Objectif

Montrer l'utilisation du logiciel d'apprentissage d'arbres de décision, XL-SIPINA, qui s'appuie sur le tableur EXCEL© pour la gestion des données.

Une large partie du processus d'extraction de connaissances repose sur la préparation et la transformation des données. Un système de gestion de données efficace est par conséquent la base incontournable d'un bon logiciel de Data Mining. C'est ce qui fait la différence entre les logiciels académiques, tournées vers les chercheurs, et les logiciels commerciaux, qui doivent être accessibles aux praticiens, compétents sur le plan des méthodes et des approches statistiques, mais réfractaires à la programmation et autres traitements informatiques de bas niveau.

Pour le chercheur, l'implémentation des méthodes a un côté passionnant, elle permet la compréhension détaillée des approches. En revanche, la programmation de la gestion de données l'intéresse rarement. Elle est fastidieuse, les bugs guettent chaque ligne de code, et surtout, elle est peu valorisante (valorisée) sur le plan académique. C'est vraisemblablement pour cette raison que les logiciels gratuits élaborés par les chercheurs et distribués librement sur le web disposent rarement d'outils sophistiqués pour l'appréhension des données, de formats et d'origines diverses.

Une solution simple serait d'implémenter directement les méthodes au sein d'outils performants d'accès et de gestion de données. Certains éditeurs de logiciels ont d'ailleurs fait ce choix, les modules de Data Mining sont directement adossés à un SGBD (IBM, Microsoft, etc.).

Sans aller jusqu'à ce stade, nous disposons dans notre vie (informatique) courante d'un outil de manipulation de données simple, accessible à tous et réellement performant : le tableur. L'outil « Tableau Croisé Dynamique », par exemple, est une merveille qui réalise avec très peu de manipulations des statistiques descriptives, des tris à plats et croisés sur plusieurs niveaux. Il permet de résumer et de mettre en évidence rapidement les interactions entre les variables. L'idéal serait de compléter les outils déjà disponibles dans un tableur par des techniques plus sophistiquées. Il existe d'ailleurs des outils commerciaux qui fonctionnent sur ce principe (XLSTAT, STATBOX, etc.).

Dans ce didacticiel, nous présentons une de nos anciennes tentatives d'intégrer directement dans Excel des modules de Data Mining, notamment l'induction des arbres de décision. L'idée n'est pas de programmer des modules « add-ons » que l'on associe au tableur, mais plutôt de produire une application qui intègre le tableur comme une des ses parties. L'implémentation s'appuie essentiellement sur la technologie OLE de WINDOWS. Nous bénéficions des avantages de cette technologie, la simplicité d'implémentation, et de ses inconvénients, principalement l'instabilité et la lenteur, que nous avons tenté de minimiser en bufférisant au maximum tous les échanges. Dans les faits, la feuille de calcul EXCEL n'est lue et parsée qu'une seule fois, au démarrage du module de construction d'arbre. Nous travaillons directement en mémoire par la suite, en utilisant nos propres bibliothèques de calcul, à savoir celles de la version recherche de SIPINA.

Notons 2 points importants :

1. **Les fonctionnalités d'exploration des données à l'aide des arbres de décision que nous présentons dans ce didacticiel sont également présentes dans la version stand-alone (Version Recherche) de SIPINA accessible sur notre site web.**

2. **Cette version spécifique XL-SIPINA ne peut fonctionner, elle, que si le tableur EXCEL est présent sur l'ordinateur.**

En exhumant ce vieux projet oublié au fond d'un tiroir, en l'occurrence d'un répertoire de sauvegarde d'un vieux CD retrouvé lors d'un déménagement, il s'agit tout simplement de montrer la faisabilité d'une telle approche dans l'implémentation des outils de Data Mining. Nous nous heurtons quand même très vite à des problèmes de performances et de stabilité. Peut être aussi, il faut le reconnaître, qu'à cause d'une maîtrise quelque peu approximative de la technologie OLE, il se peut très bien que nous n'en utilisons pas de manière optimale toutes les possibilités. Dans bien des cas, nous nous en excusons à l'avance, lorsque l'application semble se figer, la seule solution pour s'en sortir est de la redémarrer. Ce qui a le mérite de fermer proprement la session EXCEL appelée.

Assurément, une solution satisfaisante serait de pouvoir coder directement dans le tableur ou, à la rigueur, utiliser un protocole d'échange des données en mémoire s'appuyant sur un encodage plus performant que le type « variant » utilisé avec OLE. A la fin de ce didacticiel, nous ferons référence à d'autres projets, de la même teneur, qui semblent autrement plus intéressants d'explorer.

Travailler avec XL-SIPINA

Ce logiciel est avant tout conçu comme un outil exploratoire. Rien n'a donc été prévu pour l'évaluation des modèles. Ses principales fonctionnalités sont essentiellement la possibilité de visualiser l'arbre, de le modifier selon les connaissances du domaine, d'interpréter la constitution des groupes à l'aide d'une panoplie de statistiques descriptives comparatives.

