prévu particulier est plus important que tous les autres résultats possibles, vous pouvez affiner le modèle pour associer à ce résultat un coût de classification erronée plus élevé. Notez néanmoins qu'en réduisant le taux de classification erronée d'un résultat, vous augmentez ceux des autres résultats.

Construction d'un modèle d'analyse

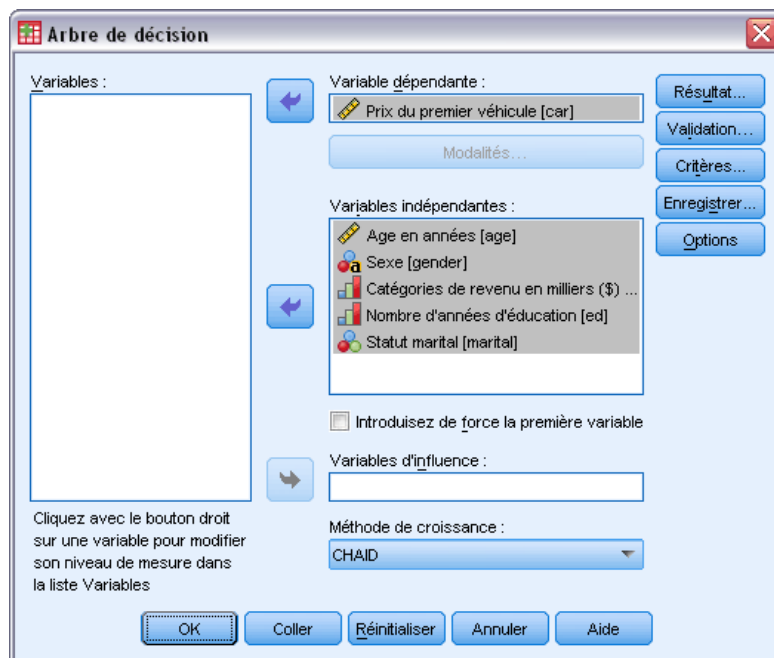
L'une des fonctions les plus puissantes et utiles de la procédure Arbre de décision réside dans la possibilité de construire des modèles pouvant ensuite être appliqués à d'autres fichiers de données pour prévoir des résultats. Par exemple, à partir d'un fichier de données contenant à la fois des informations démographiques et des informations sur le prix d'achat de véhicules, nous pouvons élaborer un modèle pouvant être utilisé pour prévoir le nombre de personnes présentant les mêmes caractéristiques démographiques qui sont susceptibles de dépenser pour l'achat d'une nouvelle voiture—, puis appliquer ce modèle à d'autres fichiers de données dans lesquels figurent des informations démographiques, mais pas d'informations sur l'achat du précédent véhicule.

Pour cet exemple, nous utiliserons le fichier de données *tree_car.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans IBM SPSS Decision Trees 20.](#)

Construction du modèle

- Pour lancer une analyse d'arbre de décision, choisissez les options suivantes dans les menus : Analyse > Classification > Arbre...

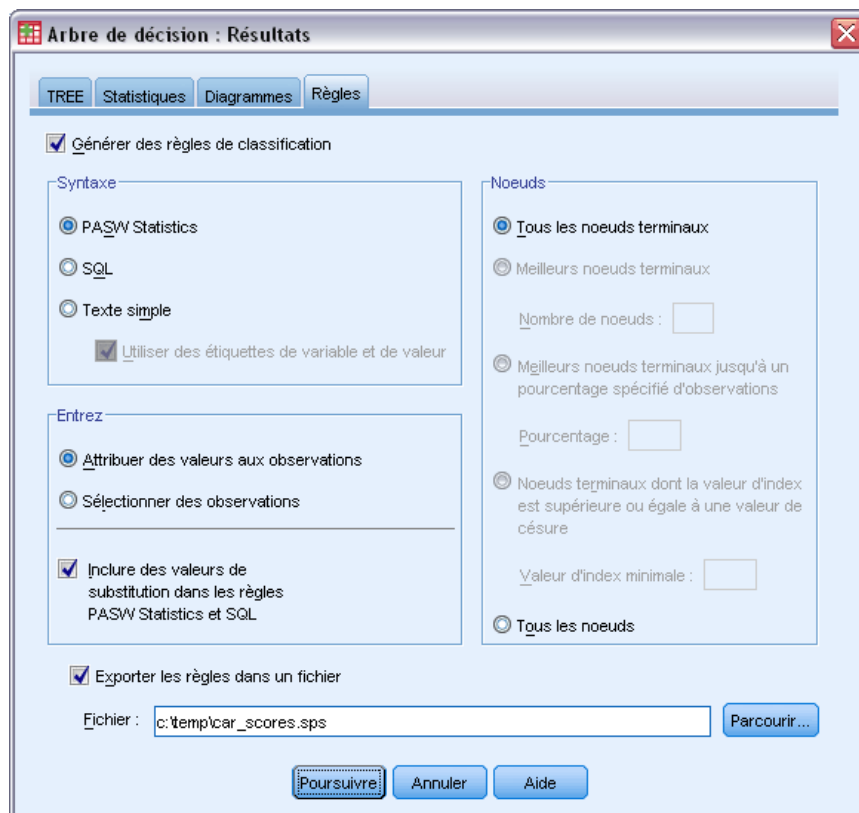
Figure 5-1
Boîte de dialogue Arbre de décision



- Sélectionnez *Prix du véhicule principal* en tant que variable dépendante.

- ▶ Sélectionnez toutes les variables restantes en tant que variables indépendantes. (La procédure exclut automatiquement les variables qui n'apportent rien au modèle final.)
- ▶ Pour la méthode de croissance, sélectionnez CRT.
- ▶ Cliquez sur Résultat.

Figure 5-2
Boîte de dialogue Résultat, onglet Règles



- ▶ Cliquez sur l'onglet Règles.
- ▶ Sélectionnez (cochez) Générer des règles de classification.
- ▶ Pour Syntaxe, sélectionnez IBM® SPSS® Statistics.
- ▶ Pour Type, sélectionnez Attribuer des valeurs aux observations.
- ▶ Sélectionnez (cochez) Exporter les règles dans un fichier et saisissez un nom de fichier et l'emplacement d'un répertoire.

Mémorisez ce nom de fichier et cet emplacement ou notez-les car vous allez en avoir besoin plus tard. Si vous n'avez pas saisi de chemin de répertoire, vous ne savez peut-être pas où le fichier a été enregistré. Vous pouvez utiliser le bouton Parcourir pour parcourir les répertoires et accéder à un emplacement spécifique (et valide).

- ▶ Cliquez sur Continuer, puis sur OK pour lancer la procédure et construire le modèle d'arbre.

Evaluation du modèle

Avant d'appliquer le modèle à d'autres fichiers de données, vous voudrez peut-être vous assurer que le modèle fonctionne relativement bien avec les données d'origine utilisées pour sa construction.

Récapitulatif des modèles

Figure 5-3
Tableau récapitulatif des modèles

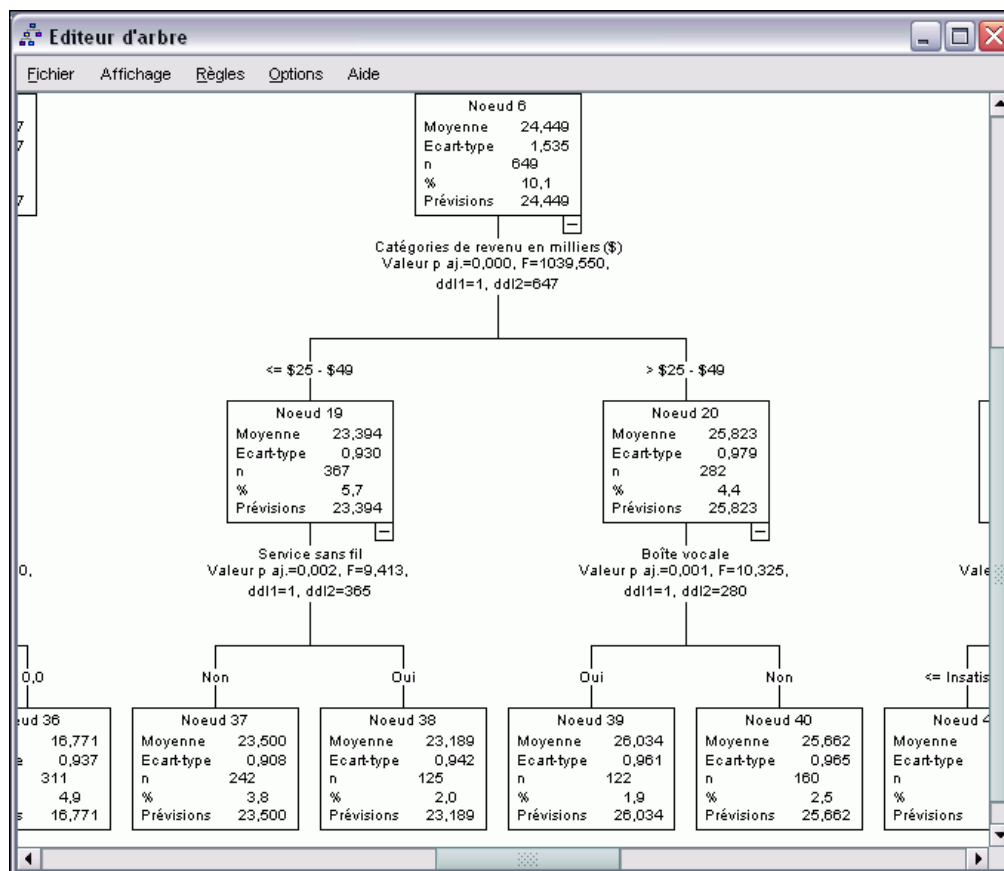
Spécifications	Méthode de développement	CRT	
	Variable dépendante :	Prix du premier véhicule	
	Variables indépendantes	Age en années, Genre, Catégorie de revenu en milliers, Niveau d'éducation, Statut marital	
	Validation	NONE	
	Profondeur maximum de l'arbre		5
	Nombre minimum d'observations d'un noeud parent		100
Résultats	Nombre minimum d'observations d'un noeud enfant		50
	Variables indépendantes incluses	Catégorie de revenu en milliers, Age en années, Niveau d'éducation	
	Nombre de noeuds		29
	Nombre de noeuds terminaux		15
	Profondeur		5

Le tableau récapitulatif des modèles indique que seulement trois des variables indépendantes sélectionnées ont apporté une contribution suffisamment significative pour être incluses dans le modèle final : les *revenus*, l'*âge* et la *formation*. Ces informations sont essentielles pour savoir si vous allez appliquer ce modèle à d'autres fichiers de données, étant donné que les variables indépendantes utilisées dans le modèle doivent être présentes dans tous les fichiers de données auxquels vous souhaitez appliquer le modèle.

