# 1 Objectif

**Montrer les fonctionnalités de SIPINA en matière d'exploration interactive des données.**

SIPINA regorge de fonctionnalités mal connues des utilisateurs. Dans ce didacticiel, nous essayons de mieux expliciter les outils que propose le logiciel pour comprendre, interpréter, et manipuler l'arbre de décision construit par l'algorithme d'apprentissage.

# 2 Données

Nous travaillons sur les données BLOOD_PRESSURE_LEVELS.XLS, accessible en ligne[1]. Le fichier comporte 399 observations. Il s'agit de comprendre l'occurrence de l'hypertension chez le patient (pression artérielle systolique > 140 mm hg) à partir d'une série de caractéristiques qui lui sont rattachées (fumer ou pas, le sexe, le surpoids, le niveau d'éducation, etc.). Voici la liste des variables, en bleu la variable à expliquer, en vert les variables explicatives :

| Attribute | Category | Informations |
|---|---|---|
| bpress_level | Discrete | 2 values |
| gender | Discrete | 2 values |
| smoke | Discrete | 2 values |
| exercise | Continue | - |
| overweight | Continue | - |
| alcohol | Continue | - |
| stress | Continue | - |
| salt | Continue | - |
| income | Continue | - |
| education | Continue | - |

Nous adoptons une démarche descriptive. Il ne s'agit pas de prédire avec le plus de précision possible la survenue de l'hypertension, mais plutôt de caractériser les individus qui en souffrent à partir des variables disponibles. Dans cette optique, la validation s'appuie avant tout sur l'expertise, ce sont les médecins qui vont nous confirmer si les règles proposées par l'arbre sont en adéquation avec les connaissances du domaine ou non.

Si nous avions voulu élaborer un modèle prédictif, il aurait fallu dès le départ prévoir la procédure de validation. Deux approches sont possibles. La première consiste à subdiviser en 2 parties le fichier : réserver une partie pour construire l'arbre de décision ; utiliser l'autre partie pour en évaluer les performances (taux d'erreur, sensibilité, précision, etc.). Nous pouvons laisser la méthode d'apprentissage construire l'arbre, puis le tester ; nous pouvons également intervenir interactivement pour définir un arbre qui tient compte des connaissances et des contraintes du domaine, qu'un algorithme automatisé ne peut (sait) pas appréhender. Il faut néanmoins rester

[1] http://eric.univ-lyon2.fr/~ricco/dataset/blood_pressure_levels.xls ; le fichier a été retravaillé, la source initiale est http://www.math.yorku.ca/Who/Faculty/Ng/ssc2003/BPMain.htm

prudent quant à la recherche interactive du meilleur arbre. A force de chercher l'arbre le plus performant sur le fichier test, nous sommes en train de le faire participer à l'apprentissage, le taux d'erreur qu'il fournit devient de plus en plus biaisé au fil des tentatives que nous effectuons.

La seconde approche d'évaluation, qui semble plus appropriée pour notre fichier, vu le peu d'observations disponibles, est d'utiliser les techniques de rééchantillonnage (validation croisée, bootstrap). Dans ce dernier cas, il ne saurait être question de construction interactive de l'arbre.

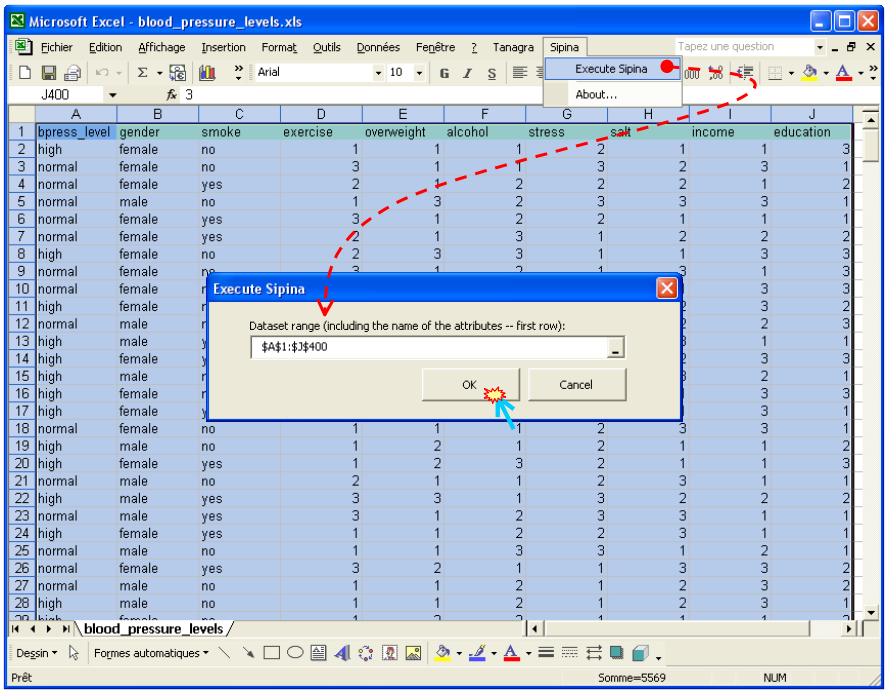
Dans ce didacticiel, nous situant plutôt dans une démarche descriptive, nous faisons intervenir la totalité du fichier pour la construction de l'arbre. Nous comptons sur l'interprétation des résultats pour statuer sur la crédibilité des solutions fournies par la méthode.

# 3 Construire un arbre de décision avec le logiciel SIPINA

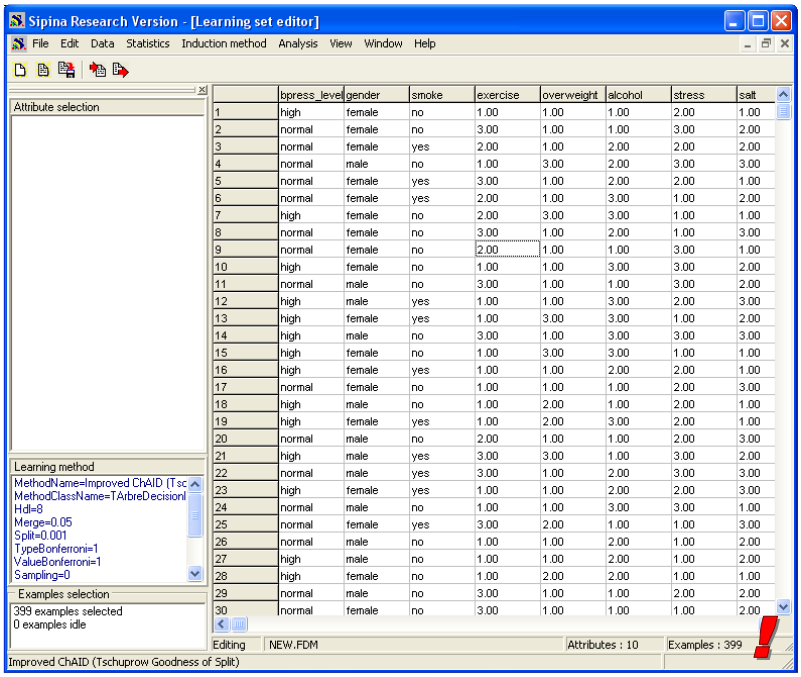
## 3.1 Charger/Importer les données

Le plus simple est de charger dans un premier temps le fichier dans le tableur EXCEL. Depuis l'élaboration de la macro complémentaire SIPINA.XLA, il est possible de faire la jonction entre ce tableur et notre logiciel. Pour savoir comment installer cette macro, le plus simple est de consulter le tutoriel disponible en ligne (http://tutoriels-data-mining.blogspot.com/2014/08/ladd-in-sipina-pour-excel-2007-et-2010.html).

Après avoir sélectionné la plage de données, nous activons le menu SIPINA/EXECUTE SIPINA disponible maintenant dans EXCEL. Une boîte de dialogue apparaît, nous confirmons en cliquant sur OK après avoir vérifié les coordonnées de la sélection. Attention, pour que SIPINA manipule correctement le fichier, il faut que la première ligne soit constituée des noms de variables.