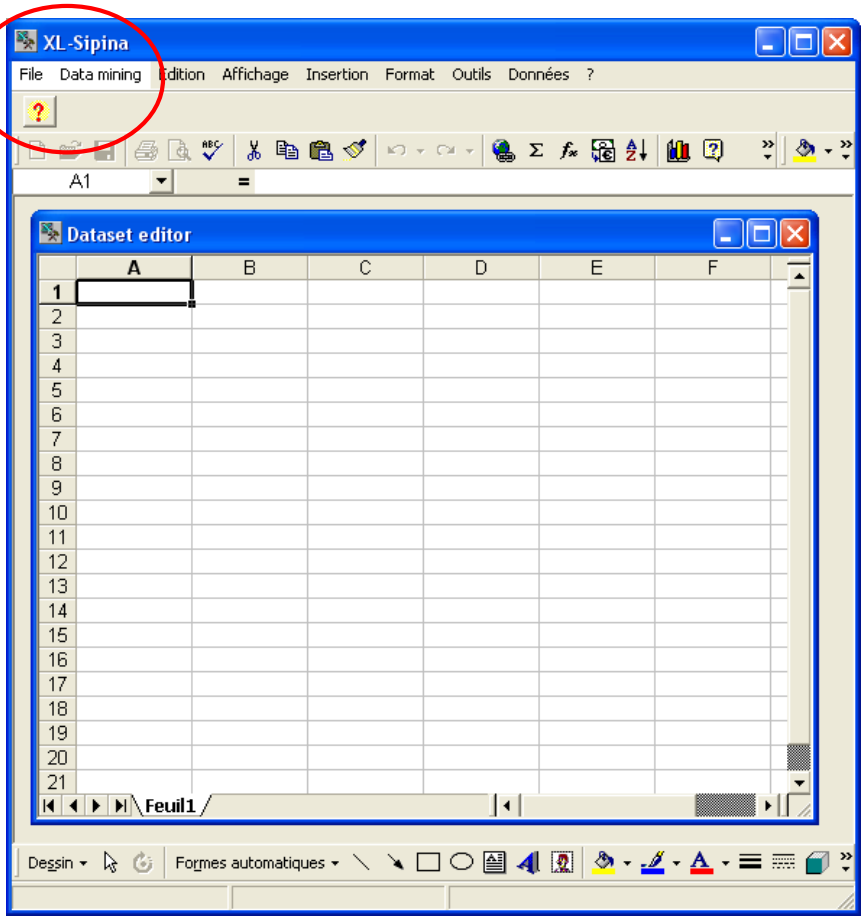
Bien entendu, il n'est pas question d'utiliser le logiciel sur de grandes bases de données. Ce n'est pas la finalité. Nous sommes de toute manière limités par les capacités d'Excel en nombre de lignes et de colonnes. Le lecteur qui souhaite se tourner vers des outils plus performants pourra par exemple utiliser la version recherche de SIPINA qui intègre également des outils d'évaluation de la qualité de l'apprentissage (subdivision apprentissage-test, validation croisée, etc.).

Données

Pour illustrer le fonctionnement du logiciel, nous avons choisi la base AUTOMOBILE DATABASE disponible sur internet. Nous l'avons nettoyée, notamment en l'épurant des variables difficiles d'interprétation. Nous avons également supprimé les variables en relation directe, par construction, avec la variable que nous avons choisie de prédire. Dans ce didacticiel, nous voulons expliquer le classement des véhicules (Risquée ou Non) réalisé par les assureurs. Nous souhaitons relier cette annotation experte avec les caractéristiques objectives des véhicules (Consommation, Puissance, Carrosserie, etc.).

Lancement du logiciel

Au lancement, le logiciel se présente comme suit.

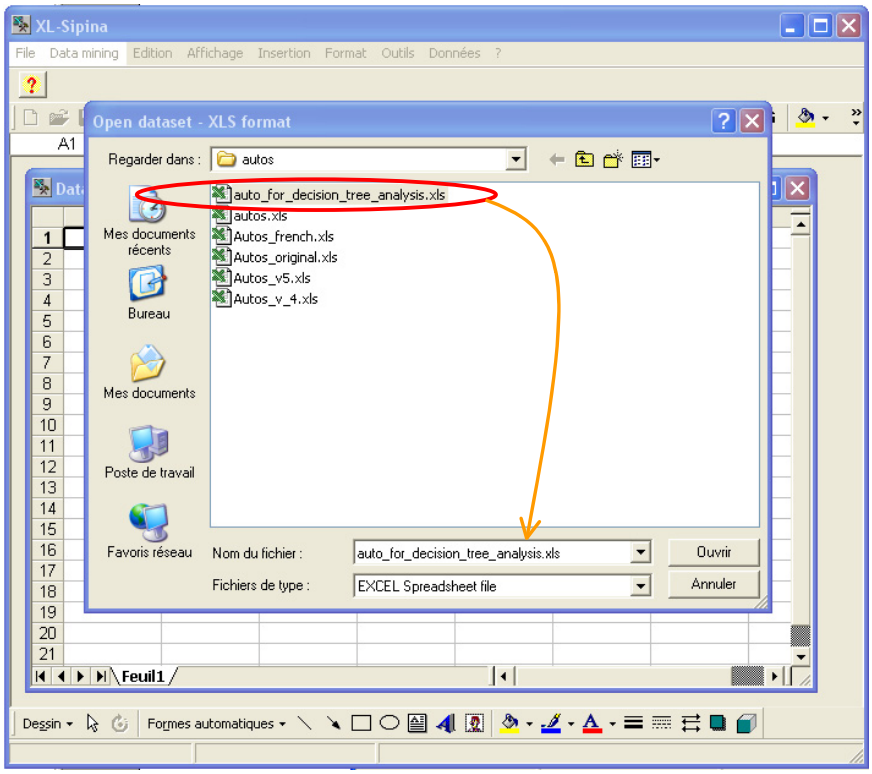


Il y a un petit temps d'attente avant que la grille du tableur ne s'affiche, cela est dû à l'initialisation du serveur OLE. Si le tableur EXCEL n'est pas présent sur le système, un message d'erreur apparaîtra, il ne sera pas possible d'aller plus loin.

Les deux premiers menus, « FILE » et « DATA MINING » sont propres au logiciel. Les autres appartiennent à EXCEL. Ils deviennent invisibles lorsque la grille n'est plus activée, ce sera le cas lorsque nous manipulerons l'arbre de décision. Un bouton d'aide résumant les fonctionnalités du logiciel est disponible dans la barre d'outils.

Chargement des données

Pour charger les données, nous activons le menu FILE / OPEN. Une boîte de dialogue permet de sélectionner le fichier de données.



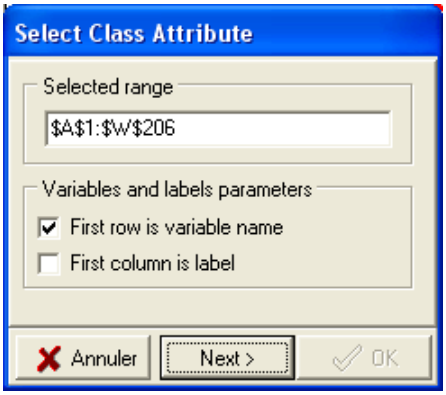
« RISKY » est la variable à prédire. Elle indique une annotation des assureurs sur les véhicules qu'ils jugent plus risqués à garantir. L'objectif de l'étude est d'expliquer les tenants de cette annotation experte à partir des caractéristiques objectives des véhicules telles que leur prix, leur consommation, leur performance, etc.