Le tableau récapitulatif indique également que le modèle d'arbre n'est peut-être pas très simple car il comporte 29 noeuds et 15 noeuds terminaux. Cela ne pose pas de problème si vous avez besoin d'un modèle fiable et facile à appliquer plutôt que d'un modèle simple et facile à décrire ou à expliquer. Bien sûr, pour des raisons pratiques, vous souhaitez probablement un modèle ne reposant pas sur de trop nombreuses variables indépendantes (qualitatives). Dans ce cas, ce n'est pas un problème car seulement trois variables indépendantes sont incluses dans le modèle final.

Diagramme de modèle d'arbre

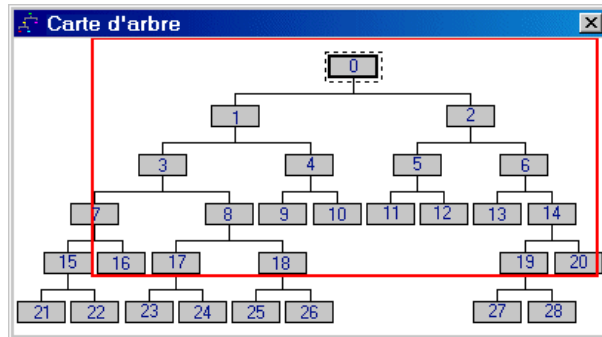
Figure 5-4
Diagramme de modèle d'arbre dans l'éditeur d'arbre



Le diagramme du modèle d'arbre comporte tellement de noeuds que l'affichage de l'intégralité du modèle risque d'être difficile. Il sera tellement petit que vous ne pourrez pas lire les informations contenues dans les noeuds. Vous pouvez utiliser la carte d'arbre pour voir l'intégralité de l'arbre :

- Dans le Viewer, double-cliquez sur l'arbre pour ouvrir l'éditeur d'arbre.
- A partir des menus de l'éditeur d'arbre, sélectionnez :
Affichage > Carte d'arbre

Figure 5-5
Carte d'arbre



- La carte d'arbre affiche l'intégralité de l'arbre. Si vous modifiez la taille de la fenêtre de la carte d'arbre, l'affichage de la carte sera agrandi ou réduit pour que l'arbre tienne dans la fenêtre.
- La zone sélectionnée dans la carte d'arbre est la zone de l'arbre affichée dans l'éditeur d'arbre.
- Vous pouvez utiliser la carte d'arbre pour parcourir l'arbre et sélectionner des noeuds.

Pour plus d'informations, reportez-vous à la section Carte d'arbre dans le chapitre 2 sur p. 43.

Pour les variables d'échelle dépendantes, chaque noeud indique la moyenne et l'écart-type de la variable dépendante. Le noeud 0 affiche le prix d'achat moyen global d'un véhicule d'environ 29,9 (en milliers), avec un écart-type d'environ 21,6.

- Le noeud 1, représentant les observations dont les revenus sont inférieurs à 75 (en milliers), dispose d'un prix moyen de véhicule de seulement 18,7.
- Au contraire, le noeud 2, représentant les observations dont les revenus sont supérieurs ou égaux à 75, dispose d'un prix moyen de véhicule de 60,9.

Un examen plus approfondi de l'arbre montrerait que l'âge et la *formation* ont également une relation avec le prix d'achat d'un véhicule, mais nous nous intéresserons pour l'instant à l'application pratique du modèle, plutôt qu'à l'examen détaillé de ses composants.

Estimation du risque

Figure 5-6
Tableau Risque

Risque	
Estimation	Erreur std.
68,485	2,985

Méthode de développement: CRT

Variable dépendante: Prix du premier véhicule

Aucun des résultats considérés jusqu'à présent n'indique s'il s'agit d'un modèle particulièrement bon. L'un des indicateurs des performances du modèle est l'estimation du risque. Pour une variable d'échelle dépendante, l'estimation du risque est la mesure de la variance intra-noeud, qui n'est pas forcément significative en elle-même. Une variance faible indique un modèle plus

adéquat, mais la variance est relative à l'unité de mesure. Si, par exemple, le prix a été enregistré à l'unité, et non en milliers, l'estimation du risque est mille fois supérieure.

Une interprétation correcte de l'estimation du risque avec une variable d'échelle dépendante demande un certain effort :

- La variance totale est égale à la variance intra-noeud (variance de l'erreur) plus la variance inter-noeuds (variance expliquée).
- La variance intra-noeud est la valeur de l'estimation du risque : 68.485.
- La variance totale est la variance des variables dépendantes avant la prise en considération des variables indépendantes, ce qui revient à la variance au niveau du noeud racine.
- L'écart-type indiqué au niveau du noeud racine est de 21,576 ; la variance totale correspond donc à cette valeur élevée au carré : 465.524.
- La proportion de la variance due à l'erreur (variance résiduelle) est de $68,485/465,524 = 0,147$.
- La proportion de la variance expliquée par le modèle est $1-0,147 = 0,853$, soit 85,3 %, ce qui indique que le modèle est relativement bon. (Il s'agit d'une interprétation similaire au taux de classification correct global d'une variable dépendante qualitative.)

Application du modèle à un autre fichier de données

Maintenant que le modèle a été jugé bon, nous pouvons l'appliquer à d'autres fichiers de données contenant des variables *âge*, *revenus* et *formation* similaires, et générer une nouvelle variable représentant le prix d'achat du véhicule prévu pour chaque observation du fichier. Ce processus est souvent appelé **analyse**.

Lorsque nous avons généré le modèle, nous avons précisé que les « règles » d'attribution des valeurs aux observations doivent être enregistrées dans un fichier texte, sous forme de syntaxe de commande. Nous allons à présent utiliser les commandes dans ce fichier pour générer des scores dans un autre fichier de données.

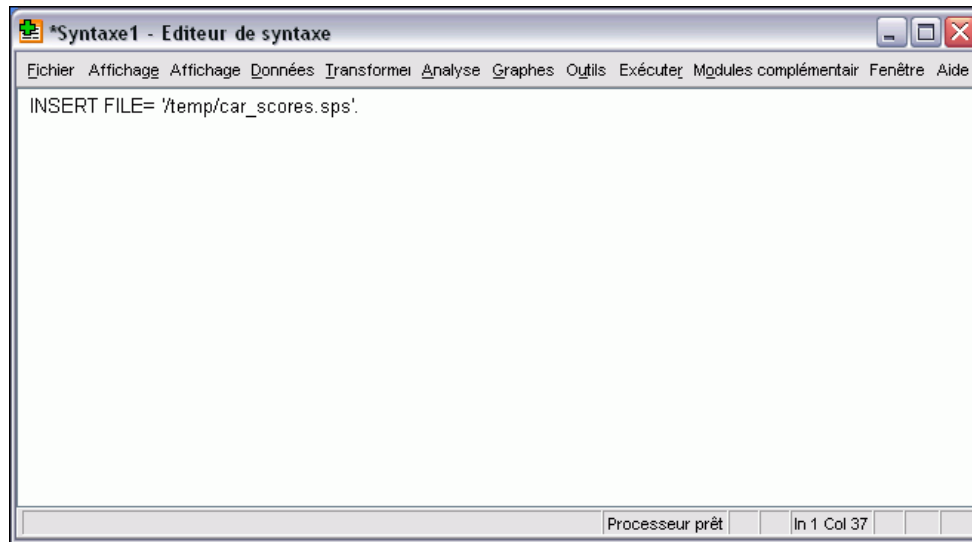
- ▶ Ouvrez le fichier de données *tree_score_car.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans IBM SPSS Decision Trees 20.](#)
- ▶ Ensuite, à partir des menus, sélectionnez :
Fichier > Nouveau > Syntaxe
- ▶ Dans la fenêtre de syntaxe de commande, entrez :

```
INSERT FILE=  
'/temp/car_scores.sps'.
```

Si vous avez utilisé un nom de fichier ou un emplacement différent, apportez les modifications nécessaires.

Figure 5-7

Fenêtre Syntaxe comportant la commande INSERT permettant d'exécuter un fichier de commande



La commande INSERT exécute les commandes dans le fichier indiqué, c'est-à-dire le fichier « règles » généré au moment de la création du modèle.