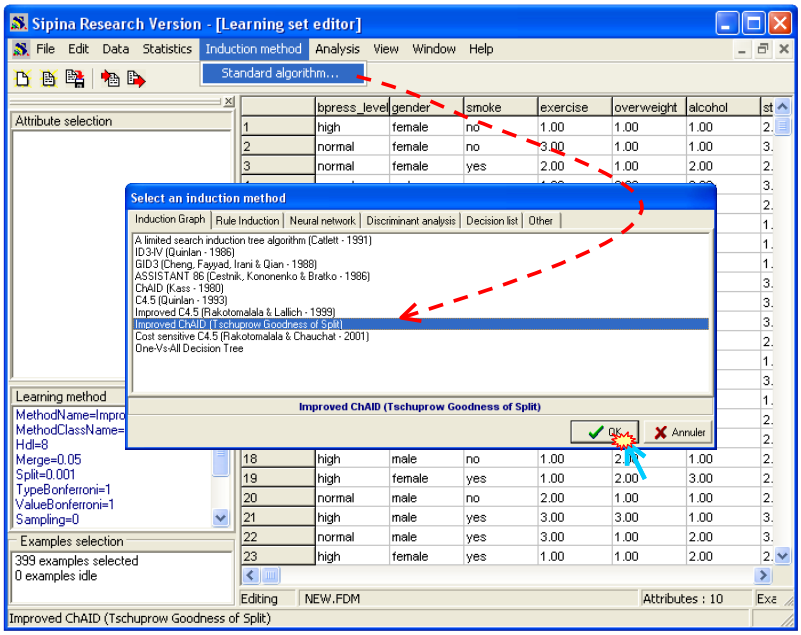


SIPINA est démarré automatiquement. Les données sont transférées via le presse-papier. Le nombre de variables et le nombre d'observations sont affichés dans la partie basse de l'éditeur (10 variables et 399 observations). Il est possible d'effectuer des modifications et de nouvelles saisies dans la grille de données SIPINA, mais ce n'est guère conseillé, EXCEL propose des fonctions d'édition autrement plus performantes. Attention, SIPINA ne sait pas gérer les données manquantes, il vous incombe de les traiter en amont avant de les charger.




## 3.2 Choisir une méthode d'apprentissage

Première démarche toujours dans tout logiciel de traitement exploratoire des données : choisir la technique d'analyse. Pour ce faire, nous activons le menu INDUCTION METHOD / STANDARD ALGORITHM, une boîte de dialogue décrivant les algorithmes disponibles apparaît.

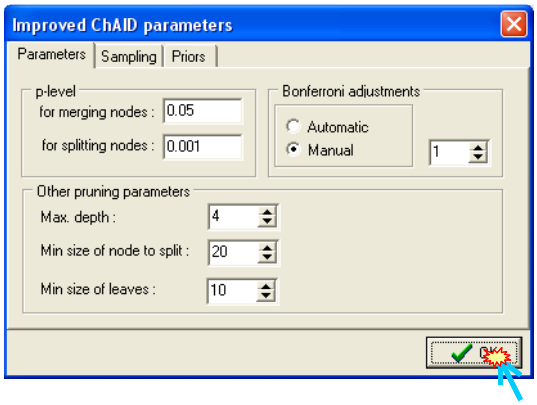


Comme nous pouvons le constater, SIPINA intègre plusieurs techniques d'apprentissage supervisées, autres que les arbres de décision. Leur intérêt dans ce logiciel est très relatif depuis que nous avons mis en ligne **TANAGRA** (http://eric.univ-lyon2.fr/~ricco/tanagra/) qui reprend, dans un cadre harmonisé, l'ensemble de ces méthodes, et en y adjoignant les techniques non supervisées (classification) et descriptives (analyse factorielle). *SIPINA n'est vraiment intéressant que pour les aspects exploratoires des arbres de décision où l'utilisateur a la possibilité d'interagir avec le logiciel lors de la construction du modèle de prédiction. C'est ce que nous présentons dans ce didacticiel.*

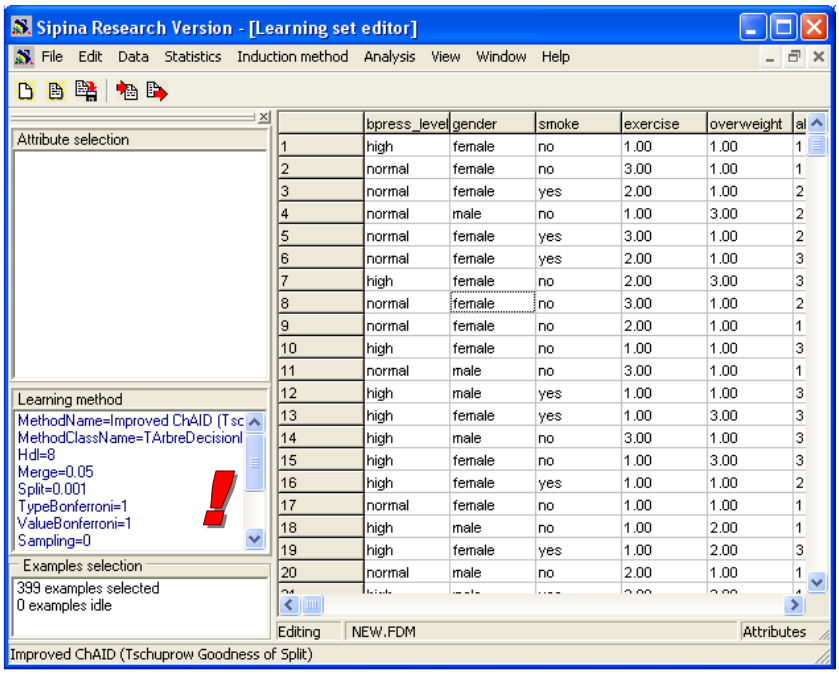
Plusieurs algorithmes de construction des arbres de décision sont disponibles, pour la plupart référencées dans la littérature. Pour un premier contact avec les données, je conseillerais la méthode IMPROVED CHAID. Elle cherche à produire un arbre de taille limitée, permettant d'initier une exploration des données. Elle n'a pas la prétention de fournir un arbre performant, il faudrait se tourner vers des méthodes éprouvées telles que C4.5 (Quinlan, 1993) ou C&RT (Breiman et al., 1984 ; disponible dans Tanagra) dans ce cas.

Nous cliquons sur OK pour valider ce choix, une seconde boîte de dialogue apparaît, elle nous permet de fixer les paramètres de la méthode. Les plus simples à appréhender pour un néophyte sont les contraintes sur les effectifs. Nous conseillons la lecture des supports en ligne sur les arbres de décision pour mieux comprendre le rôle de ces paramètres (http://eric.univ-lyon2.fr/~ricco/cours/slides/Arbres_de_decision_Introduction.pdf ; http://eric.univ-lyon2.fr/~ricco/doc/tutoriel_arbre_revue_modulad_33.pdf).

Nous nous contentons de valider les paramètres par défaut proposés par le logiciel.

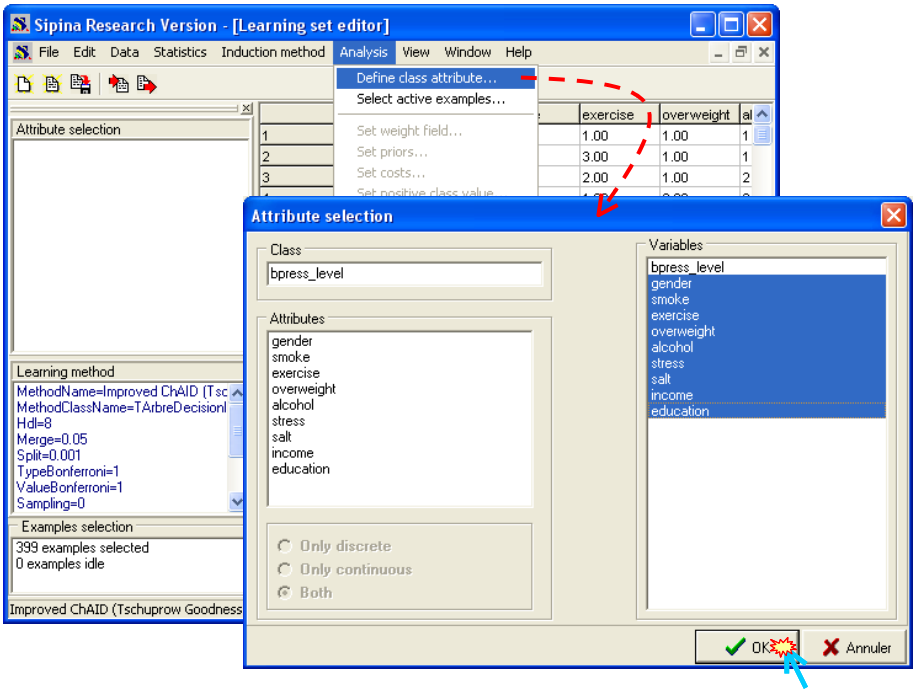



Les choix de l'utilisateur (méthode + paramètres) sont retranscrits dans la section médiane de la partie gauche du logiciel.