Le fichier comprend 205 observations. La variable à prédire prend deux valeurs possibles « POSITIVE » et « NEGATIVE ». Il y a 22 prédicteurs, 8 d'entre eux sont discrets. C'est un fichier particulièrement intéressant. L'objectif n'est pas tant de prédire avec précision l'annotation mais plutôt de comprendre les idées qui la sous-tendent, avec un certain nombre de variables concurrentes et/ou complémentaires.

Lancer le traitement

Il est nécessaire de **sélectionner l'ensemble des données**, y compris l'en-tête des variables, avant de lancer une analyse. Les données doivent être d'un seul bloc, il n'est pas possible d'effectuer des sélections multiples.

Puis, nous activons le menu **DATA MINING / START LEARNING**. Une boîte de dialogue apparaît, elle indique la plage de données sélectionnée.

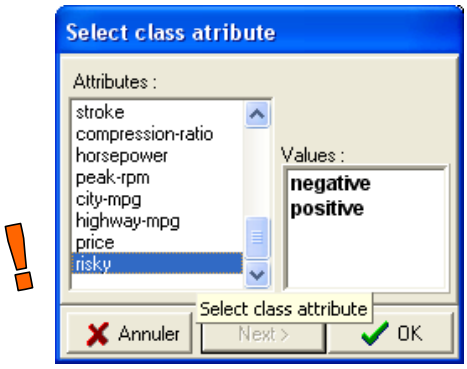


Deux options peuvent être cochées. La première nous demande si la première ligne des données correspond aux noms des variables, ce sera le cas souvent. La seconde nous permet d'indiquer si la première colonne correspond aux identifiants des observations.

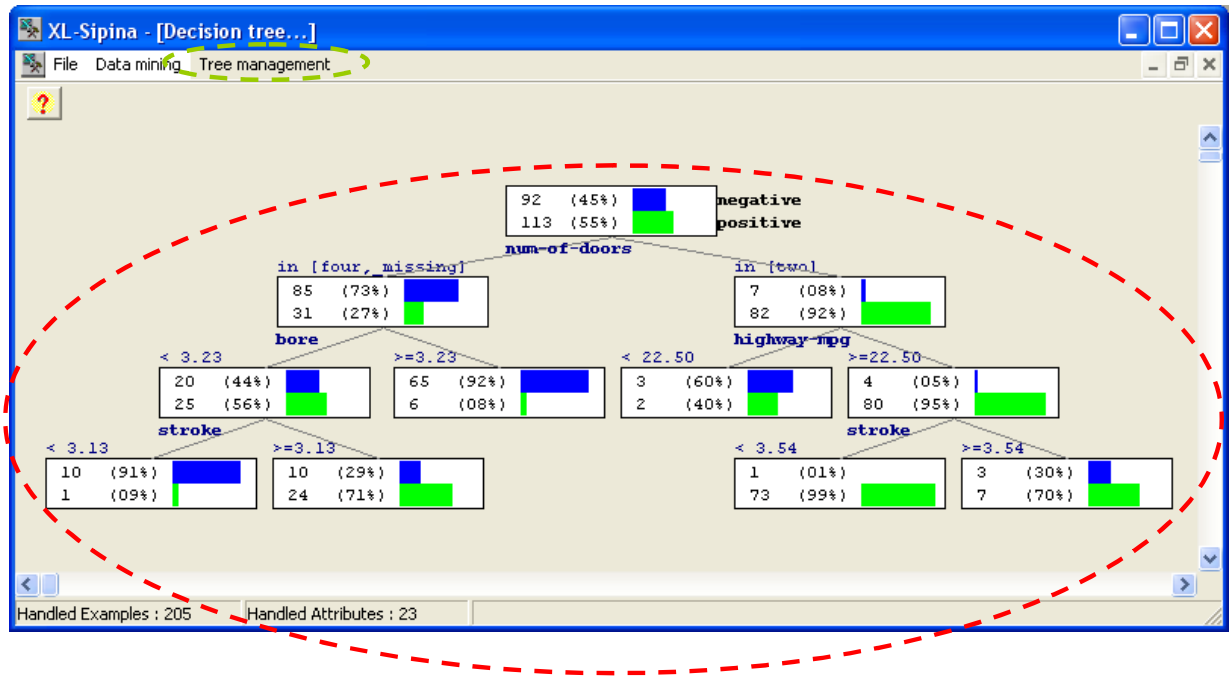
Nous cliquons sur le bouton **NEXT, les données sont alors scannées**. Le logiciel identifie à ce stade les variables catégorielles des variables numériques. La stratégie utilisée est très simple. Si la première observation peut être transformée en valeur numérique, il est décidé qu'il s'agit d'une variable continue. Elle est codée en variable discrète dans le cas contraire. Cette opération délicate peut se révéler assez lente si la plage de données est importante.

Notons que les données manquantes ne sont pas traitées par le logiciel. Il nous revient d'effectuer les pré-traitements adéquats avant de lancer une analyse.

Lorsque les données sont correctement encodées, nous devons **indiquer au logiciel la variable à prédire. Elle est forcément discrète**. Dans notre exemple, nous choisissons la variable RISKY puis nous cliquons sur OK.



Le calcul est automatiquement exécuté. L'arbre s'affiche dans une nouvelle fenêtre, un nouveau menu « TREE MANAGEMENT » vient se greffer sur la barre de menu. Les menus en provenance d'EXCEL sont masqués maintenant.



La méthode utilisée est dérivée de CHAID, elle est explicitée dans un didacticiel publié dans la revue MODULAD. Dans notre paramétrage par défaut, l'arbre est volontairement limité à 4 niveaux. A charge pour l'utilisateur de modifier ce paramétrage (voir Modification du paramétrage plus bas) ou de poursuivre la construction de l'arbre manuellement.