- ▶ A partir des menus de la fenêtre de syntaxe de commande, sélectionnez :
Exécuter > Tous

Figure 5-8

Prévisions ajoutées au fichier de données

1 : car	inccat	ed	marital	nod_001	pre_001	vs
1	3,00	1	1	10,00	30,56	
2	4,00	1	0	27,00	61,08	
3	2,00	3	1	24,00	17,13	
4	2,00	4	1	23,00	15,58	
5	1,00	2	0	21,00	9,39	
6	3,00	2	0	9,00	29,78	
7	1,00	1	0	22,00	10,22	
8	4,00	3	1	12,00	54,08	

Deux nouvelles variables sont ainsi ajoutées au fichier de données :

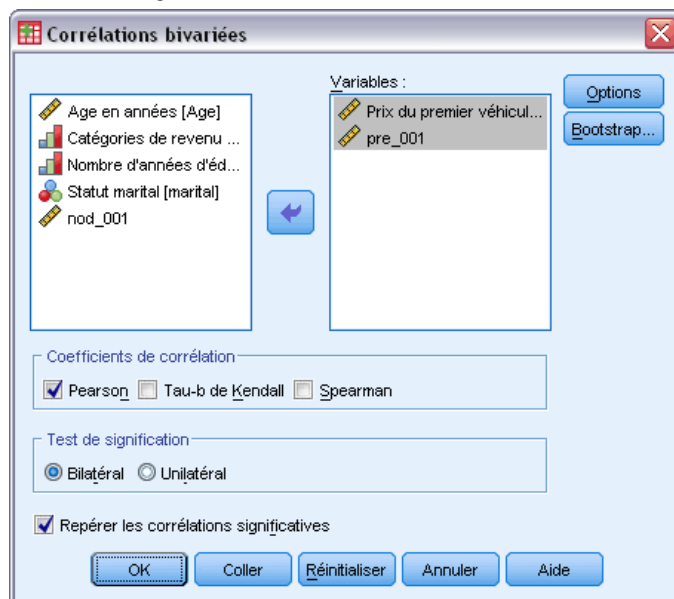
- *nod_001* contient le nombre de noeuds terminaux prévus par le modèle pour chaque observation.
- *pre_001* contient les prévisions du prix d'achat d'un véhicule pour chaque observation.

Étant donné que nous avons demandé des règles pour l'attribution de valeurs aux nœuds terminaux, le nombre de valeurs attendues possibles est identique au nombre de nœuds terminaux (15 dans ce cas). Par exemple, chaque observation disposant d'un nombre de nœuds prévus de 10 auront le même prix d'achat de véhicule prévu : 30.56. Il s'agit (non par hasard) de la valeur moyenne reportée pour le nœud terminal 10 dans le modèle d'origine.

Bien que le modèle soit normalement appliqué aux données pour lesquelles la valeur de la variable dépendante est inconnue, dans cet exemple, le fichier de données auquel le modèle est appliqué contient déjà ces informations ; vous pouvez ainsi comparer les prévisions du modèle aux valeurs réelles.

- ▶ A partir des menus, sélectionnez :
Analyse > Corrélation > Bivariée
- ▶ Sélectionnez *Prix du véhicule principal* et *pre_001*.

Figure 5-9
Boîte de dialogue *Corrélations bivariées*



- ▶ Cliquez sur OK pour exécuter la procédure.

Figure 5-10
Corrélation entre le prix prévu et le prix réel du véhicule

		pre_001	Prix du premier véhicule
pre_001	Corrélation de Pearson	1	,919**
	Sig. (bilatérale)		,000
	N	3290	3290
Prix du premier véhicule	Corrélation de Pearson	,919**	1
	Sig. (bilatérale)	,000	
	N	3290	3290

** . La corrélation est significative au niveau 0.01 (bilatéral).

La corrélation de 0,92 indique une corrélation positive très élevée entre le prix prévu et le prix réel du véhicule, signifiant que le modèle fonctionne bien.

Récapitulatif

Vous pouvez utiliser la procédure Arbre de segmentation pour construire des modèles pouvant ensuite être appliqués à d'autres fichiers de données, afin de prévoir des résultats. Le fichier de données cible doit contenir des variables portant le même nom que les variables indépendantes incluses dans le modèle final, mesurées dans la même unité et avec les mêmes valeurs manquantes éventuelles spécifiées par l'utilisateur. Cependant, la variable dépendante et les variables indépendantes exclues du modèle final ne doivent pas obligatoirement être présentes dans le fichier de données cible.

Valeurs manquantes dans les modèles d'arbre

Les diverses méthodes de croissance traitent les valeurs manquantes des variables indépendantes (explicatives) de différentes manières :

- CHAID et Exhaustive CHAID traitent toutes les valeurs manquantes par défaut et spécifiées par l'utilisateur pour chaque variable indépendante en tant que modalité unique. Pour les variables d'échelle indépendantes ou ordinales et en fonction des critères de croissance, cette modalité peut être fusionnée par la suite avec d'autres modalités de cette variable indépendante.
- CRT et QUEST utilisent des **valeurs de substitution** pour les variables indépendantes (explicatives). 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 variables indépendantes de rechange sont appelées valeurs de substitution.

L'exemple suivant montre la différence existant entre CHAID et CRT lorsque des valeurs manquantes de variables indépendantes sont utilisées dans le modèle.

Pour cet exemple, nous utiliserons le fichier de données *tree_missing_data.sav*. [Pour plus d'informations, reportez-vous à la section Fichiers d'exemple dans l'annexe A dans *IBM SPSS Decision Trees 20*.](#)

Remarque : Pour les variables indépendantes et dépendantes nominales, vous pouvez choisir de traiter les valeurs **manquantes spécifiées** comme des valeurs valides, auquel cas elles sont traitées comme n'importe quelle autre valeur non manquante. [Pour plus d'informations, reportez-vous à la section Valeurs manquantes dans le chapitre 1 sur p. 23.](#)

Valeurs manquantes avec CHAID

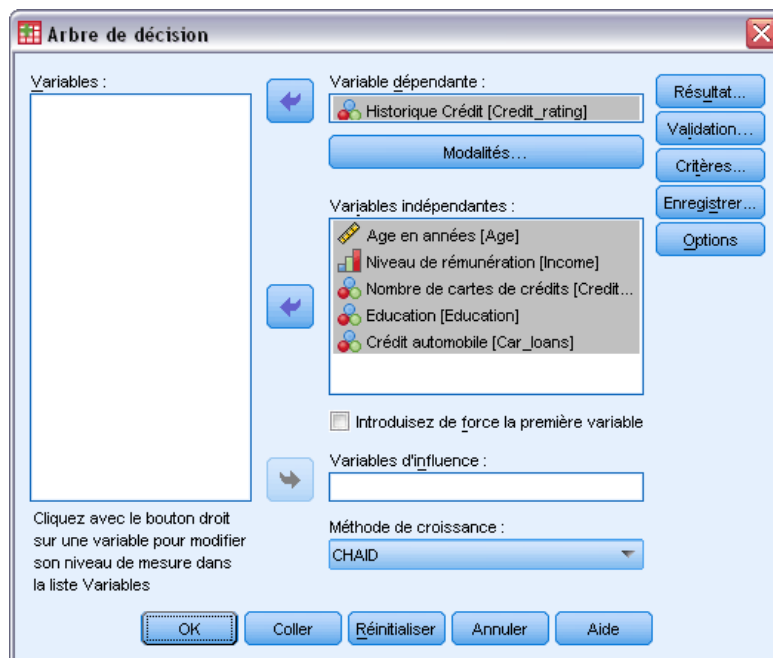
Figure 6-1
Données de crédit avec des valeurs manquantes

	Credit_rating	Age	Income	Credit_cards	Educ:
1	0	36	2,00	.	.
2	0,00	22	2,00	.	.
3	0,00	29	.	2,00	.
4	0,00	33	.	2,00	.
5	0,00	37	2,00	.	.
6	0,00	39	2,00	2,00	.
7	0,00	32	2,00	2,00	.
8	0,00	35	.	2,00	.
9	0,00	32	1,00	2,00	.
10	0,00	25	2,00	.	.
11	0,00	23	.	2,00	.

A l'instar de l'exemple du risque de crédit (pour plus d'informations, reportez-vous au [le chapitre 4](#)), cet exemple tente de construire un modèle permettant de classer les bons et les mauvais risques de crédit. La différence principale réside dans le fait que ce fichier de données contient des valeurs manquantes pour certaines variables indépendantes utilisées dans le modèle.

- Pour lancer une analyse d'arbre de décision, choisissez les options suivantes dans les menus :
Analyse > Classification > Arbre...

Figure 6-2
Boîte de dialogue Arbre de décision

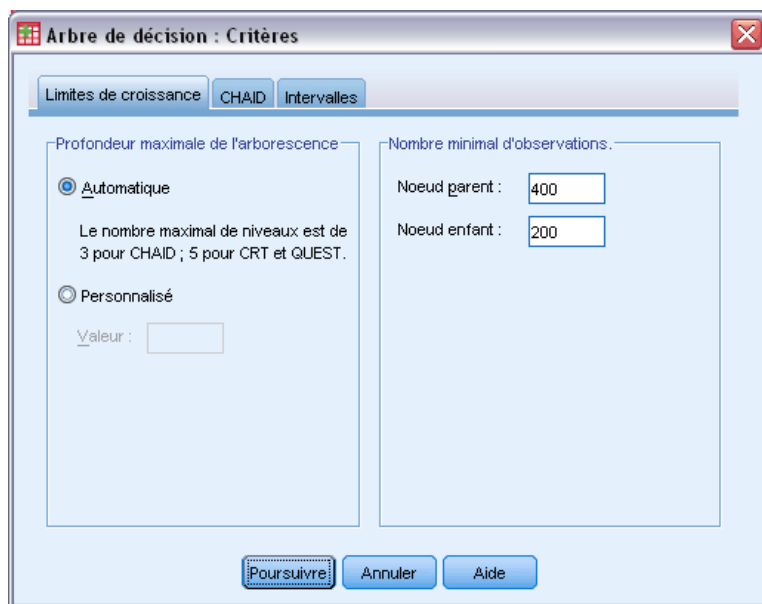


- ▶ Sélectionnez la variable dépendante *Cote de solvabilité*.
- ▶ Sélectionnez toutes les variables restantes en tant que variables indépendantes. (La procédure exclut automatiquement les variables qui n'apportent rien au modèle final.)
- ▶ Pour la méthode de croissance, sélectionnez CHAID.