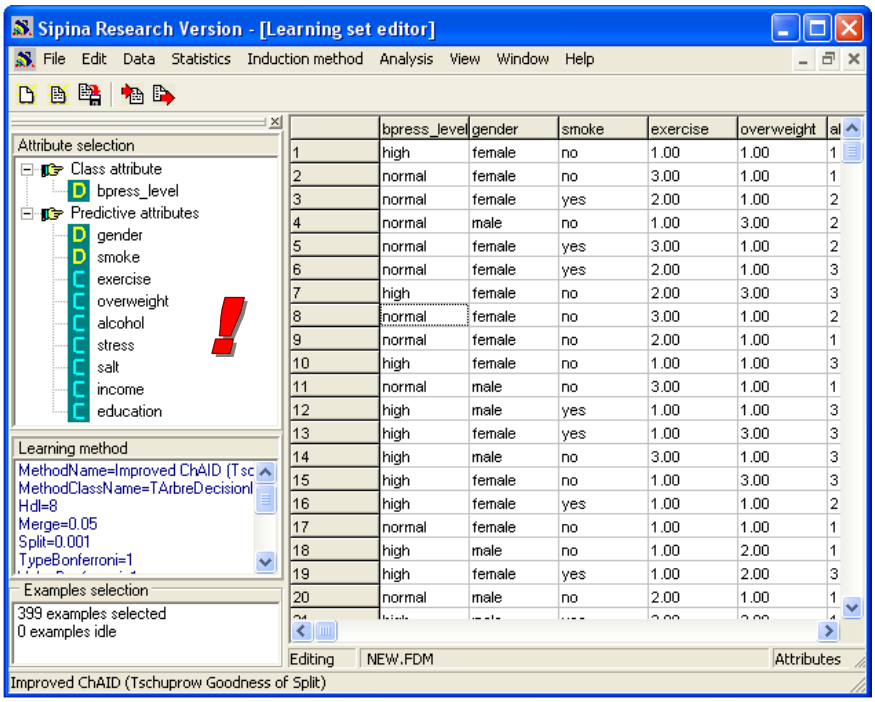



## 3.3 Définir le problème

L'étape suivante consiste à spécifier la variable à expliquer et les variables explicatives candidates. Pour cela, nous activons le menu ANALYSIS/DEFINE CLASS ATTRIBUTE. Une boîte de dialogue de sélection apparaît. Par *glisser déposer*, nous plaçons en CLASS la variable BPRESS_LEVEL, en ATTRIBUTES les autres. La sélection multiple est possible.



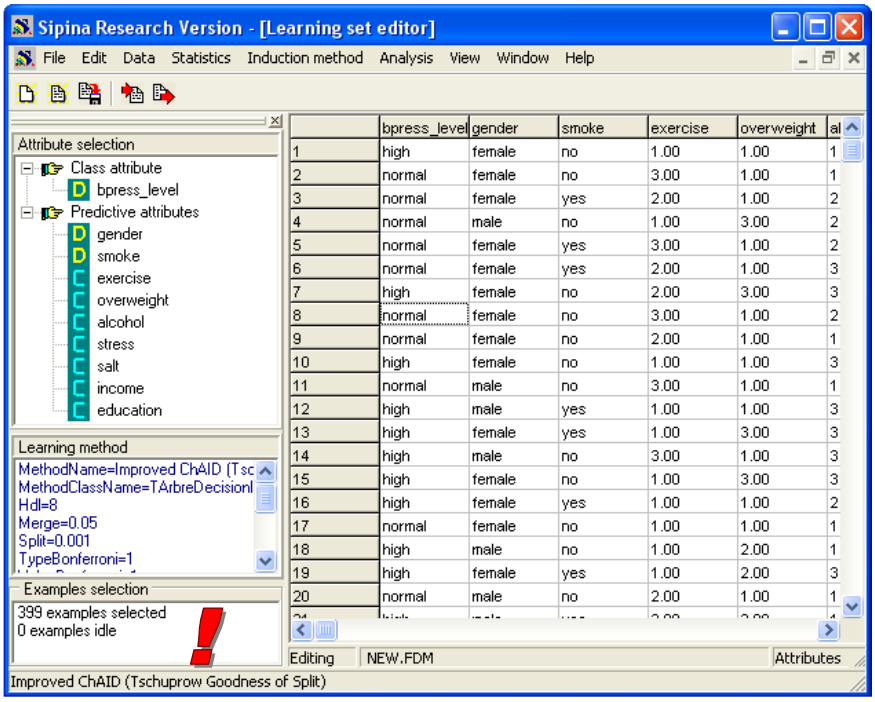

Nous validons en cliquant sur OK. Les choix opérés par l'utilisateur sont retranscrits dans la section haute de la partie gauche de la fenêtre principale. La variable cible est forcément qualitative nominale (D pour discrète), les explicatives peuvent être qualitatives ou quantitatives (C pour continue).




Les variables explicatives ne sont pas toutes forcément pertinentes. Un des intérêts des arbres justement est que la méthode va sélectionner pour nous les variables les plus intéressantes dans le processus d'explication.

## 3.4 Choix des observations

Dans notre cas, nous voulons exploiter l'ensemble des observations pour l'élaboration de l'arbre. C'est la spécification par défaut retenue par SIPINA lors de l'importation des données, elle est retranscrite dans la section basse de la description de l'étude.



Si nous avions voulu modifier la sélection, nous aurions activé le menu ANALYSIS / SELECT ACTIVE EXAMPLES, et défini dans la boîte de sélection qui apparaît la subdivision appropriée des données. Nous verrons cela détail dans d'autres didacticiels.

## 3.5 Analyse automatique

Il nous reste maintenant à construire l'arbre de décision à partir de la méthode sélectionnée. Nous activons le menu ANALYSIS / LEARNING. A partir de l'arbre de décision (Figure 1), nous obtenons les règles suivantes :

| Règle | Confiance | Lift | Support |
|---|---|---|---|
| Si overweight >= 2.5 Alors Blood pressure = high | 71% | 1.25 | 40%[2] |
| Si overweight < 2.5 ET Exercice < 1.5 Alors Blood Pressure = high | 62% | 1.08 | 24% |
| Si overweight < 2.5 ET Exercice >= 1.5 Alors Blood pressure = low | 61% | 1.40 | 36% |

Une règle est de bonne qualité si elle est précise, avec une confiance élevée, et touche un nombre conséquent d'observations c.-à-d. avec un support élevé.

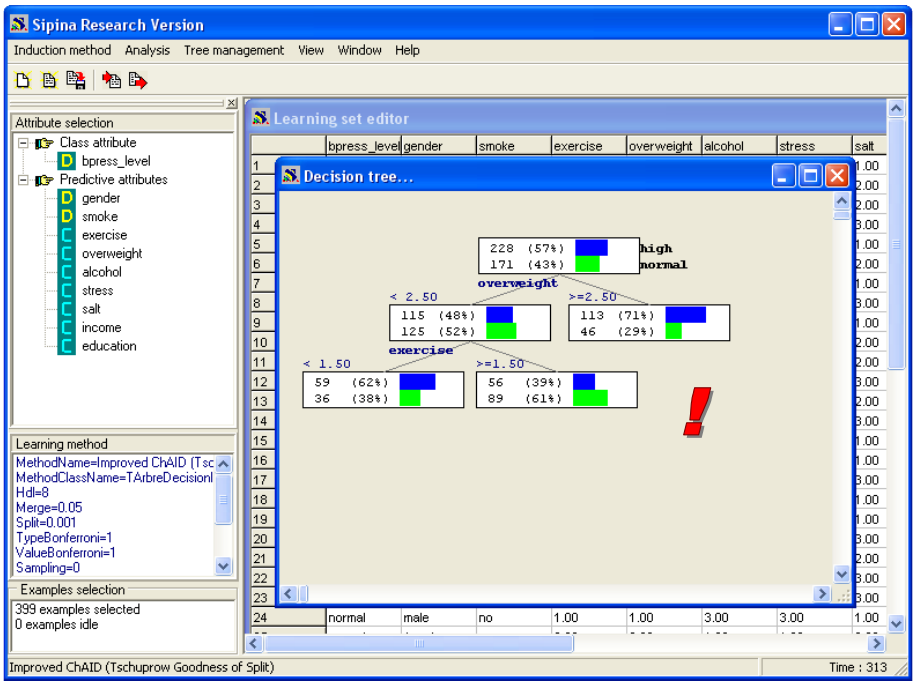



**Figure 1 - Arbre de décision sur les données "blood pressure"**

[2] La définition du support dans l'induction des règles en apprentissage supervisé est quelque peu différente de celle utilisée dans la recherche des règles d'association. Ici, le support est égal au nombre d'individus couverts par la règle divisé par l'effectif total c.-à-d. 40% = (113+46)/399.