Explorer l'arbre

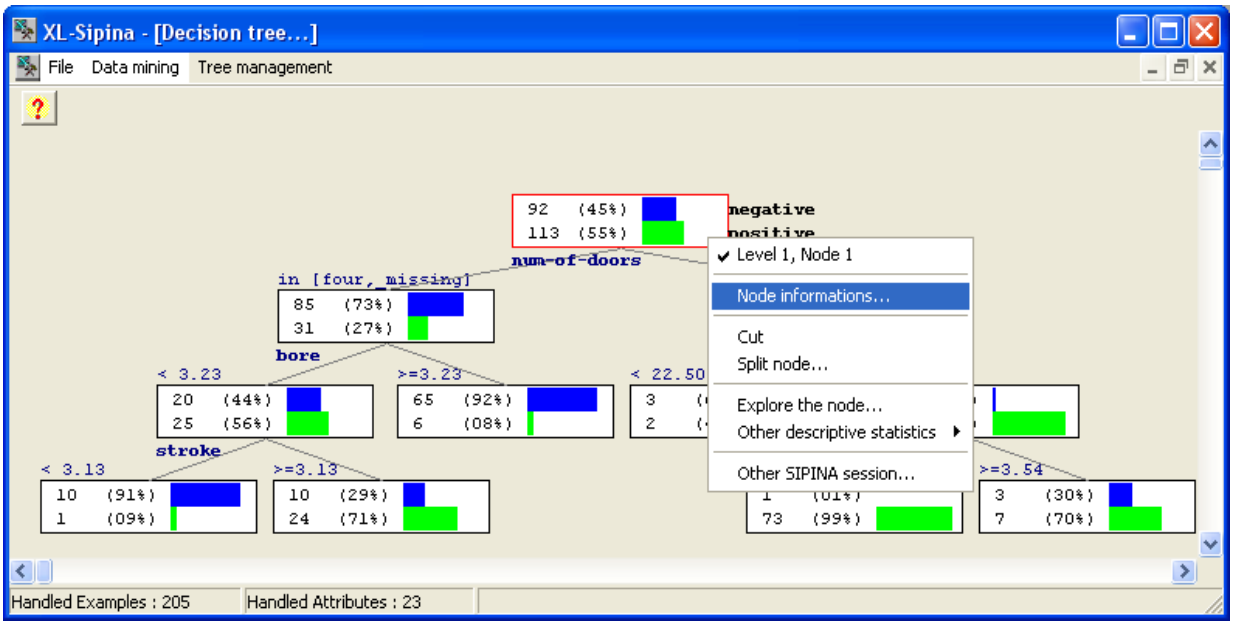
Le principal atout des arbres de décision par rapport aux autres techniques de Data Mining est la possibilité d'explorer en détail les résultats proposés par la méthode et, éventuellement, de la guider vers des solutions plus en adéquation avec les connaissances du domaine.

Voir les variables de segmentation candidates

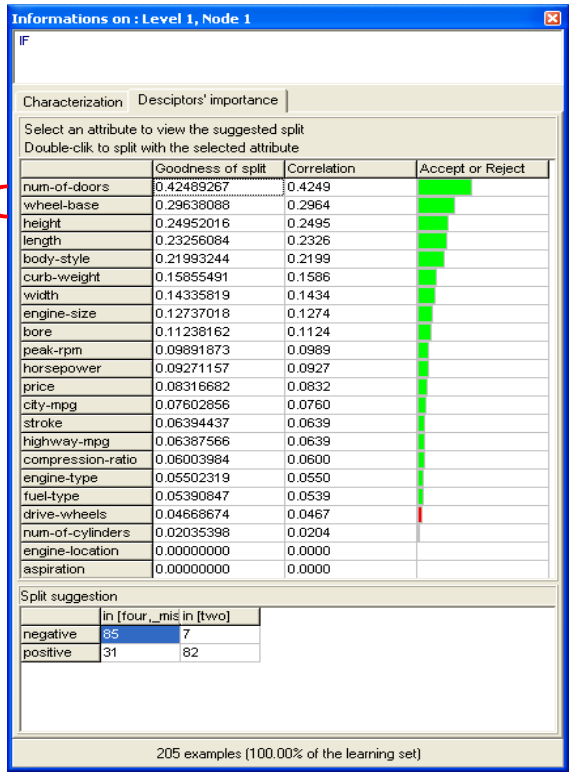
Nous constatons sur la racine de l'arbre que 55% des observations ont été annotées « risquées » (POSITIVE). La première variable de segmentation est « nombre de portes » (NUM-OF-DOORS) : sur le sommet à droite au second niveau de l'arbre, 89 disposent de 2 portières, la proportion des véhicules risqués passe à 92% dans ce cas.

La variable NUM-OF-DOORS a été automatiquement choisie par la méthode. La question est qu'en est-il des autres variables ? Est-ce qu'elles auraient permis d'isoler de manière différente les véhicules à risque ?

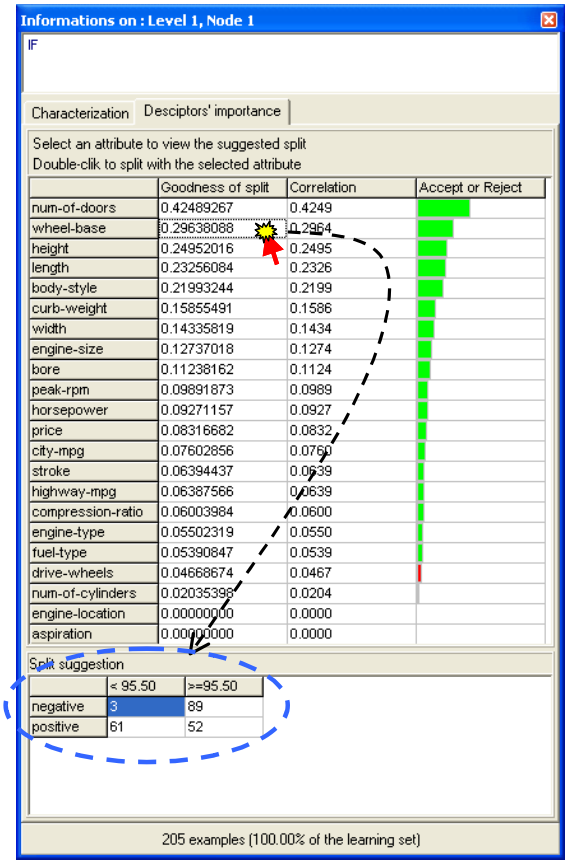
Pour le savoir, nous explorons le sommet initial, la racine de l'arbre. Nous le sélectionnons. Le rectangle encadrant le sommet passe à la couleur rouge. Puis nous faisons apparaître le menu contextuel (clic avec le bouton droit de la souris). Nous activons alors le menu « NODE INFORMATIONS… ».