Pour cet exemple, nous avons voulu présenter un arbre relativement simple ; nous limiterons donc la croissance de l'arbre en augmentant le nombre minimum d'observations dans les noeuds parent et enfant.

- ▶ Dans la boîte de dialogue Arbre de décision principale, cliquez sur Critères.

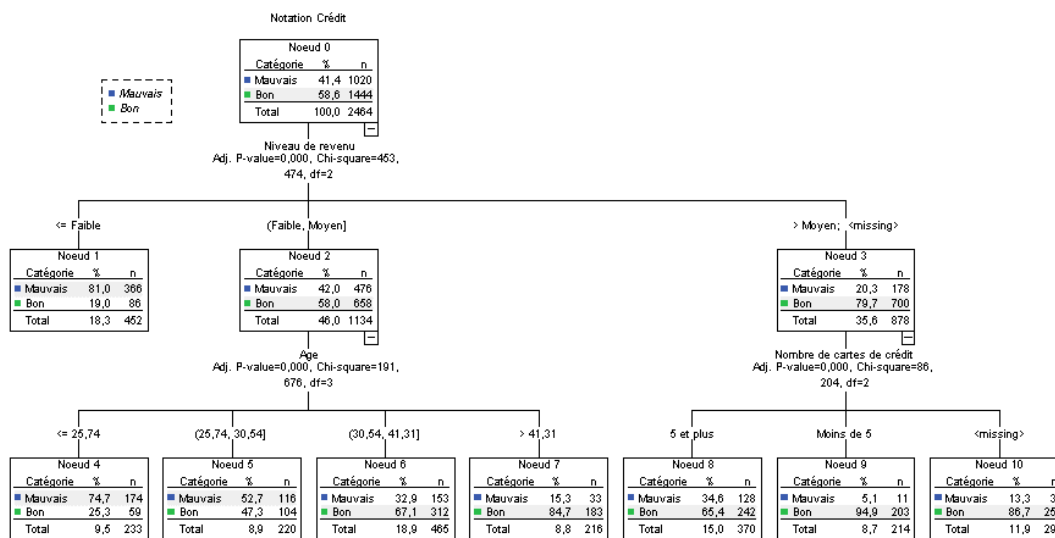
Figure 6-3
Boîte de dialogue Critères, onglet Limites de croissance



- Pour Nombre minimal d'observations, saisissez 400 pour Noeud parent et 200 pour Noeud enfant.
- Cliquez sur Continuer, puis sur OK pour lancer la procédure.

Résultats CHAID

Figure 6-4
Arbre CHAID avec valeurs de variable indépendante manquantes



Pour le noeud n°3, la valeur du *niveau de revenu* s'affiche de la manière suivante :

>Moyen;<manquant>. Cela signifie que le noeud contient des observations dans la modalité de revenus élevés et des observations avec valeurs manquantes pour le *niveau de revenu*.

Le noeud terminal n°10 contient des observations avec valeurs manquantes pour le *nombre de cartes de crédit*. Si vous cherchez à déterminer les bons risques de crédit, il s'agit du deuxième meilleur noeud terminal, ce qui risque d'être problématique si vous voulez utiliser ce modèle pour prévoir les bons risques de crédit. Vous ne voudrez certainement pas qu'un modèle prévoie une bonne cote de solvabilité simplement parce que vous ne savez pas de combien de cartes de crédit une observation dispose, et parce que certaines des observations ont des informations sur leur niveau de revenu manquantes.

Figure 6-5

Tableaux de classement et de risques pour le modèle CHAID

Risque	
Estimation	Erreur std.
,249	,009

Méthode de développement: CHAID

Variable dépendante: Notation Crédit

Observations	Prévisions		
	Mauvais	Bon	Pourcentage correct
Mauvais	656	364	64,3%
Bon	249	1195	82,8%
Pourcentage global	36,7%	63,3%	75,1%

Méthode de développement: CHAID

Variable dépendante: Notation Crédit

Les tableaux de classement et de risques indiquent que le modèle CHAID classe correctement environ 75 % des observations. Ce résultat n'est pas mauvais, mais il n'est pas suffisant. De plus, nous pouvons raisonnablement suspecter que le taux de classifications correctes pour les bonnes observations de crédit est trop optimiste, car il est en partie basé sur la supposition que le manque d'informations concernant deux variables indépendantes (*niveau de revenu* et *nombre de cartes de crédit*) est le signe de bonnes conditions de crédit.

Valeurs manquantes avec CRT

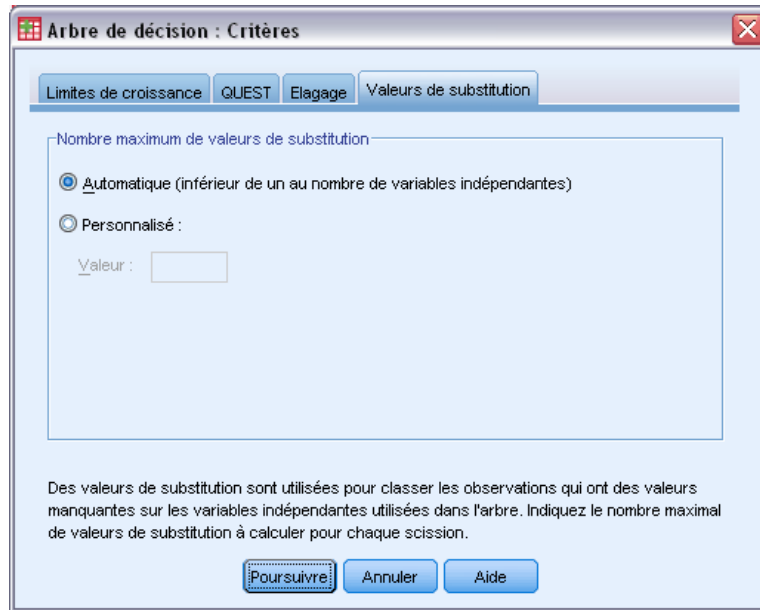
Nous allons déployer la même analyse de base, sauf que la méthode de croissance utilisée est CRT.

- ▶ Dans la boîte de dialogue principale Arbre de décision sur la méthode de croissance, sélectionnez CRT.
- ▶ Cliquez sur Critères.
- ▶ Vérifiez que le nombre minimum d'observations est toujours de 400 pour les noeuds parent et de 200 pour les noeuds enfant.

- Cliquez sur l'onglet Valeurs de substitution.

Remarque : L'onglet Valeurs de substitution ne s'affiche pas tant que vous n'avez pas sélectionné CRT ou QUEST comme méthode de croissance.

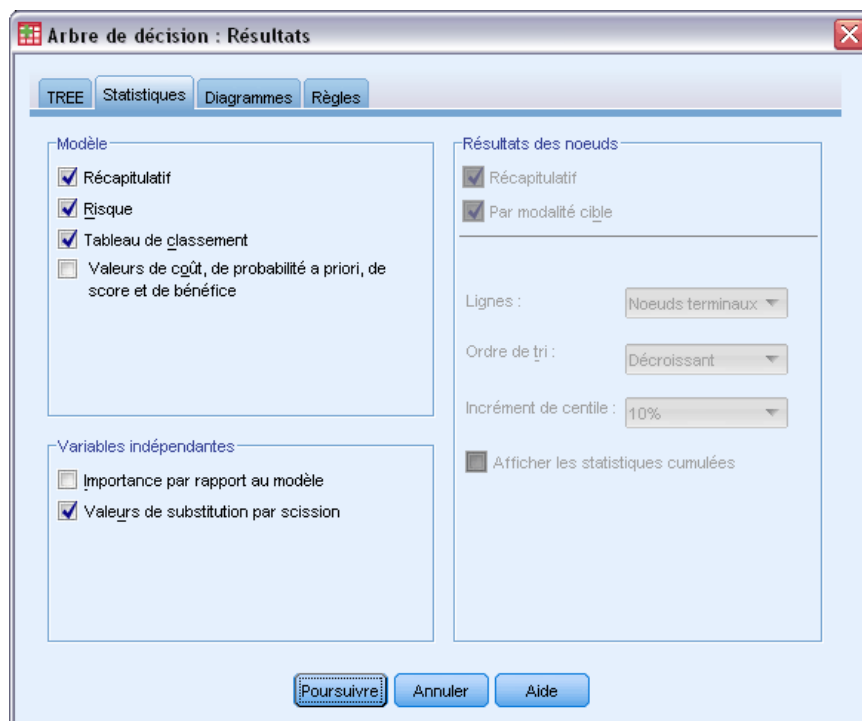
Figure 6-6
Boîte de dialogue Critères, onglet Valeurs de substitution



Pour chaque scission de noeud de variable indépendante, le paramètre Automatique considère toutes les autres variables indépendantes indiquées comme des valeurs de substitution possibles pour le modèle. Etant donné que cet exemple ne comporte pas beaucoup de variables indépendantes, le paramètre Automatique convient tout à fait.

- Cliquez sur Poursuivre.
- Dans la boîte de dialogue Arbre de décision principale, cliquez sur Résultat.