La partie « SI » constitue la prémisse de la règle, la partie « ALORS », la conclusion.

Grosso modo, les personnes ayant un embonpoint souffrent également majoritairement d'hypertension, plus que dans la population globale en tous les cas. Chez les personnes minces, l'absence d'activité physique est également un facteur d'hypertension.

Pour évaluer pleinement l'apport informationnel d'une règle, on utilise l'indicateur LIFT qui est le rapport entre la probabilité conditionnelle de la règle (la confiance) et la probabilité marginale de la modalité de la conclusion. Dans le cas de la première règle, elle est égale à 1.40 = 71%/57%.

# 4 Exploration des solutions

Les experts jugeront de la qualité de ce modèle explicatif. Pour qu'ils puissent l'apprécier au mieux, nous avons la possibilité d'explorer en détail chaque nœud de l'arbre et les solutions alternatives que nous aurions pu mettre en place.

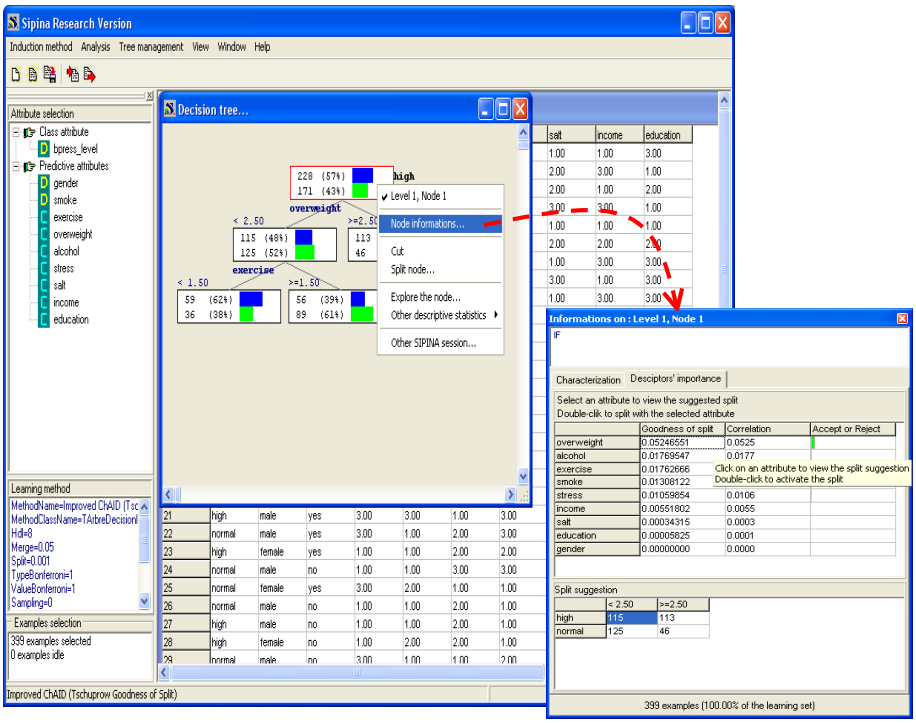
## 4.1 Segmentations alternatives

Première question que l'on peut invoquer : la variable de segmentation à la racine de l'arbre est « overweight », était-ce la seule solution possible ?

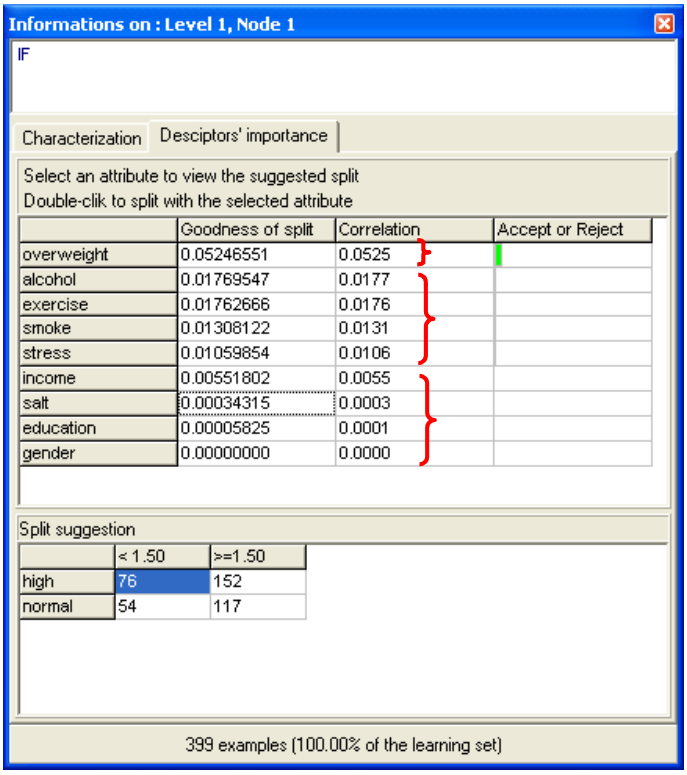
En effet, la technique choisit simplement la meilleure variable au sens d'un critère donné, au risque de masquer les autres variables candidates, alors qu'elles sont susceptibles d'apporter une qualité d'information quasi équivalente. Ce n'est absolument pas anodin. Si l'on choisit d'introduire une autre variable à la place de celle détectée par la méthode, il se peut très bien que les autres variables intervenant dans les parties basses de l'arbre soient complètement différentes. On aboutit à un système de règles (en apparence[3]) très différent.

Nous allons étudier les solutions alternatives à la variable « overweight » lors de la segmentation de la racine de l'arbre. Pour cela, nous effectuons un clic avec le bouton droit de la souris sur ce sommet. Dans le menu contextuel qui apparaît, nous activons l'option NODE INFORMATIONS…

[3] En apparence seulement car, même si les systèmes de règles semblent très différents en ne faisant pas intervenir les mêmes variables, il est très vraisemblable qu'ils classent les individus de la même manière.




Une fenêtre apparaît. Nous observons la liste des variables candidates et leurs apports respectifs si nous les avions introduites dans la segmentation. Cette méthode s'appuie sur le de T de TSCHUPROW pour caractériser une segmentation.

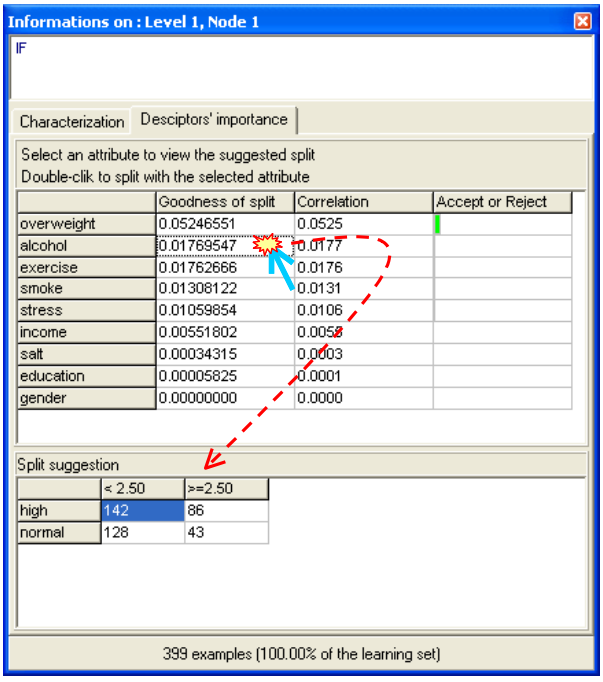



Nous observons principalement 3 groupes de variables :

1. OVERWEIGHT est sans conteste la variable la plus intéressante pour segmenter la racine. Dans la partie basse de la fenêtre apparaît le partitionnement associé, c'est le même que celui que nous observons dans la représentation graphique de l'arbre.
2. Ensuite vient un groupe de variables formé par ALCOHOL, EXERCISE, SMOKE et STRESS. Elles sont moins déterminantes que OVERWEIGHT, mais leur introduction dans la segmentation à ce stade n'est pas choquante, tout dépend de l'objectif de l'étude et des contraintes que veut introduire l'expert.
3. Enfin, nous avons un troisième groupe de variables constitué de INCOME, SALT, EDUCATION et GENDER. Il semble que ces variables, à ce stade, pèsent très peu pour détecter les personnes souffrant de l'hypertension.