Dans la fenêtre d'information qui apparaît, nous constatons que NUM-OF-DOORS était bien de loin la plus pertinente pour mettre en évidence les véhicules risqués. Le t de TSCHUPROW – qui peut s'interpréter comme le carré du coefficient de corrélation pour variables catégorielles binaires – associé à la segmentation est égal à 0.42489 ; la variable candidate suivante, WHEEL-BASE (empattement) propose un indice égal 0.29638.



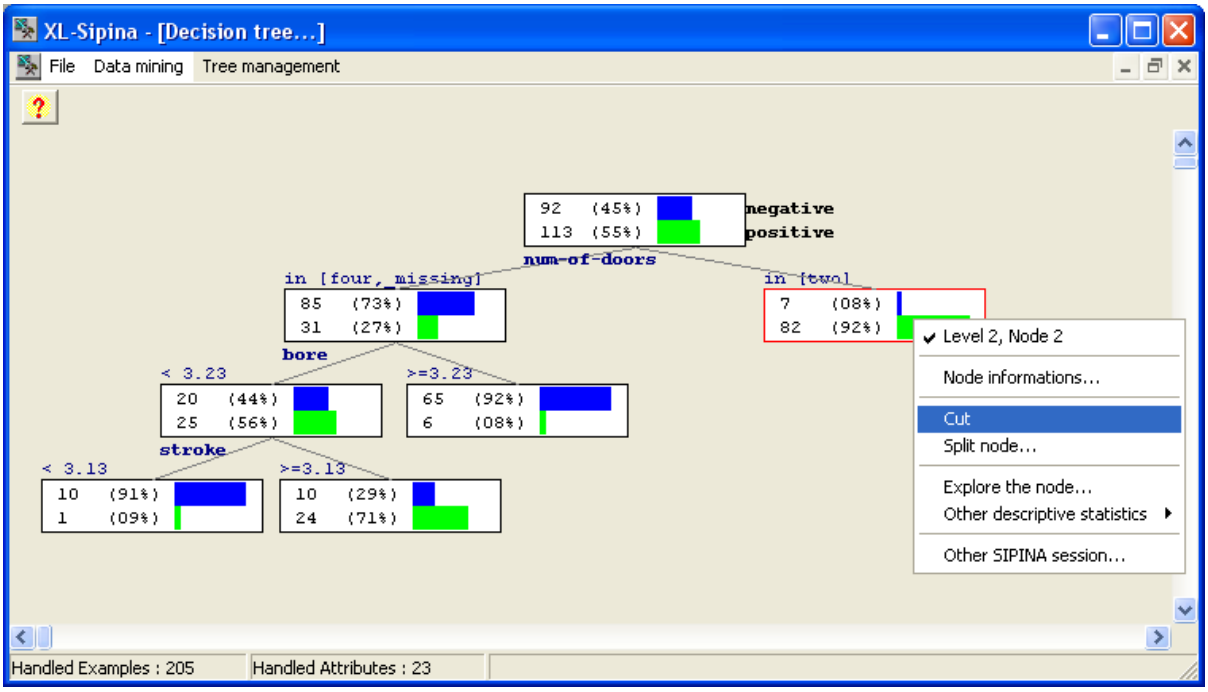
Lorsque nous sélectionnons la variable candidate dans la liste, plus précisément le rectangle contenant la valeur de l'indice de qualité, la subdivision proposée apparaît dans la partie basse de la fenêtre (SPLIT SUGGESTION). Dans le cas de WHEEL-BASE, le seuil de discrétisation est de 95.5 si nous segmentons avec cette variable.



Dans le cas de la méthode implémentée dans ce logiciel, une variante de CHAID, la colonne CORRELATION et GOODNESS OF SPLIT renvoient des valeurs identiques. La dernière colonne nous indique la situation du partitionnement par rapport à la règle d'arrêt : l'histogramme est de couleur verte (resp. rouge) si la segmentation proposée est autorisée (resp. refusée).