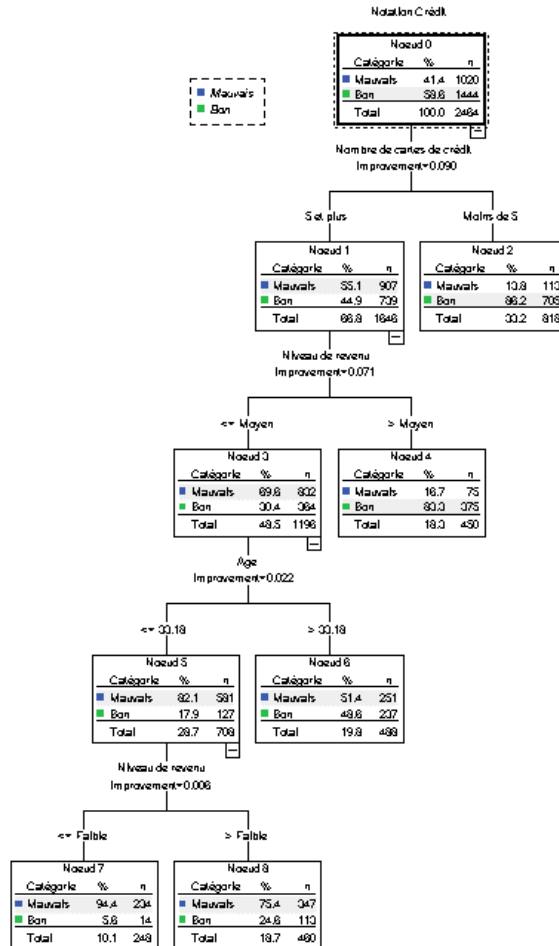
Figure 6-7
Boîte de dialogue Résultat, onglet Statistiques



- ▶ Cliquez sur l'onglet Statistiques.
- ▶ Sélectionnez Valeurs de substitution par division.
- ▶ Cliquez sur Continuer, puis sur OK pour lancer la procédure.

Résultats CRT

Figure 6-8
Arbre CRT avec valeurs de variable indépendante manquantes



Vous remarquerez immédiatement que cet arbre ne ressemble pas beaucoup à l'arbre CHAID. Mais cela n'est pas significatif en soi. Dans un modèle d'arbre CRT, toutes les scissions sont binaires, c'est-à-dire que chaque noeud parent est scindé en seulement deux noeuds enfant. Dans un modèle CHAID, les noeuds parent peuvent être scindés en de nombreux noeuds enfant. Ainsi, les arbres auront souvent une apparence différente bien qu'ils représentent le même modèle sous-jacent.

Il existe cependant un certain nombre de différences significatives :

- La variable indépendante (explicative) la plus importante dans le modèle CRT est le *nombre de cartes de crédit*, alors que dans le modèle CHAID, il s'agissait du *niveau de revenu*.
- Pour les observations comportant moins de cinq cartes de crédit, le *nombre de cartes de crédit* est la seule variable indépendante significative de la cote de solvabilité, et le noeud n°2 est un noeud terminal.
- Comme dans le modèle CHAID, le *niveau de revenu* et l'*âge* sont inclus dans le modèle, même si le *niveau de revenu* est désormais la deuxième variable indépendante, et non la première.

- Aucun noeud ne comporte de modalité <manquante> car la méthode CRT utilise des variables indépendantes de substitution plutôt que des valeurs manquantes dans le modèle.

Figure 6-9

Tableaux de classement et de risques pour le modèle CRT

Risque	
Estimation	Erreur std.
,224	,008

Méthode de développement: CRT
Variable dépendante: Notation Crédit

Observations	Prévisions		
	Mauvais	Bon	Pourcentage correct
Mauvais	832	188	81,6%
Bon	364	1080	74,8%
Pourcentage global	48,5%	51,5%	77,6%

Méthode de développement: CRT
Variable dépendante: Notation Crédit

- Les tableaux de risques et de classification montrent un taux de classifications correctes global d'environ 78 %, légèrement supérieur au modèle CHAID (75 %).
- Le taux de classifications correctes des observations de mauvais crédit est bien supérieur pour le modèle CRT (81,6 %) que pour le modèle CHAID (64,3 %).
- Le taux de classifications correctes des observations de bon crédit, lui, est passé de 82,8 % pour CHAID à 74,8 % pour CRT.

Valeurs de substitution

Les différences entre les modèles CHAID et CRT sont dues en partie à l'utilisation de valeurs de substitution dans le modèle CRT. Le tableau des valeurs de substitution indique comment les valeurs de substitution ont été utilisées dans le modèle.

Figure 6-10

Tableau des valeurs de substitution

Noeud parent	Variable indépendante :		Amélioration	Association
0	Principal	Nombre de cartes de crédit	,090	
	Surrogate	Crédit automobile	,052	,643
		Age	,001	,004
1	Principal	Niveau de revenu	,071	
	Surrogate	Age	,001	,004
3	Principal	Age	,022	
5	Principal	Niveau de revenu	,006	
	Surrogate	Age	3,93E-005	,009

Growing Method: CRT
Dependent Variable: Notation Crédit

- Au niveau du noeud racine (noeud 0), la meilleure variable indépendante (explicative) est le *nombre de cartes de crédit*.

- Pour toutes les observations avec valeurs manquantes pour le *nombre de cartes de crédit*, les *prêts auto* sont utilisés en tant que variable indépendante de substitution, puisque cette variable a un degré d'association relativement élevé (0,643) avec le *nombre de cartes de crédit*.
- Si une observation comporte également une valeur manquante pour les *prêts auto*, c'est l'*âge* qui est utilisé en tant que valeur de substitution (bien que cette variable n'ait qu'une valeur d'association de 0,004).
- L'*âge* est également la valeur de substitution du *niveau de revenu* pour les noeuds 1 et 5.

Récapitulatif

Les différentes méthodes de croissance n'ont pas la même manière de gérer les données manquantes. Si les données utilisées pour créer le modèle contiennent plusieurs valeurs manquantes ou si vous souhaitez appliquer ce modèle à d'autres fichiers de données comportant des valeurs manquantes, vous devez évaluer les effets des valeurs manquantes sur les différents modèles. Si vous souhaitez utiliser des valeurs de substitution dans le modèle pour compenser les valeurs manquantes, utilisez la méthode CRT ou QUEST.

Fichiers d'exemple

Les fichiers d'exemple installés avec le produit figurent dans le sous-répertoire *Echantillons* du répertoire d'installation. Il existe un dossier distinct au sein du sous-répertoire *Echantillons* pour chacune des langues suivantes : Anglais, Français, Allemand, Italien, Japonais, Coréen, Polonais, Russe, Chinois simplifié, Espagnol et Chinois traditionnel.

Seuls quelques fichiers d'exemples sont disponibles dans toutes les langues. Si un fichier d'exemple n'est pas disponible dans une langue, le dossier de langue contient la version anglaise du fichier d'exemple.

Descriptions

Voici de brèves descriptions des fichiers d'exemple utilisés dans divers exemples à travers la documentation.

- **accidents.sav.** Ce fichier de données d'hypothèse concerne une société d'assurance qui étudie les facteurs de risque liés à l'âge et au sexe dans les accidents de la route survenant dans une région donnée. Chaque observation correspond à une classification croisée de la catégorie d'âge et du sexe.
- **adl.sav.** Ce fichier de données d'hypothèse concerne les mesures entreprises pour identifier les avantages d'un type de thérapie proposé aux patients qui ont subi une attaque cardiaque. Les médecins ont assigné de manière aléatoire les patients du sexe féminin ayant subi une attaque cardiaque à un groupe parmi deux groupes possibles. Le premier groupe a fait l'objet de la thérapie standard tandis que le second a bénéficié en plus d'une thérapie émotionnelle. Trois mois après les traitements, les capacités de chaque patient à effectuer les tâches ordinaires de la vie quotidienne ont été notées en tant que variables ordinales.
- **advert.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- **aflatoxin.sav.** Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'alfatoxine en parties par milliard (PPB).
- **anorectic.sav.** En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.

- **bankloan.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondent aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- **bankloan_binning.sav.** Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- **behavior.sav.** Dans un exemple classique, on a demandé à 52 étudiants de noter les combinaisons établies à partir de 15 situations et de 15 comportements sur une échelle de 0 à 9, où 0 = « extrêmement approprié » et 9 = « extrêmement inapproprié ». En effectuant la moyenne des résultats de l'ensemble des individus, on constate une certaine différence entre les valeurs.
- **behavior_ini.sav.** Ce fichier de données contient la configuration initiale d'une solution bidimensionnelle pour *behavior.sav*.
- **brakes.sav.** Ce fichier de données d'hypothèse concerne le contrôle qualité effectué dans une usine qui fabrique des freins à disque pour des voitures haut de gamme. Le fichier de données contient les mesures de diamètre de 16 disques de 8 machines de production. Le diamètre cible des freins est de 322 millimètres.
- **breakfast.sav.** Au cours d'une étude classique, on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1 = « aliment préféré » à 15 = « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de « Préférence générale » à « En-cas avec boisson uniquement ».
- **breakfast-overall.sav.** Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, « Préférence générale ».
- **broadband_1.sav.** Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- **broadband_2.sav.** Ce fichier de données est identique au fichier *broadband_1.sav* mais contient les données relatives à trois mois supplémentaires.
- **car_insurance_claims.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modélisé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- **car_sales.sav.** Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de *edmunds.com* et des sites des constructeurs.
- **car_sales_uprepared.sav.** Il s'agit d'une version modifiée de *car_sales.sav* qui n'inclut aucune version transformée des champs.