On peut se demander quelle aurait été le partitionnement proposé si nous avions choisi une autre variable, la variable ALCOHOL par exemple. Tout simplement parce qu'elle nous semble importante, ou parce qu'on se positionne dans une campagne contre l'alcoolisme, etc. Bref, on souhaite s'appuyer – aussi – sur des critères métiers pour approfondir l'étude.

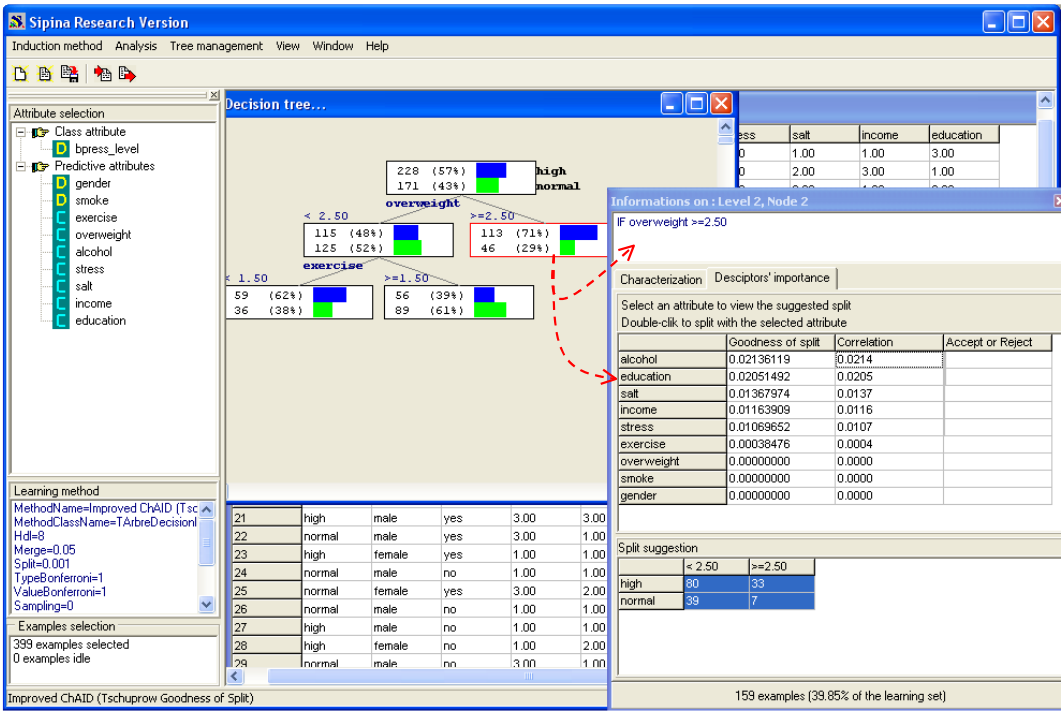
Dans un premier temps, il ne s'agit pas de remettre en cause l'arbre. Nous verrons plus loin ce qu'il en est dans l'induction interactive. Nous voulons seulement avoir une idée de ce que l'on pourrait obtenir. Pour cela, il faut cliquer sur la valeur GOODNESS OF SPLIT pour chaque variable. Dans le cas de ALCOHOL, nous activons la case correspondante, la partie basse de la fenêtre reflète la segmentation associée.




En termes de pureté des feuilles, la solution est bien entendu moins intéressante que la précédente, c'est ce que reflète le T de Tschuprow d'ailleurs. L'expert lui, avec ses connaissances, saura à même de nous dire si finalement cette alternative est pertinente ou pas.

## 4.2 Description d'un sommet de l'arbre

Penchons-nous maintenant sur le sommet au niveau suivant à droite, le groupe des personnes avec embonpoint (OVERWEIGHT >= 2.5). Nous cliquons sur le sommet en question sans avoir à fermer au préalable la fenêtre de description, les informations sont automatiquement actualisées (ou si vous avez fermé la fenêtre, vous cliquez de nouveau sur le menu contextuel NODE INFORMATIONS après avoir sélectionné le sommet).




Vient alors une question importante qui légitime pleinement l'analyse interactive dans l'induction des arbres : il est entendu que les personnes associées au sommet sont accortes et hypertendues, mais qu'en est-il des autres variables, quelles sont les autres caractéristiques de ces individus ?

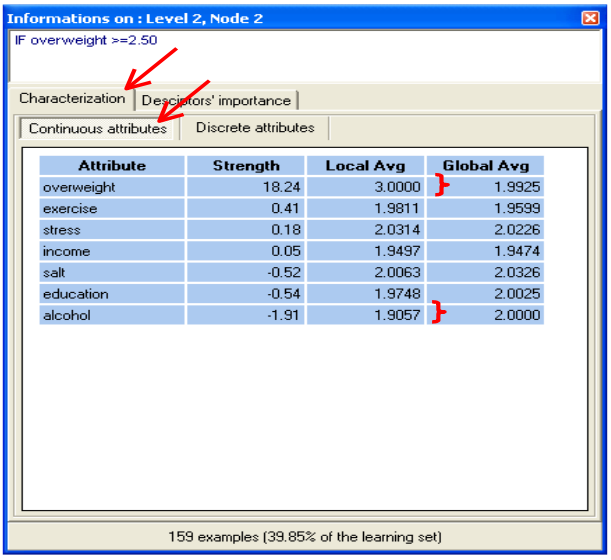
En opérant une sélection automatique, l'algorithme nous simplifie grandement la tâche, surtout lorsque le fichier comporte un très grand nombre de variables candidates. Mais ce faisant, elle masque malheureusement le rôle des autres variables, notamment lorsqu'il s'agit de caractériser les groupes d'individus couverts par une règle.

Le second onglet CHARACTERIZATION sert à caractériser les groupes associés à chaque nœud de l'arbre. SIPINA procède simplement à des statistiques comparatives entre le sommet initial de l'arbre, représentant l'ensemble de la population, et le sommet courant, représentant une sous population définie par la règle. Pour évaluer l'importance de l'écart, la valeur test[4] qui est la statistique du test de comparaison à un standard est mise en œuvre : une comparaison de moyennes lorsque la variable est continue, une comparaison de proportions lorsque la variable est

[4] Sur le calcul et la lecture de la valeur test, voir http://tutoriels-data-mining.blogspot.com/2008/04/interprter-la-valeur-test.html.

discrète. Ce n'est pas à proprement parler un test statistique puisque les échantillons ne sont pas indépendants, mais son intérêt et sa souplesse sont indéniables dans la pratique.

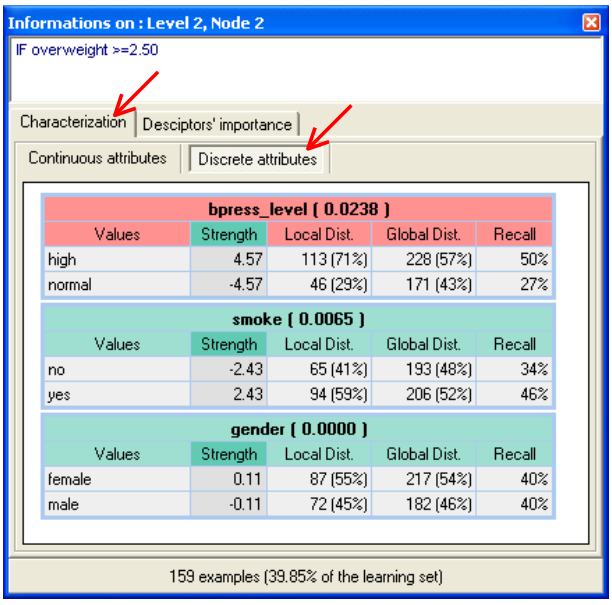
Nous activons l'onglet idoine. Nous obtenons dans un premier temps les statistiques comparatives pour les variables continues (quantitatives).




| Attribute | Strength | Local Avg | Global Avg |
|---|---|---|---|
| overweight | 18.24 | 3.0000 | 1.9925 |
| exercise | 0.41 | 1.9811 | 1.9599 |
| stress | 0.18 | 2.0314 | 2.0226 |
| income | 0.05 | 1.9497 | 1.9474 |
| salt | -0.52 | 2.0063 | 2.0326 |
| education | -0.54 | 1.9748 | 2.0025 |
| alcohol | -1.91 | 1.9057 | 2.0000 |

Les individus concernés ont effectivement un embonpoint certain (valeur test = +18.24), tous de niveau 3 puisque la moyenne sur ce groupe est de 3, c'est également le niveau maximum que l'on peut atteindre. Nous constatons également, ce n'était pas évident dans l'arbre, que ces personnes ont une consommation d'alcool moindre par rapport à la totalité de la population (valeur test = -1.91). Encore une fois, nous ne pouvons pas dire si l'écart est statistiquement significatif ou pas. Nous constatons en revanche qu'il est autrement plus fort pour cette variable par rapport aux autres variables de l'étude (excepté OVERWEIGHT bien entendu).