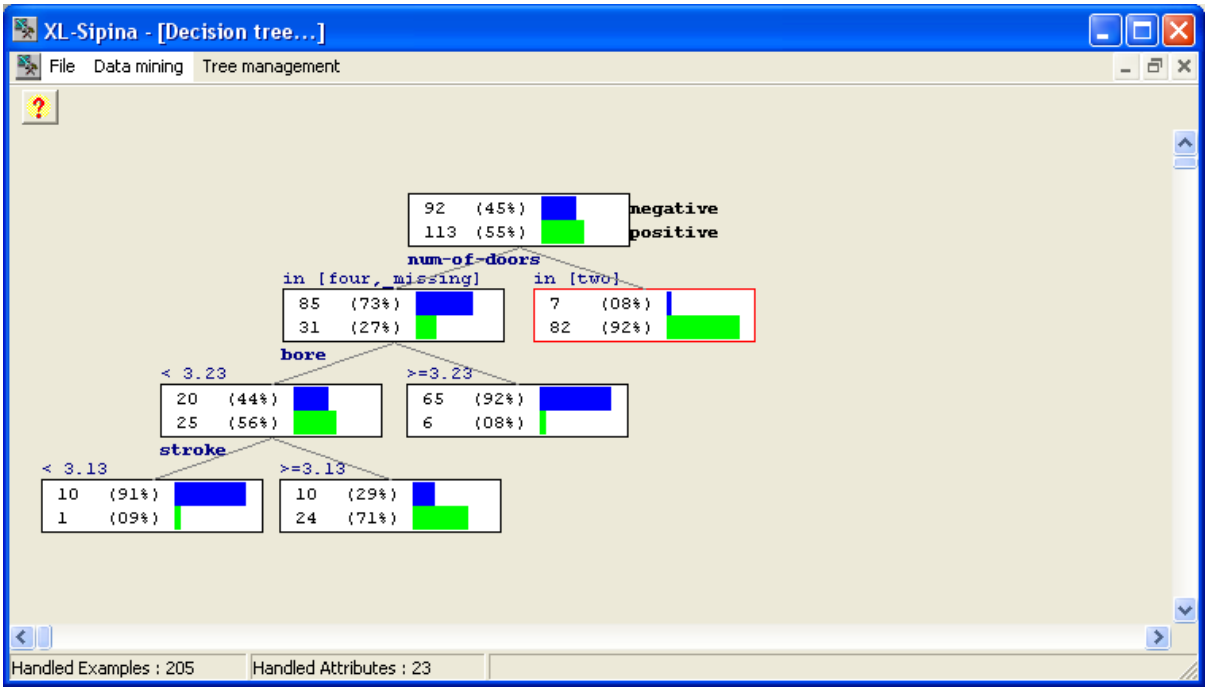
Elaguer manuellement l'arbre

Intéressons-nous maintenant au sommet de droite sur le deuxième niveau de l'arbre. Il décrit 89 observations, dont 82 (92%) sont positifs. Nous décidons de valider ce groupe, et par conséquent d'élaguer le sous-arbre qui lui est consécutif. De nouveau, nous faisons apparaître le menu contextuel à l'aide du clic avec le bouton droit de la souris. Pour supprimer les branches situées sous un sommet, nous activons le menu CUT. Le sommet traité devient ainsi un sommet terminal, une feuille de l'arbre.



Il est possible de réorganiser la disposition des sommets en appuyant sur la touche F5 ou en activant le menu TREE MANAGEMENT / REORGANIZE… Cela peut être particulièrement intéressant lorsque l'arbre est très large, avec de nombreuses feuilles, et qu'il est difficile de visualiser l'ensemble des sommets. Notons que plusieurs options de zoom sont également disponibles, toujours dans le menu TREE MANAGEMENT.

La fenêtre de l'application doit alors se présenter comme suit.



Exploration d'un sommet (d'un sous-ensemble d'observations)

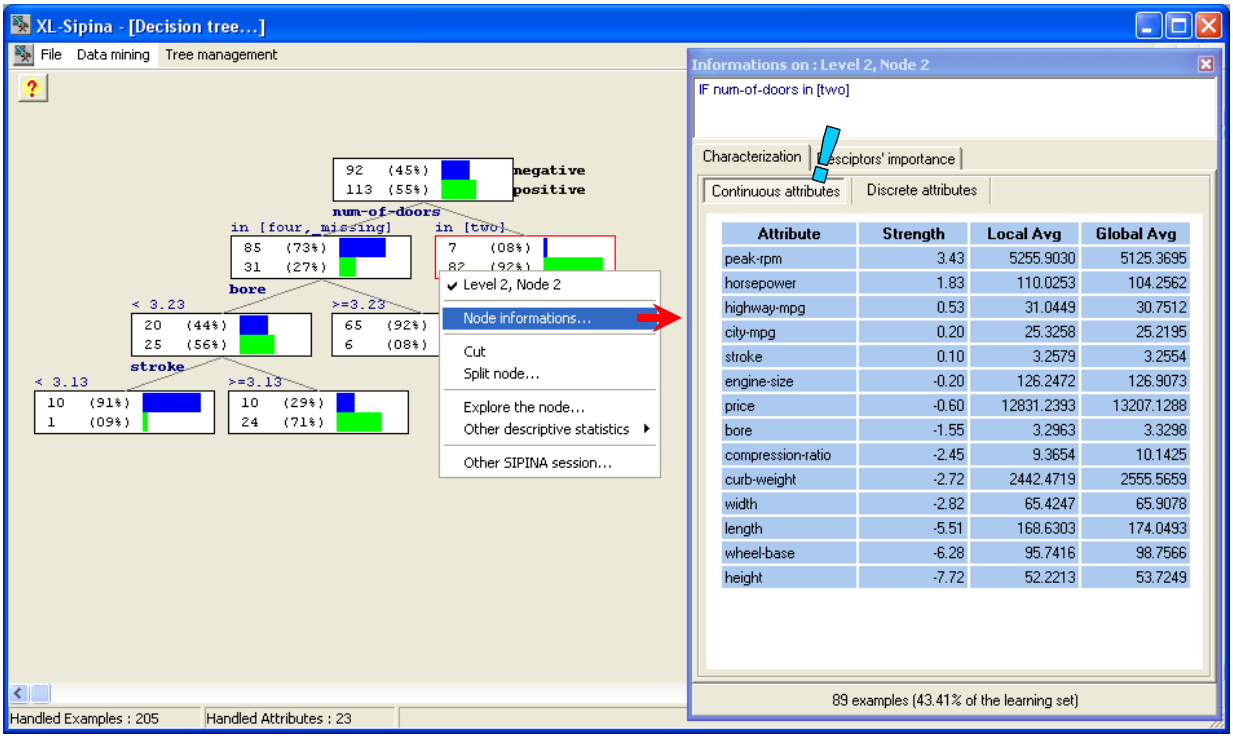
La règle NUM-OF-DOORS = TWO définit un sous-ensemble d'individus où les véhicules risqués sont largement majoritaires. C'est entendu. Quelles sont les autres caractéristiques de ce sous-ensemble d'observations ?

Un arbre propose une règle de description immédiate, très facilement lisible. Elle masque en revanche le rôle des autres variables. La liste de segmentations concurrentes nous permet de mieux comprendre le partitionnement des données. Il nous manque toutefois la description approfondie d'un nœud de l'arbre.

Nous disposons de 3 outils, proposés dans le menu contextuel, pour explorer un sommet dans XL-SIPINA. La première est activée à l'aide du menu **EXPLORE THE NODE**. Elle affiche les données locales, correspondant au nœud, dans une grille. Il nous est donc possible de visualiser, voire de sauvegarder pour une analyse ultérieure, les observations situées sur chaque nœud de l'arbre.

Si les observations locales sont nombreuses, la visualisation des données apporte quand même une information assez limitée. La seconde option permet de calculer des statistiques descriptives sur le sous-ensemble d'individus circonscrits par un sommet. Elle est accessible via le menu **OTHER DESCRIPTIVE STATISTICS**. Elle est assez performante. En effet, nous disposons de tous les outils de statistique descriptive de la bibliothèque de calcul du logiciel (statistiques univariées et bivariées).

La répétition de l'accès au menu et le paramétrage des calculs lors du passage d'un nœud à l'autre peuvent néanmoins se révéler fastidieux à la longue. Un troisième outil permet de calculer des statistiques descriptives, très simplifiées. Elles permettent de caractériser la sous-population par rapport à l'ensemble des observations associées au sommet initial, la racine de l'arbre. L'outil est accessible via le menu contextuel **NODE INFORMATIONS**, nous sélectionnons alors l'onglet CHARACTERIZATION dans la fenêtre qui apparaît.