- **carpet.sav.** Dans un exemple courant, une société intéressée par la commercialisation d'un nouveau nettoyeur de tapis souhaite examiner l'influence de cinq critères sur la préférence du consommateur : la conception du conditionnement, la marque, le prix, une étiquette *Economique* et une garantie satisfait ou remboursé. Il existe trois niveaux de critère pour la conception du conditionnement, suivant l'emplacement de l'applicateur, trois marques (*K2R*, *Glory* et *Bissell*), trois niveaux de prix et deux niveaux (non ou oui) pour chacun des deux derniers critères. Dix consommateurs classent 22 profils définis par ces critères. La variable *Préférence* indique le classement des rangs moyens de chaque profil. Un rang faible correspond à une préférence élevée. Cette variable reflète une mesure globale de préférence pour chaque profil.
- **carpet_prefs.sav.** Ce fichier de données repose sur le même exemple que celui décrit pour *carpet.sav*, mais contient les classements réels issus de chacun des 10 clients. On a demandé aux consommateurs de classer les 22 profils de produits, du préféré au moins intéressant. Les variables *PREF1* à *PREF22* contiennent les identificateurs des profils associés, tels qu'ils sont définis dans *carpet_plan.sav*.
- **catalog.sav.** Ce fichier de données contient des chiffres de ventes mensuelles hypothétiques relatifs à trois produits vendus par une entreprise de vente par correspondance. Les données relatives à cinq variables explicatives possibles sont également incluses.
- **catalog_seasfac.sav.** Ce fichier de données est identique à *catalog.sav* mais contient en plus un ensemble de facteurs saisonniers calculés à partir de la procédure de désaisonnalisation, ainsi que les variables de date correspondantes.
- **cellular.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un opérateur téléphonique pour réduire les taux de désabonnement. Des scores de propension au désabonnement sont attribués aux comptes, de 0 à 100. Les comptes ayant une note égale ou supérieure à 50 sont susceptibles de changer de fournisseur.
- **ceramics.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fabricant pour déterminer si un nouvel alliage haute qualité résiste mieux à la chaleur qu'un alliage standard. Chaque observation représente un test séparé de l'un des deux alliages ; le degré de chaleur auquel l'alliage ne résiste pas est enregistré.
- **cereal.sav.** Ce fichier de données d'hypothèse concerne un sondage de 880 personnes interrogées sur leurs préférences de petit-déjeuner et sur leur âge, leur sexe, leur situation familiale et leur mode de vie (actif ou non actif, selon qu'elles pratiquent une activité physique au moins deux fois par semaine). Chaque observation correspond à un répondant distinct.
- **clothing_defects.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de textile. Dans chaque lot produit à l'usine, les inspecteurs prélèvent un échantillon de vêtements et comptent le nombre de vêtements qui ne sont pas acceptables.
- **coffee.sav.** Ce fichier de données concerne l'image perçue de six marques de café frappé. Pour chacun des 23 attributs d'image de café frappé, les personnes sollicitées ont sélectionné toutes les marques décrites par l'attribut. Les six marques sont appelées AA, BB, CC, DD, EE et FF à des fins de confidentialité.
- **contacts.sav.** Ce fichier de données d'hypothèse concerne les listes de contacts d'un groupe de représentants en informatique d'entreprise. Chaque contact est classé selon le service de l'entreprise où il travaille et le classement de son entreprise. Sont également enregistrés le

montant de la dernière vente effectuée, le temps passé depuis la dernière vente et la taille de l'entreprise du contact.

- **creditpromo.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un grand magasin pour évaluer l'efficacité d'une promotion récente de carte de crédit. A cette fin, 500 détenteurs de carte ont été sélectionnés au hasard. La moitié a reçu une publicité faisant la promotion d'un taux d'intérêt réduit sur les achats effectués dans les trois mois à venir. L'autre moitié a reçu une publicité saisonnière standard.
- **customer_dbase.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour utiliser les informations figurant dans sa banque de données et proposer des offres spéciales aux clients susceptibles d'être intéressés. Un sous-groupe de la base de clients a été sélectionné au hasard et a reçu des offres spéciales. Les réponses des clients ont été enregistrées.
- **customer_information.sav.** Un fichier de données d'hypothèse qui contient les informations postales du client, telles que le nom et l'adresse.
- **customer_subset.sav.** Un sous-ensemble de 80 observations de *customer_dbase.sav*.
- **debate.sav.** Ce fichier de données d'hypothèse concerne des réponses appariées à une enquête donnée aux participants à un débat politique avant et après le débat. Chaque observation représente un répondant distinct.
- **debate_aggregate.sav.** Il s'agit d'un fichier de données d'hypothèse qui rassemble les réponses dans le fichier *debate.sav*. Chaque observation correspond à une classification croisée de préférence avant et après le débat.
- **demo.sav.** Ce fichier de données d'hypothèse concerne une base de données clients achetée en vue de diffuser des offres mensuelles. Les données indiquent si le client a répondu ou non à l'offre et contiennent diverses informations démographiques.
- **demo_cs_1.sav.** Ce fichier de données d'hypothèse concerne la première mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à une ville différente. La région, la province, le quartier et la ville sont enregistrés.
- **demo_cs_2.sav.** Ce fichier de données d'hypothèse concerne la seconde mesure entreprise par une société pour compiler une base de données contenant des informations d'enquête. Chaque observation correspond à un ménage différent issu des villes sélectionnées à la première étape. La région, la province, le quartier, la ville, la sous-division et l'identification sont enregistrés. Les informations d'échantillonnage des deux premières étapes de la conception sont également incluses.
- **demo_cs.sav.** Ce fichier de données d'hypothèse concerne des informations d'enquête collectées via une méthode complexe d'échantillonnage. Chaque observation correspond à un ménage différent et diverses informations géographiques et d'échantillonnage sont enregistrées.
- **dmdata.sav.** Ceci est un fichier de données d'hypothèse qui contient des informations démographiques et des informations concernant les achats pour une entreprise de marketing direct. *dmdata2.sav* contient les informations pour un sous-ensemble de contacts qui ont reçu un envoi d'essai, et *dmdata3.sav* contient des informations sur les contacts restants qui n'ont pas reçu l'envoi d'essai.

- **dietstudy.sav.** Ce fichier de données d'hypothèse contient les résultats d'une étude portant sur le régime de Stillman. Chaque observation correspond à un sujet distinct et enregistre son poids en livres avant et après le régime, ainsi que ses niveaux de triglycérides en mg/100 ml.
- **dvdplayer.sav.** Ce fichier de données d'hypothèse concerne le développement d'un nouveau lecteur DVD. A l'aide d'un prototype, l'équipe de marketing a collecté des données de groupes spécifiques. Chaque observation correspond à un utilisateur interrogé et enregistre des informations démographiques sur cet utilisateur, ainsi que ses réponses aux questions portant sur le prototype.
- **german_credit.sav.** Ce fichier de données provient de l'ensemble de données « German credit » figurant dans le référentiel Machine Learning Databases de l'université de Californie, Irvine.
- **grocery_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *grocery_coupons.sav* dans lequel les achats hebdomadaires sont organisés par client distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, le montant dépensé enregistré est à présent la somme des montants dépensés au cours des quatre semaines de l'enquête.
- **grocery_coupons.sav.** Il s'agit d'un fichier de données d'hypothèse qui contient des données d'enquête collectées par une chaîne de magasins d'alimentation qui cherchent à déterminer les habitudes de consommation de ses clients. Chaque client est suivi pendant quatre semaines et chaque observation correspond à une semaine distincte. Les informations enregistrées concernent les endroits où le client effectue ses achats, la manière dont il les effectue, ainsi que les sommes dépensées en provisions au cours de cette semaine.
- **guttman.sav.** Bell a présenté un tableau pour illustrer les groupes sociaux possibles. Guttman a utilisé une partie de ce tableau, dans lequel cinq variables décrivant des éléments tels que l'interaction sociale, le sentiment d'appartenance à un groupe, la proximité physique des membres et la formalité de la relation, ont été croisées avec sept groupes sociaux théoriques, dont les foules (par exemple, le public d'un match de football), l'audience (par exemple, au cinéma ou dans une salle de classe), le public (par exemple, les journaux ou la télévision), les bandes (proche d'une foule, mais qui serait caractérisée par une interaction beaucoup plus intense), les groupes primaires (intimes), les groupes secondaires (volontaires) et la communauté moderne (groupement lâche issu d'une forte proximité physique et d'un besoin de services spécialisés).
- **health_funding.sav.** Ce fichier de données d'hypothèse concerne des données sur le financement des soins de santé (montant par groupe de 100 individus), les taux de maladie (taux par groupe de 10 000 individus) et les visites chez les prestataires de soins de santé (taux par groupe de 10 000 individus). Chaque observation représente une ville différente.
- **hivassay.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un laboratoire pharmaceutique pour développer une analyse rapide de détection d'infection HIV. L'analyse a pour résultat huit nuances de rouge, les nuances les plus marquées indiquant une plus forte probabilité d'infection. Un test en laboratoire a été effectué sur 2 000 échantillons de sang, la moitié de ces échantillons étant infectée par le virus HIV et l'autre moitié étant saine.
- **hourlywagedata.sav.** Ce fichier de données d'hypothèse concerne les salaires horaires d'infirmières occupant des postes administratifs et dans les services de soins, et affichant divers niveaux d'expérience.