Pour les variables discrètes, nous obtenons la description suivante.




| bpress_level ( 0.0238 ) | | | | |
|---|---|---|---|---|
| Values | Strength | Local Dist. | Global Dist. | Recall |
| high | 4.57 | 113 (71%) | 228 (57%) | 50% |
| normal | -4.57 | 46 (29%) | 171 (43%) | 27% |

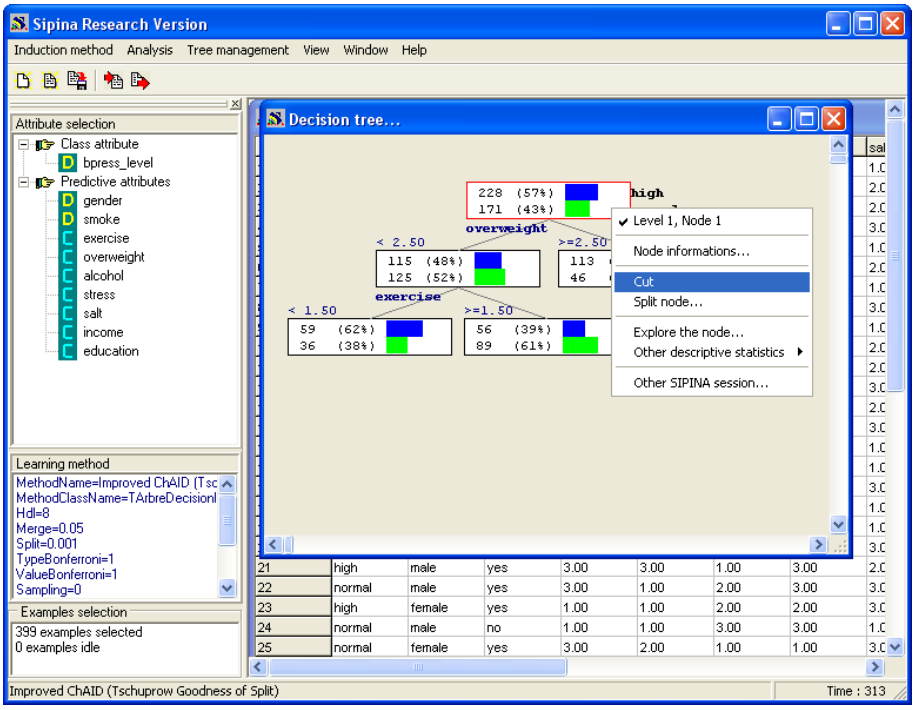
| smoke ( 0.0065 ) | | | | |
|---|---|---|---|---|
| Values | Strength | Local Dist. | Global Dist. | Recall |
| no | -2.43 | 65 (41%) | 193 (48%) | 34% |
| yes | 2.43 | 94 (59%) | 206 (52%) | 46% |

| gender ( 0.0000 ) | | | | |
|---|---|---|---|---|
| Values | Strength | Local Dist. | Global Dist. | Recall |
| female | 0.11 | 87 (55%) | 217 (54%) | 40% |
| male | -0.11 | 72 (45%) | 182 (46%) | 40% |

Les individus sur ce sommet souffrent majoritairement d'hypertension (BPRESS_LEVEL = HIGH), nous le savions à la lecture de l'arbre. En revanche, information moins évidente, la proportion des fumeurs est plus élevée dans ce groupe (59% contre 52% dans la population initiale). Rappelons d'ailleurs que la variable SMOKE était une variable crédible, en 4ème position, pour la segmentation du premier sommet. Le rôle de l'expert métier est très important à ce stade. Finalement, est-ce que le seul embonpoint est la cause de l'hypertension, ou bien masque-t-elle le rôle du tabac, ou est-ce la conjonction des deux ? Nous pouvons bien entendu mettre en œuvre des techniques statistiques pour répondre clairement à ces questions. En tous les cas, si nous nous étions contentés du modèle initialement proposé par l'arbre de décision, sans essayer d'approfondir les résultats, nous serions passés à côté de ces questions.

## 4.3 Intervention de l'utilisateur dans la construction de l'arbre

La possibilité pour l'utilisateur de guider à son gré l'exploration des solutions est certainement un des aspects les plus séduisants des arbres de décision. Reprenons la variable SMOKE qui apparaît finalement comme très importante. Sans un être un grand spécialiste de la médecine, il suffit de lire ce qu'il y a écrit sur les paquets de cigarettes pour savoir que le tabac ne fait certainement pas du bien à nos artères. Nous décidons de l'introduire comme première variable de segmentation dans notre analyse, pour étudier ensuite le rôle des autres variables. Nous revenons sur la racine de l'arbre. Pour l'élaguer manuellement, nous effectuons un clic avec le bouton droit de la souris, dans le menu contextuel nous sélectionnons l'option CUT.




Avec le menu contextuel, nous activons l'option SPLIT NODE, nous aboutissons dans la même fenêtre que pour l'option NODE INFORMATIONS. Nous sélectionnons l'onglet DESCRIPTORS' IMPORTANCE. Dans la liste des descripteurs candidats, nous sélectionnons la variable SMOKE (Figure 2).

Pour introduire une segmentation, il suffit de double cliquer sur la case contenant la valeur de GOODNESS OF SPLIT pour la variable choisie. Dans ce cas, le partitionnement est effectué même si les conditions d'acceptation ne sont pas réunies (Figure 3).

Nous constatons qu'il y a effectivement sur représentation des hypertendus chez les fumeurs (63% contre 57% dans la population globale), les proportions sont équilibrées en ce qui concerne les non-fumeurs, il a y presque autant d'hypertendus que de non hypertendus.