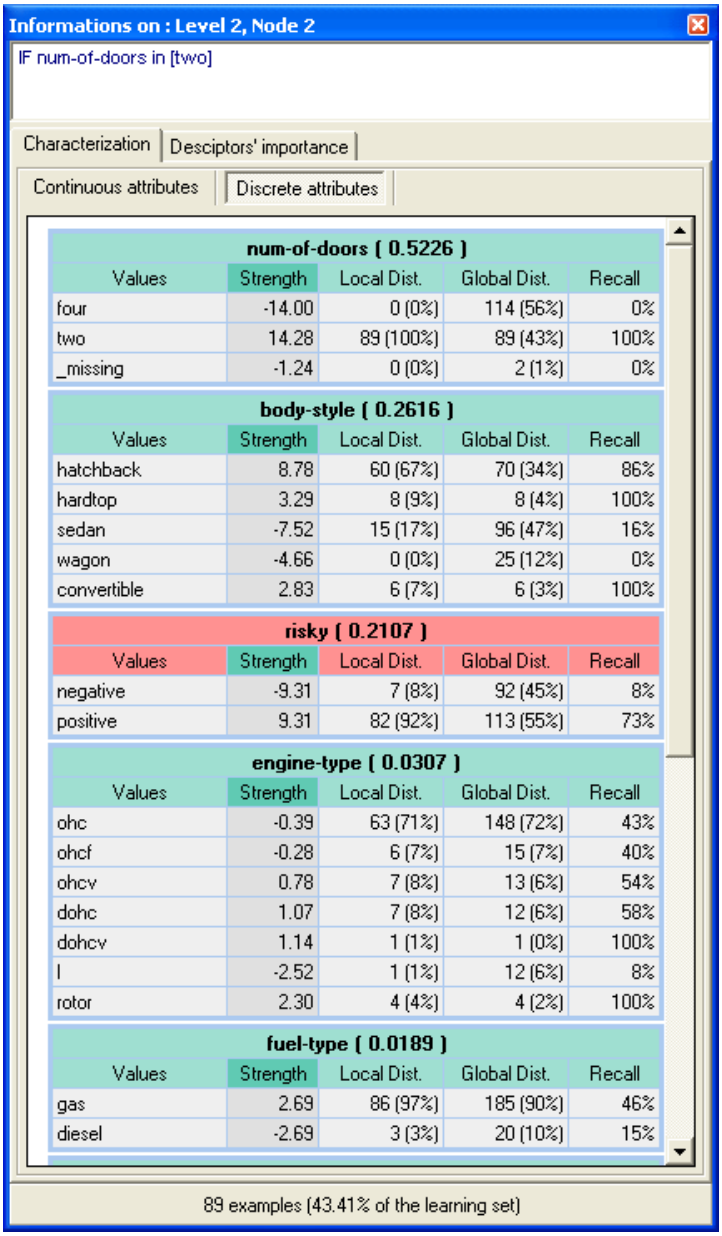
Attribute	Strength	Local Avg	Global Avg
peak-rpm	3.43	5255.9030	5125.3695
horsepower	1.83	110.0253	104.2562
highway-mpg	0.53	31.0449	30.7512
city-mpg	0.20	25.3258	25.2195
stroke	0.10	3.2579	3.2554
engine-size	-0.20	126.2472	126.9073
price	-0.60	12831.2393	13207.1288
bore	-1.55	3.2963	3.3298
compression-ratio	-2.45	9.3654	10.1425
curb-weight	-2.72	2442.4719	2555.5659
width	-2.82	65.4247	65.9078
length	-5.51	168.6303	174.0493
wheel-base	-6.28	95.7416	98.7566
height	-7.72	52.2213	53.7249

Deux sous-onglets sont disponibles, elles correspondent à la caractérisation du sommet selon les variables continues et discrètes.

Dans notre exemple, nous constatons qu'en moyenne les véhicules comportant deux portes, et qui sont pour la plupart classées risquées, possèdent un régime moteur maximal plus élevé (PEAK-RPM = 5255 contre 5125 dans la population globale) ; ils sont légèrement plus puissants (HORSEPOWER = 110.8 contre 104) ; ils sont plus petits (HEIGHT = 52.2 contre 53.7, WHEEL-BASE = 95.7 contre 98.7, etc.).

L'indicateur STRENGTH correspond au T de Student de conformité de la valeur de la moyenne locale à un standard, la moyenne dans la population globale estimée sur la totalité de l'échantillon. Il ne s'agit pas réellement d'un test au sens rigoureux du terme. Il faut avant tout voir STRENGTH comme un simple indicateur normalisé de l'importance de l'écart. En effet, étant exprimées dans des unités différentes, les différences entre les moyennes ne sont pas directement comparables d'une variable à l'autre.

Si nous sélectionnons le sous-onglet DISCRETE ATTRIBUTES, nous visualisons les comparaisons concernant les variables discrètes.



La plus forte différenciation se fait sur le nombre de portes, c'est normal puisqu'elle définit le sommet. Nous constatons que 100% des véhicules sur ce sommet ont bien 2 portes. Plus intéressant en revanche est la deuxième variable. Il s'agit du style de carrosserie (BODY-STYLE). Nous observons une sur-représentation des fourgonnettes (HATCHBACK = 67% contre 34% dans la totalité de l'échantillon) et… décapotables (HARDTOP = 9% contre 4% ; CONVERTIBLE = 7% contre 3%). Cette dernière caractéristique semble très marquante puisque 100% des HARDTOP (8/8 : RECALL = 100%) et CONVERTIBLES (RECALL = 6/6 = 100%) sont dans ce groupe.

L'indicateur STRENGTH dans ce cas correspond à la valeur Z d'un test de comparaison de proportions. S'il est positif avec une valeur élevée, cela indique que la modalité est comparativement plus fréquente sur le sommet ; s'il est négatif et élevé en valeur absolue, la modalité est plus rare.