- **insurance_claims.sav.** Il s'agit d'un fichier de données hypothétiques qui concerne une compagnie d'assurance souhaitant développer un modèle pour signaler des réclamations suspectes, potentiellement frauduleuses. Chaque observation correspond à une réclamation distincte.
- **insure.sav.** Ce fichier de données d'hypothèse concerne une compagnie d'assurance qui étudie les facteurs de risque indiquant si un client sera amené à déclarer un incident au cours d'un contrat d'assurance vie d'une durée de 10 ans. Chaque observation figurant dans le fichier de données représente deux contrats, l'un ayant enregistré une réclamation et l'autre non, appariés par âge et sexe.
- **judges.sav.** Ce fichier de données d'hypothèse concerne les scores attribués par des juges expérimentés (plus un juge enthousiaste) à 300 performances de gymnastique. Chaque ligne représente une performance distincte ; les juges ont examiné les mêmes performances.
- **kinship_dat.sav.** Rosenberg et Kim se sont lancés dans l'analyse de 15 termes de parenté (cousin/cousine, fille, fils, frère, grand-mère, grand-père, mère, neveu, nièce, oncle, père, petite-fille, petit-fils, sœur, tante). Ils ont demandé à quatre groupes d'étudiants (deux groupes de femmes et deux groupes d'hommes) de trier ces termes en fonction des similarités. Deux groupes (un groupe de femmes et un groupe d'hommes) ont été invités à effectuer deux tris, en basant le second sur un autre critère que le premier. Ainsi, un total de six "sources" a été obtenu. Chaque source correspond à une matrice de proximité 15×15 , dont le nombre de cellules est égal au nombre de personnes dans une source moins le nombre de fois où les objets ont été partitionnés dans cette source.
- **kinship_ini.sav.** Ce fichier de données contient une configuration initiale d'une solution tridimensionnelle pour *kinship_dat.sav*.
- **kinship_var.sav.** Ce fichier de données contient les variables indépendantes *sexe*, *génér(ation)* et *degré* (de séparation) permettant d'interpréter les dimensions d'une solution pour *kinship_dat.sav*. Elles permettent en particulier de réduire l'espace de la solution à une combinaison linéaire de ces variables.
- **marketvalues.sav.** Ce fichier de données concerne les ventes de maisons dans un nouvel ensemble à Algonquin (Illinois) au cours des années 1999–2000. Ces ventes relèvent des archives publiques.
- **nhis2000_subset.sav.** Le NHIS (National Health Interview Survey) est une enquête de grande envergure concernant la population des États-Unis. Des entretiens ont lieu avec un échantillon de ménages représentatifs de la population américaine. Des informations démographiques et des observations sur l'état de santé et le comportement sanitaire sont recueillies auprès des membres de chaque ménage. Ce fichier de données contient un sous-groupe d'informations issues de l'enquête de 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Fichier de données et documentation d'usage public. ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/. Accès en 2003.
- **ozone.sav.** Les données incluent 330 observations portant sur six variables météorologiques pour prévoir la concentration d'ozone à partir des variables restantes. Des chercheurs précédents , , ont décelé parmi ces variables des non-linéarités qui pénalisent les approches standard de la régression.

- **pain_medication.sav.** Ce fichier de données d'hypothèse contient les résultats d'un essai clinique d'un remède anti-inflammatoire traitant les douleurs de l'arthrite chronique. On cherche notamment à déterminer le temps nécessaire au médicament pour agir et les résultats qu'il permet d'obtenir par rapport à un médicament existant.
- **patient_los.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux de patients admis à l'hôpital pour suspicion d'infarctus du myocarde suspecté (ou « attaque cardiaque »). Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **patlos_sample.sav.** Ce fichier de données d'hypothèse contient les dossiers médicaux d'un échantillon de patients sous traitement thrombolytique après un infarctus du myocarde. Chaque observation correspond à un patient distinct et enregistre de nombreuses variables liées à son séjour à l'hôpital.
- **poll_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un enquêteur pour déterminer le niveau de soutien du public pour un projet de loi avant législature. Les observations correspondent à des électeurs enregistrés. Chaque observation enregistre le comté, la ville et le quartier où habite l'électeur.
- **poll_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des électeurs répertoriés dans le fichier *poll_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *poll_csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. Toutefois, ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS – Probability-Proportional-to-Size), il existe également un fichier contenant les probabilités de sélection conjointes (*poll_jointprob.sav*). Les variables supplémentaires correspondant à la répartition démographique des électeurs et à leur opinion sur le projet de loi proposé ont été collectées et ajoutées au fichier de données une fois l'échantillon prélevé.
- **property_assess.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur au niveau du comté pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés vendues dans le comté au cours de l'année précédente. Chaque observation du fichier de données enregistre la ville où se trouve la propriété, l'évaluateur ayant visité la propriété pour la dernière fois, le temps écoulé depuis cette évaluation, l'évaluation effectuée à ce moment-là et la valeur de vente de la propriété.
- **property_assess_cs.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un contrôleur du gouvernement pour maintenir les évaluations de valeur de propriété à jour sur des ressources limitées. Les observations correspondent à des propriétés de l'état. Chaque observation du fichier de données enregistre le comté, la ville et le quartier où se trouve la propriété, le temps écoulé depuis la dernière évaluation et l'évaluation alors effectuée.
- **property_assess_cs_sample.sav.** Ce fichier de données d'hypothèse contient un échantillon des propriétés répertoriées dans le fichier *property_assess_cs.sav*. L'échantillon a été prélevé selon le plan spécifié dans le fichier de plan *property_assess_csplan* et ce fichier de données enregistre les probabilités d'inclusion et les pondérations d'échantillon. La variable supplémentaire *Valeur courante* a été collectée et ajoutée au fichier de données une fois l'échantillon prélevé.

- **recidivism.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis, ainsi que le temps écoulé jusqu'à la seconde arrestation si elle s'est produite dans les deux années suivant la première.
- **recidivism_cs_sample.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une agence administrative d'application de la loi pour interpréter les taux de récidive dans la juridiction. Chaque observation correspond à un récidiviste libéré suite à la première arrestation en juin 2003 et enregistre les informations démographiques qui lui sont propres, certains détails sur le premier délit commis et les données relatives à la seconde arrestation, si elle a eu lieu avant fin juin 2006. Les récidivistes ont été choisis dans plusieurs départements échantillonnés conformément au plan d'échantillonnage spécifié dans *recidivism_cs.csplan*. Ce plan faisant appel à une méthode d'échantillonnage de probabilité proportionnelle à la taille (PPS - Probability proportional to size), il existe également un fichier contenant les probabilités de sélection conjointes (*recidivism_cs_jointprob.sav*).
- **rfm_transactions.sav.** Un fichier de données d'hypothèse qui contient les données de transaction d'achat, y compris la date d'achat, le/les élément(s) acheté(s) et le montant monétaire pour chaque transaction.
- **salesperformance.sav.** Ce fichier de données d'hypothèse concerne l'évaluation de deux nouveaux cours de formation en vente. Soixante employés, divisés en trois groupes, reçoivent chacun une formation standard. En outre, le groupe 2 suit une formation technique et le groupe 3 un didacticiel pratique. À l'issue du cours de formation, chaque employé est testé et sa note enregistrée. Chaque observation du fichier de données représente un stagiaire distinct et enregistre le groupe auquel il a été assigné et la note qu'il a obtenue au test.
- **satisf.sav.** Il s'agit d'un fichier de données d'hypothèse portant sur une enquête de satisfaction effectuée par une société de vente au détail au niveau de quatre magasins. Un total de 582 clients ont été interrogés et chaque observation représente la réponse d'un seul client.
- **screws.sav.** Ce fichier de données contient des informations sur les descriptives des vis, des boulons, des écrous et des clous..
- **shampoo_ph.sav.** Ce fichier de données d'hypothèse concerne le processus de contrôle qualité observé dans une usine de produits capillaires. À intervalles réguliers, six lots de sortie distincts sont mesurés et leur pH enregistré. La plage cible est 4,5–5,5.
- **ships.sav.** Il s'agit d'un ensemble de données présenté et analysé ailleurs et concernant les dommages causés à des cargos par les vagues. Les effectifs d'incidents peuvent être modélisés comme des incidents se produisant selon un taux de Poisson en fonction du type de navire, de la période de construction et de la période de service. Les mois de service totalisés pour chaque cellule du tableau formé par la classification croisée des facteurs fournissent les valeurs d'exposition au risque.
- **site.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société pour choisir de nouveaux sites pour le développement de ses activités. L'entreprise a fait appel à deux consultants pour évaluer séparément les sites. Ces consultants, en plus de fournir un rapport approfondi, ont classé chaque site comme constituant une éventualité « bonne », « moyenne » ou « faible ».

- **smokers.sav.** Ce fichier de données est extrait de l'étude National Household Survey of Drug Abuse de 1998 et constitue un échantillon de probabilité des ménages américains. (<http://dx.doi.org/10.3886/ICPSR02934>) Ainsi, la première étape dans l'analyse de ce fichier doit consister à pondérer les données pour refléter les tendances de population.
- **stocks.sav** Ce fichier de données hypothétiques contient le cours et le volume des actions pour un an.
- **stroke_clean.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois celle-ci purgée via des procédures de l'option Validation de données.
- **stroke_invalid.sav.** Ce fichier de données d'hypothèse concerne l'état initial d'une base de données médicales et comporte plusieurs erreurs de saisie de données.
- **stroke_survival.** Ce fichier de données d'hypothèse concerne les temps de survie de patients qui quittent un programme de rééducation à la suite d'un accident ischémique et rencontrent un certain nombre de problèmes. Après l'attaque, l'occurrence d'infarctus du myocarde, d'accidents ischémiques ou hémorragiques est signalée, et le moment de l'événement enregistré. L'échantillon est tronqué à gauche car il n'inclut que les patients ayant survécu durant le programme de rééducation mis en place suite à une attaque.
- **stroke_valid.sav.** Ce fichier de données d'hypothèse concerne l'état d'une base de données médicales une fois les valeurs vérifiées via la procédure Validation de données. Elle contient encore des observations anormales potentielles.
- **survey_sample.sav.** Ce fichier de données concerne des informations d'enquête dont des données démographiques et des mesures comportementales. Il est basé sur un sous-ensemble de variables de la 1998 NORC General Social Survey, bien que certaines valeurs de données aient été modifiées et que des variables supplémentaires fictives aient été ajoutées à titre de démonstration.
- **telco.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une société de télécommunications pour réduire les taux de désabonnement de sa base de clients. Chaque observation correspond à un client distinct et enregistre diverses informations démographiques et d'utilisation de service.
- **telco_extra.sav.** Ce fichier de données est semblable au fichier de données *telco.sav* mais les variables de permanence et de dépenses des consommateurs transformées log ont été supprimées et remplacées par des variables de dépenses des consommateurs transformées log standardisées.
- **telco_missing.sav.** Ce fichier de données est un sous-ensemble du fichier de données *telco.sav* mais certaines des valeurs de données démographiques ont été remplacées par des valeurs manquantes.
- **testmarket.sav.** Ce fichier de données d'hypothèse concerne une chaîne de fast foods et ses plans marketing visant à ajouter un nouveau plat à son menu. Trois campagnes étant possibles pour promouvoir le nouveau produit, le nouveau plat est introduit sur des sites sur plusieurs marchés sélectionnés au hasard. Une promotion différente est effectuée sur chaque site et les ventes hebdomadaires du nouveau plat sont enregistrées pour les quatre premières semaines. Chaque observation correspond à un site-semaine distinct.
- **testmarket_1month.sav.** Ce fichier de données d'hypothèse est le fichier de données *testmarket.sav* dans lequel les ventes hebdomadaires sont organisées par site distinct. Certaines variables qui changeaient toutes les semaines disparaissent. En outre, les ventes

enregistrées sont à présent la somme des ventes réalisées au cours des quatre semaines de l'enquête.