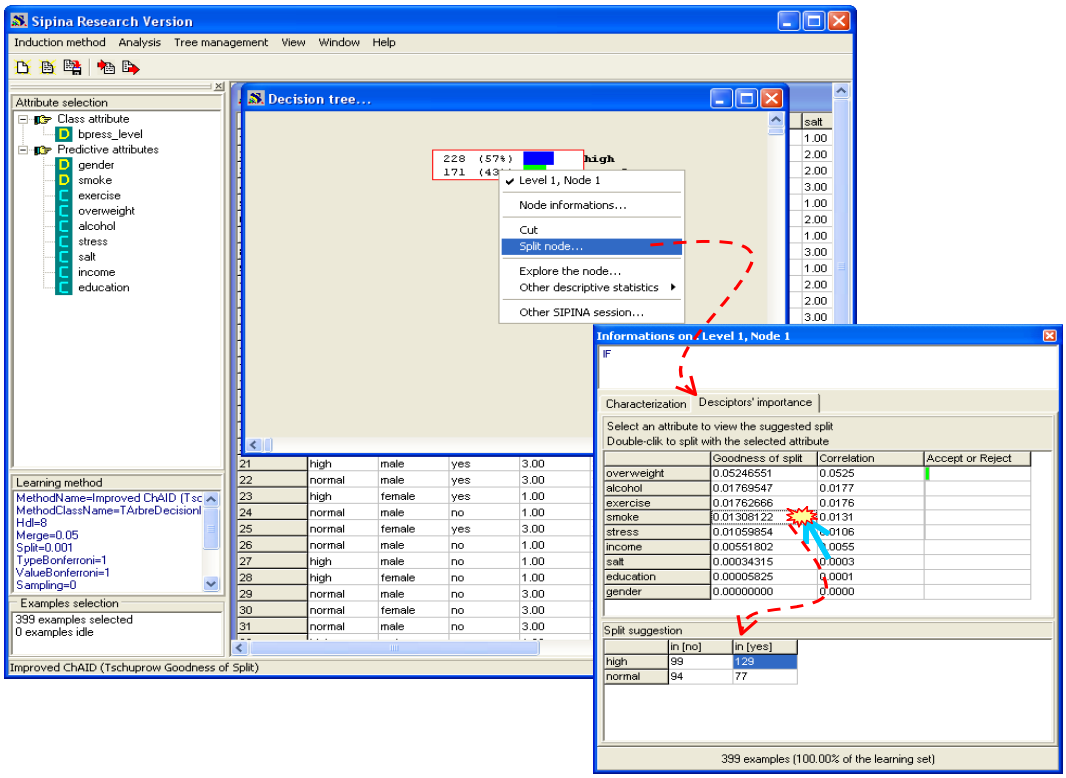



**Figure 2 - Prévisualisation de la segmentation à l'aide de la variable SMOKE**

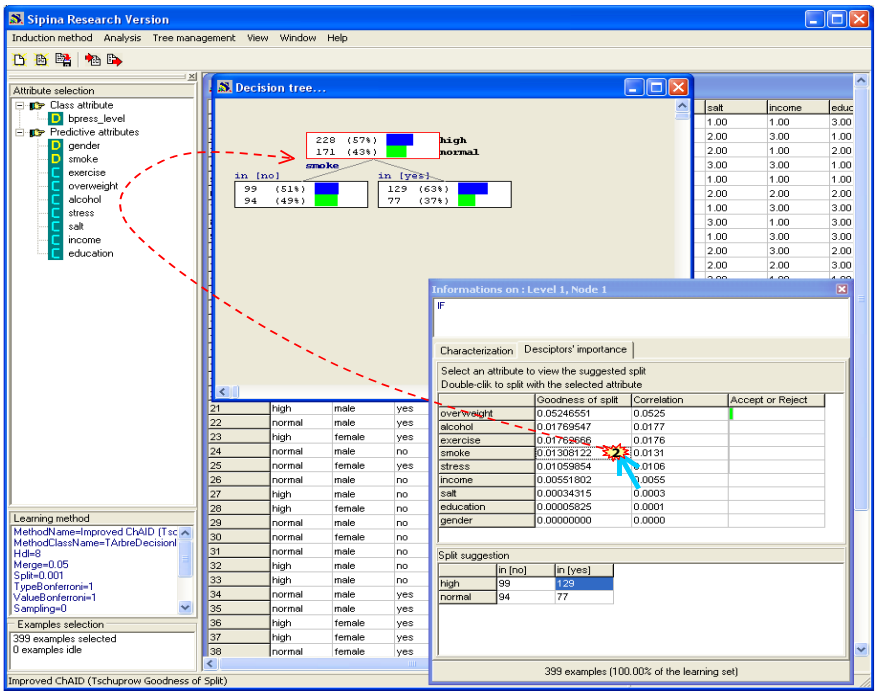
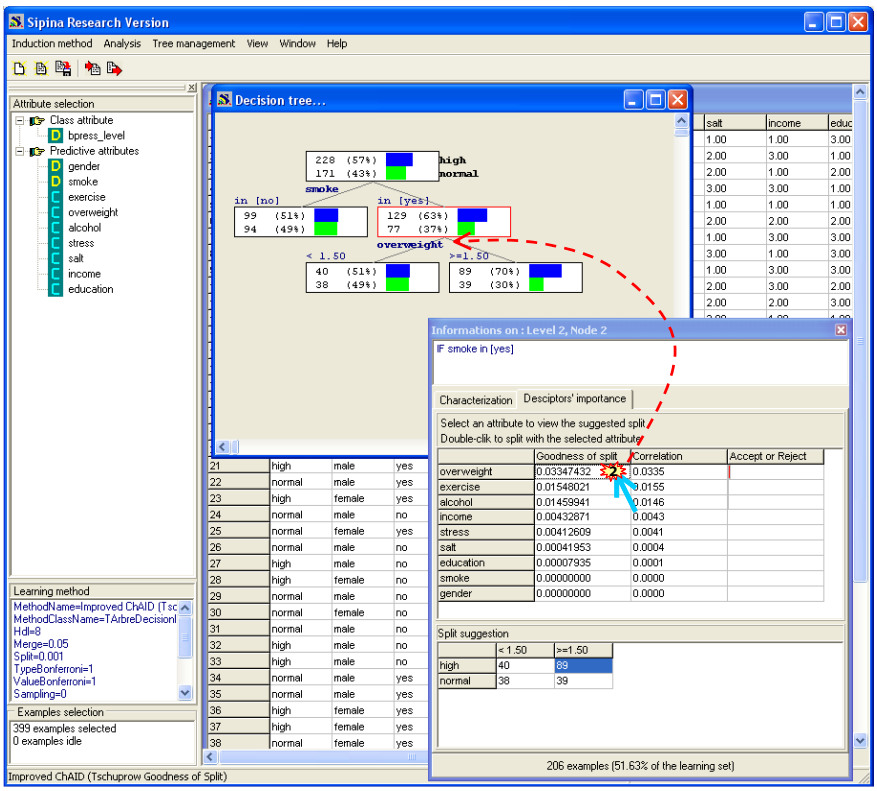



**Figure 3 - Segmentation avec la variable SMOKE**

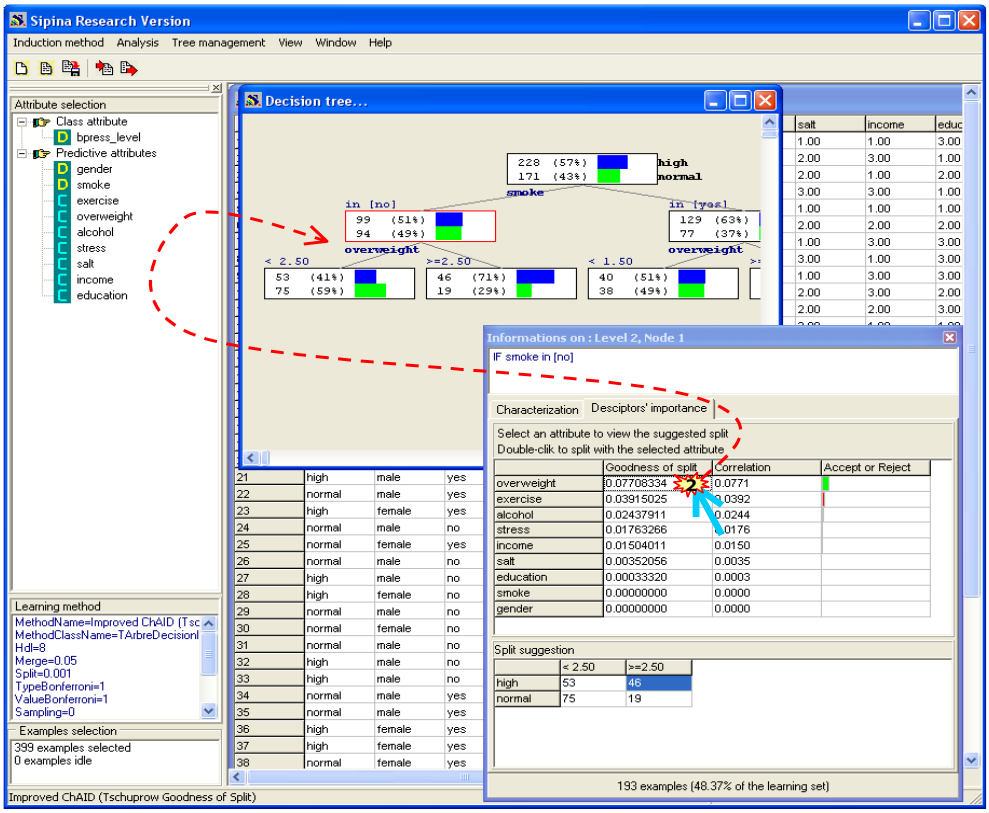
A partir de maintenant, il nous revient de construire complètement l'arbre en accord avec nos connaissances, nos intuitions et… les indicateurs statistiques que nous fournit le logiciel.

**Poursuivons l'exploration.** Sélectionnons le groupe des fumeurs à droite, sur le second niveau de l'arbre. Observons les variables candidates à la segmentation. Le surpoids (overweight) apparaît de nouveau en première position. Nous double cliquons pour introduire le découpage.