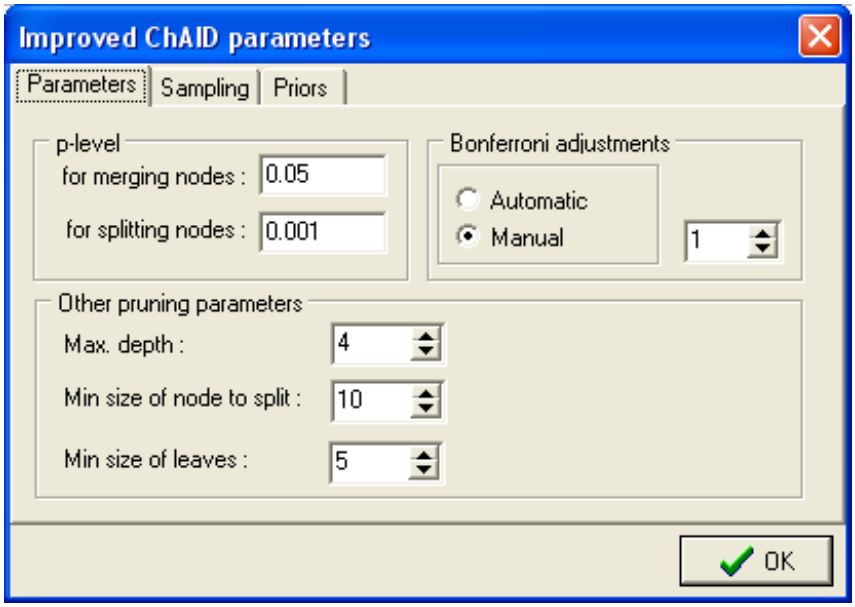
Que des fourgonnettes et des véhicules décapotables possèdent deux portes n'a rien d'étonnant. Qu'elles soient unies dans les annotations des assureurs paraît plus étrange. C'est à ce stade que le rôle de l'expert du domaine devient très important, nous éclairant sur cette apparente contradiction, proposant

les bonnes variables permettant de les différencier, identifiant la variable manquante dans l'analyse, etc. Les arbres de décision constituent un outil privilégié d'exploration de données dans ce contexte.

Modifier les paramètres de traitement

Pour modifier les paramètres de la méthode d'induction d'arbre de décision, nous devons tout d'abord interrompre l'étude courante en activant le menu DATA MINING / STOP LEARNING.

Puis, nous cliquons sur le menu DATA MINING / PARAMETERS, la boîte de dialogue suivante apparaît.



Les paramètres les plus importants sont regroupés dans l'onglet PARAMETERS. Ils correspondent peu ou prou aux paramètres de la méthode CHAID. Ils permettent avant tout de contrôler la profondeur de l'arbre lors du traitement automatique.

Conclusion

Ce logiciel est un exercice de style abandonné il y a bien longtemps. A l'époque, vers 1998-1999, la technologie utilisée pouvait paraître novatrice, elle l'est nettement moins aujourd'hui. Il s'appuie de plus sur les structures internes de SIPINA devenues obsolètes. Par la suite, je me suis lancé dans le projet TANAGRA auquel je consacre maintenant tous mes efforts. Un de mes enjeux dans ce nouveau projet étant de maximiser le rapport entre le code de calcul et le code d'interface, les choix technologiques sont totalement différents : application stand-alone aussi légère que possible ; autonomie totale de l'application, aucun accès à une bibliothèque extérieure d'interface n'est nécessaire ; interface graphique réduite à sa plus simple expression, etc.

Néanmoins, l'idée de marier intimement les outils de Data Mining avec un tableur semble très intéressante. Les utilisateurs sont très souvent familiarisés avec les tableurs. Leur offrir dans cet environnement de travail des fonctionnalités étendues de traitement statistique des données est tout à fait naturelle. A ce titre, je ne peux qu'attirer l'attention du lecteur sur le projet GNUMERIC.

Il s'agit d'un tableur libre, de qualité quasi-équivalente aux logiciels commerciaux, tourné vers le calcul scientifique. Il est « open source », nous pouvons lui intégrer nos propres modules de calcul. Toute la partie gestion de données est disponible, notre seule contrainte est de lire les données dans le classeur courant avant de lancer les calculs et afficher les résultats dans d'autres feuilles du même classeur. De

fait, les techniques de Data Mining deviendront des fonctions supplémentaires, complétant les innombrables fonctionnalités déjà disponibles dans le tableur.

Par rapport au schéma présenté dans ce didacticiel, les avantages sont nombreux : nous ne sommes plus assujettis à la présence hypothétique d'un logiciel commercial sur la machine de l'utilisateur ; accéder directement au code source permet d'optimiser le calcul, nous n'avons plus à passer par des appels à des protocoles compliqués, voire instables ; en mutualisant les contributions de chacun, la bibliothèque pourra être très rapidement enrichie ; l'accessibilité du code source garantit la pérennité du projet au sein de la communauté scientifique.

Un projet à suivre… auquel je pense contribuer d'une manière ou d'une autre dans l'avenir. Mon grand bonheur serait d'en faire un outil collaboratif en ligne, dans la lignée de GOOGLE SPREADSHEET et autres. Quand je vois par exemple que EDITGRID s'appuie en grande partie sur GNUMERIC, il y a assurément matière à faire des choses intéressantes.

Epilogue et versions plus récentes d'Excel (29/08/2010)

Avec un peu de recul maintenant (29/08/2010), je me rends compte que **la solution** « macro complémentaire » **qui permet d'intégrer un menu SIPINA dans Excel est finalement la plus viable**. Elle est simple, fiable (les deux vont de pair souvent), et très performante (quelques secondes suffisent pour transférer une base comportant 100.000 observations et 22 variables). C'est donc la solution que j'utilise moi-même lorsque je traite mes fichiers de données.

Un autre aspect a titillé ma curiosité récemment. Je voulais savoir comment réagissait cette variante spécifique de SIPINA (XL-SIPINA ne fonctionne que si Excel est présent sur la machine, rappelons le) avec les dernières versions d'Excel (Office 2007 et Office 2010). J'ai constaté que le dispositif fonctionne sans écueils. Toutes les fonctionnalités d'exploration des données sont actives. Nous avons pu reproduire toutes les opérations d'induction d'arbres de décision décrites dans ce didacticiel. L'interface est en revanche modifiée. Le ruban d'Office prend place dans le logiciel lorsqu'il est démarré. Voici une copie d'écran pour Excel 2010.