- **tree_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_credit.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire.
- **tree_missing_data.sav** Ce fichier de données d'hypothèse concerne des données démographiques et d'historique de prêt bancaire avec un grand nombre de valeurs manquantes.
- **tree_score_car.sav.** Ce fichier de données d'hypothèse concerne des données démographiques et de prix d'achat de véhicule.
- **tree_textdata.sav.** Ce fichier de données simples ne comporte que deux variables et vise essentiellement à indiquer l'état par défaut des variables avant affectation du niveau de mesure et des étiquettes de valeurs.
- **tv-survey.sav.** Ce fichier de données d'hypothèse concerne une enquête menée par un studio de télévision qui envisage de prolonger la diffusion d'un programme ou de l'arrêter. On a demandé à 906 personnes si elles regarderaient le programme dans diverses situations. Chaque ligne représente un répondant distinct et chaque colonne une situation distincte.
- **ulcer_recurrence.sav.** Ce fichier contient des informations partielles d'une enquête visant à comparer l'efficacité de deux thérapies de prévention de la récurrence des ulcères. Il fournit un bon exemple de données censurées par intervalle et a été présenté et analysé ailleurs .
- **ulcer_recurrence_recoded.sav.** Ce fichier réorganise les informations figurant dans le fichier *ulcer_recurrence.sav* pour que vous puissiez modéliser la probabilité d'événement pour chaque intervalle de l'enquête plutôt que la probabilité d'événement de fin d'enquête. Il a été présenté et analysé ailleurs .
- **verd1985.sav.** Ce fichier de données concerne une enquête . Les réponses de 15 sujets à 8 variables ont été enregistrées. Les variables présentant un intérêt sont divisées en trois ensembles. Le groupe 1 comprend l'âge et la *situation familiale*, le groupe 2 les *animaux domestiques* et la *presse*, et le groupe 3 la *musique* et l'*habitat*. A la variable *animal domestique* est appliqué un codage nominal multiple et à *âge*, un codage ordinal ; toutes les autres variables ont un codage nominal simple.
- **virus.sav.** Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un fournisseur de services Internet pour déterminer les effets d'un virus sur ses réseaux. Il a suivi le pourcentage (approximatif) de trafic de messages électroniques infectés par un virus sur ses réseaux sur la durée, de la découverte à la circonscription de la menace.
- **wheeze_steubenville.sav.** Il s'agit d'un sous-ensemble d'une enquête longitudinale des effets de la pollution de l'air sur la santé des enfants . Les données contiennent des mesures binaires répétées de l'état asthmatique d'enfants de la ville de Steubenville (Ohio), âgés de 7, 8, 9 et 10 ans, et indiquent si la mère fumait au cours de la première année de l'enquête.
- **workprog.sav.** Ce fichier de données d'hypothèse concerne un programme de l'administration visant à proposer de meilleurs postes aux personnes défavorisées. Un échantillon de participants potentiels au programme a ensuite été prélevé. Certains de ces participants ont

été sélectionnés au hasard pour participer au programme. Chaque observation représente un participant au programme distinct.

- **worldsales.sav** Ce fichier de données hypothétiques contient les revenus des ventes par continent et par produit.

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'ETAT » 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.

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.

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>.

Adobe, le logo Adobe, PostScript et le logo PostScript sont des marques déposées ou des marques commerciales de Adobe Systems Incorporated aux États-Unis et/ou dans d'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 commerciales ou des marques déposées de Intel Corporation ou de ses filiales aux États-Unis et dans d'autres pays.

Java et toutes les marques et logos Java sont des marques commerciales de Sun Microsystems, Inc. aux États-Unis et/ou 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 États-Unis et/ou dans d'autres pays.

UNIX est une marque déposée de The Open Group aux États-Unis et dans d'autres pays.

Ce produit utilise WinWrap Basic, Copyright 1993-2007, Polar Engineering and Consulting, <http://www.winwrap.com/>.

Les autres noms de produits et de services peuvent être des marques d'IBM ou d'autres sociétés.

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.



Index

- arbres, 1
 - affichage et masquage des statistiques de branche, 26
 - application de modèles, 84
 - arbre sous forme de tableau, 70
 - attribut de texte, 46
 - bénéfices, 18
 - carte d'arbre, 43
 - contenu des arbres dans un tableau, 26
 - contrôle de la taille de noeud, 9
 - contrôle de l'affichage des arbres, 26, 45
 - Couleurs, 46
 - couleurs des diagrammes de noeud, 46
 - coûts de classification erronée, 17
 - coûts personnalisés, 79
 - Critères de croissance CHAID, 10
 - diagrammes, 32
 - effets des étiquettes de valeur, 57
 - effets du niveau de mesure, 53
 - élagage, 15
 - Enregistrement de prévisions, 75
 - enregistrement des variables du modèle, 24
 - estimations du risque, 28
 - estimations du risque pour les variables d'échelle dépendantes, 88
 - génération des règles, 38, 49
 - importance des valeurs prédites, 28
 - intervalles des variables d'échelle indépendantes, 12
 - limitation du nombre de niveaux, 9
 - manipulation de grands arbres, 42
 - masquage de branches et de noeuds, 41
 - Méthode CRT, 13
 - mise à l'échelle de l'affichage de l'arbre, 44
 - modification, 41
 - notation, 84
 - orientation de l'arbre, 26
 - Polices, 46
 - Probabilité a priori, 20
 - scores, 21
 - sélection de plusieurs noeuds, 41
 - sélection d'observations dans les noeuds, 76
 - statistiques des noeuds terminaux, 28
 - tableau de gains pour les noeuds, 72
 - tableau des mauvaises réaffectations, 28
 - tableau récapitulatif des modèles, 68
 - valeurs de substitution, 93, 100
 - valeurs d'index, 28
 - Valeurs manquantes, 23, 93
 - validation croisée, 8
 - validation par partition, 8
 - variables d'échelle dépendantes, 84
- arbres de décision, 1
 - introduction forcée de la première variable dans le modèle, 1
 - Méthode CHAID, 1
 - Méthode CRT, 1
 - Méthode Exhaustive CHAID, 1
- Méthode QUEST, 1, 14
 - niveau de mesure, 1
- bénéfices
 - arbres, 18, 28
 - Probabilité a priori, 20
- CHAID, 1
 - ajustement de Bonferroni, 10
 - critères de scission et de fusion, 10
 - intervalles des variables d'échelle indépendantes, 12
 - nombre maximum d'itérations, 10
 - scission des modalités fusionnées, 10
- classification erronée
 - arbres, 28
 - coûts, 17
 - taux, 74
- coûts
 - classification erronée, 17
 - modèles d'arbre, 79
- CRT, 1
 - élagage, 15
 - mesures d'impureté, 13
- diagramme des gains, 73
- diagramme des index, 73
- élagage d'arbres décision
 - et masquage des noeuds, 15
- estimations du risque
 - arbres, 28
 - pour les variables dépendantes dans la procédure Arbre de décision, 88
 - variables dépendantes qualitatives, 74
- Etiquettes de valeurs
 - arbres, 57
- fichiers d'exemple
 - emplacement, 103
 - fusion de branches d'arbre, 41
- gain, 72
- Génération de nombres aléatoires
 - validation d'arbre de décision, 8
- Gini, 13
- impureté
 - Arbres CRT, 13
- index
 - modèles d'arbre, 72

- marques commerciales, 115
- masquage de branches d'arbre, 41
- masquage des noeuds
 - et élagage, 15
- mentions légales, 114
- modèles d'arbre, 72

- niveau de mesure
 - arbres de décision, 1
 - dans les modèles d'arbre, 53
- niveau de signification pour scinder les noeuds, 14
- noeuds
 - sélection de plusieurs noeuds d'arbre, 41
- nombre de noeuds
 - enregistrement en tant que variable à partir des arbres de décision, 24
- notation
 - modèles d'arbre, 84

- Pondération d'observations
 - pondérations fractionnelles dans les arbres de décision, 1
- Prévisions
 - enregistrement en tant que variable à partir des arbres de décision, 24
 - enregistrement pour les modèles d'arbre, 75
- probabilité prédite
 - enregistrement en tant que variable à partir des arbres de décision, 24

- QUEST, 1, 14
 - élagage, 15

- règles
 - création d'une syntaxe de sélection et d'analyse pour les arbres de décision, 38, 49
- réponse
 - modèles d'arbre, 72

- scores
 - arbres, 21
- sélection de plusieurs noeuds d'arbre, 41
- SQL
 - création de la syntaxe SQL pour la sélection et l'analyse, 38, 49
- Syntaxe
 - création d'une syntaxe de sélection et d'analyse pour les arbres de décision, 38, 49
- Syntaxe de commande
 - création d'une syntaxe de sélection et d'analyse pour les arbres de décision, 38, 49

- tableau de classement, 74
- tableau récapitulatif des modèles
 - modèles d'arbre, 68

- twoing, 13
- twoing ordonné, 13

- valeurs de substitution
 - dans les modèles d'arbre, 93, 100
- valeurs d'index
 - arbres, 28
- Valeurs manquantes
 - arbres, 23
 - dans les modèles d'arbre, 93
- validation
 - arbres, 8
- validation croisée
 - arbres, 8
- validation par partition
 - arbres, 8
- Variables d'échelle
 - variables dépendantes dans la procédure Arbre de décision, 84