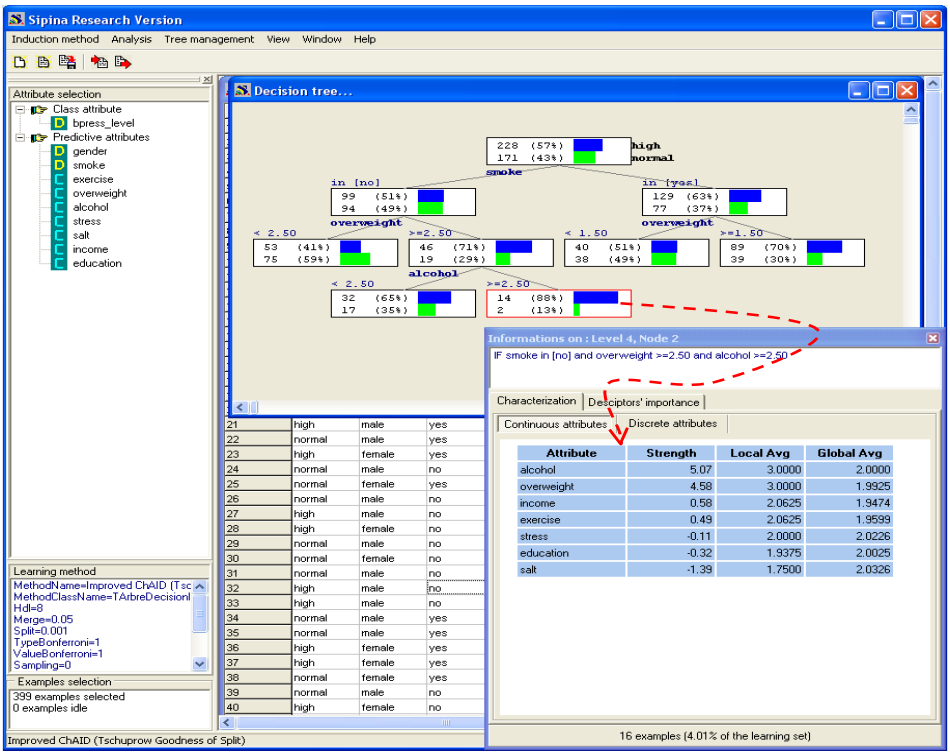


Les hypertendus sont sur représentés chez les fumeurs en surpoids. Mais à la différence de la segmentation sur l'ensemble de l'échantillon (sur la racine de l'arbre, Figure 1), le seuil de discrétisation (de découpage) de la variable OVERWEIGHT est différent : il était de 2.5 dans la population globale, il faut avoir un surpoids sévère pour être hypertendu, alors qu'il est de 1.5 chez les fumeurs, une surcharge pondérale moyenne a des conséquences néfaste sur la tension artérielle chez les fumeurs.

Voyons ce qu'il en est chez les non-fumeurs. Nous sélectionnons le sommet à gauche sur le second niveau de l'arbre. La meilleure variable explicative est encore le surpoids. Nous effectuons l'opération, les personnes en surpoids souffrent plus de l'hypertension (71%). Le seuil de découpage est en revanche plus élevé, nous retrouvons la valeur 2.5 : il faut une surcharge pondérale sévère pour que l'hypertension survienne chez les non-fumeurs.

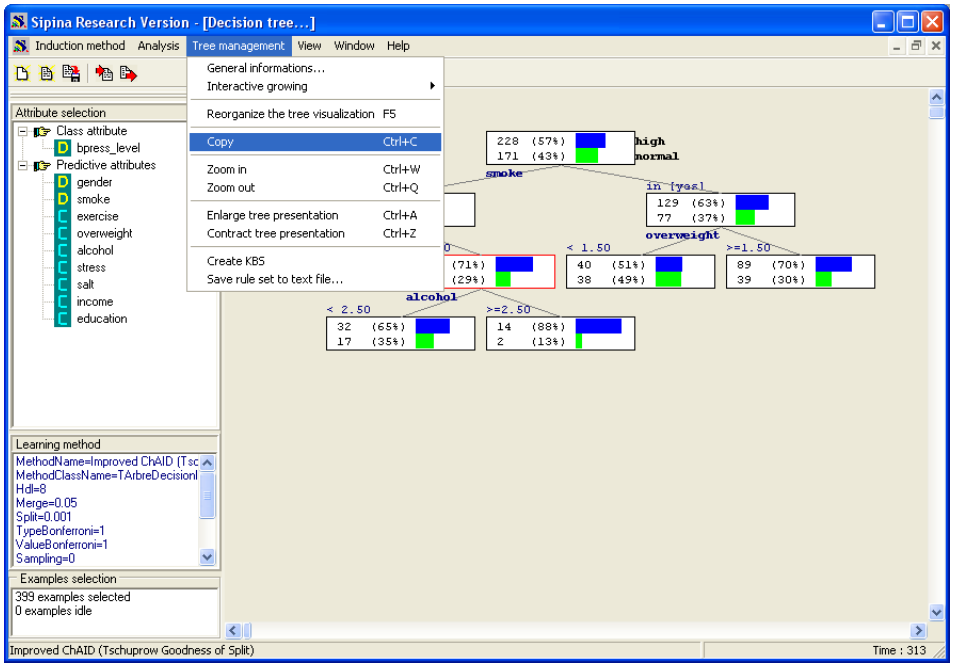



Si nous poursuivons l'analyse, nous pouvons explorer différentes configurations. Parmi les solutions possibles, nous pouvons produire la suivante :




Les personnes non-fumeur, en surcharge pondérale et qui abusent de l'alcool, ça ne va pas du tout. On y retrouve 14/228 ~ 6% des hypertendus, et ils sont fortement majoritaires dans ce groupe (88%).

Même si SIPINA est loin de disposer de toutes les fonctionnalités d'édition des logiciels commerciaux (SPAD Interactive Decision Tree, SPSS Answer Tree, SAS EM, etc.), il propose quand même quelques options destinées à améliorer la présentation des résultats. Elles sont disponibles dans le menu TREE MANAGEMENT. Nous pouvons, entre autres, copier le dessin de l'arbre pour le coller dans un traitement de texte ou un logiciel de présentation.



# 5 Conclusion

L'induction des arbres de décision ne constitue certainement pas une technique récente. Elle était largement connue et mise en œuvre dans les années 1970 dans le domaine du marketing en France (voir par exemple la méthode E.L.I.S.E.E. « Exploration des Liaisons et Interactions par Segmentation d'un Ensemble Expérimental » de CELLARD et AL. (1967), très populaire en son temps[5]).

Vers le début des années 1990, il y eu un double renouveau : **Côté science**, avec les travaux de Quinlan qui a littéralement inondé la planète Machine Learning de publications sur la méthode ID3 et ses dérivés dont le fameux C4.5 à la fin des années 1980, mais aussi avec de très nombreuses thèses remarquables comme celles de Wehenkel (1990), Buntine (1992), Murthy (1995), et tant d'autres… qui ont permis de circonscrire au mieux les caractéristiques de la technique ; **Côté logiciels**, lorsque les ressources en matière de programmation des interfaces graphiques ont été suffisamment performantes pour que l'on puisse facilement programmer des logiciels « user-friendly », donnant à utilisateur la faculté d'interagir efficacement avec l'outil et guider à sa convenance la découverte des connaissances.

Les possibilités d'interprétation et d'exploration que nous avons présentés dans ce didacticiel, qui restent quand même très limités dans SIPINA car il est destiné avant tout à la recherche, font en grande partie la renommée de cette méthode auprès du grand public. De nos jours, on imagine très mal un logiciel commercial de Data Mining sans un arbre de décision graphique et interactif. Côté logiciels libres, l'offre est très limitée. A part SIPINA, disponible sur Internet depuis 1995, ORANGE Machine Learning (https://orange.biolab.si) est un des rares outils à proposer des fonctionnalités qui s'en rapprochent (https://docs.biolab.si/2/widgets/rst/classify/interactivetreebuilder.html).

[5] NDA : Merci, merci, internet, et merci aux éditeurs de la **Revue de Statistique Appliquée**, de nous donner accès à des documents dont on ne soupçonne même pas l'existence, et pourtant, je pensais m'y connaître quand même un peu s'agissant des arbres de décision. Cette référence est réellement instructive sur les connaissances et les préoccupations de l'époque concernant les arbres de segmentation. Des préoccupations guères éloignées de celles des années 1990 finalement, lorsque les publications sur les arbres monopolisaient les actes des conférences Machine Learning, et plus tard Data Mining : A. BACCINI, A. POUSSE, « Segmentation aux moindres carrés : un aspect synthétique », Revue de Statistique Appliquée, N°3, pp. 17-35, 1975.

http://archive.numdam.org/ARCHIVE/RSA/RSA_1975__23_3/RSA_1975__23_3_17_0/RSA_1975__23_3_17_0.pdf