Statistiques descriptives avec SIPINA

1 Objectif

Montrer les outils de « statistiques descriptives » de SIPINA.

SIPINA propose des fonctionnalités de statistiques descriptives. Peu de personnes le savent. En soi, l'information n'est pas éblouissante, il existe un grand nombre de logiciels libres capables de produire les indicateurs de la description statistique.

L'affaire devient plus intéressante lorsque l'on couple ces outils avec l'induction d'un arbre de décision. La richesse de la phase exploratoire est décuplée. En effet, chaque nœud d'un arbre correspond à une sous population décrite par une règle. Ce groupe a été constitué de manière à ce que seule une des modalités de la variable à prédire soit représentée. C'est l'objectif de l'apprentissage. Mais qu'en est-il des autres variables ?

L'arbre a une qualité rare, il met en avant les meilleures variables dans l'induction. Mais il a le défaut de ses qualités, il ne donne pas directement d'informations sur les variables qui ont été écartées, encore moins sur les relations entre ces variables. La possibilité de calculer simplement des statistiques descriptives sur les sous populations permet à l'utilisateur d'étudier finement les

spécificités de ces groupes, et par là même de mieux caractériser la règle produite par l'induction. C'est ce que nous essayions de mettre en valeur dans ce didacticiel.

2 Données

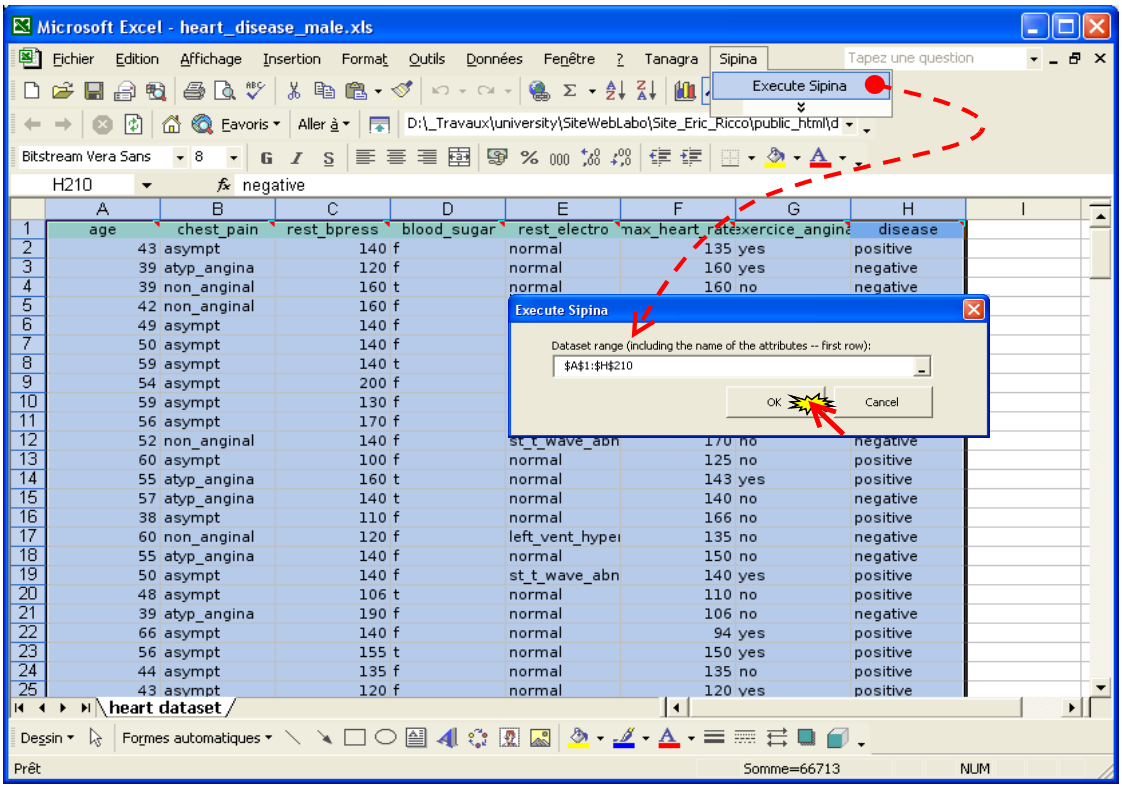
Nous utilisons les données HEART_DISEASE_MALE.XLS[1]. Il s'agit de prédire l'occurrence d'un maladie cardiaque (DISEASE) à partir des caractéristiques des individus (AGE, SUCRE dans le sang, etc.). Les données, 209 observations, sont restreintes aux individus de sexe masculin.

3 Statistiques descriptives

3.1 Charger les données dans SIPINA

Une macro complémentaire permet de faire la jonction entre EXCEL et SIPINA, il faut l'installer[2], un menu supplémentaire SIPINA doit apparaître dans le tableur.

Nous sélectionnons les données, puis nous activons le menu SIPINA / EXECUTE SIPINA. Une boîte de dialogue apparaît, demandant confirmation des coordonnées de la plage de cellules. Nous validons si elles sont correctes.

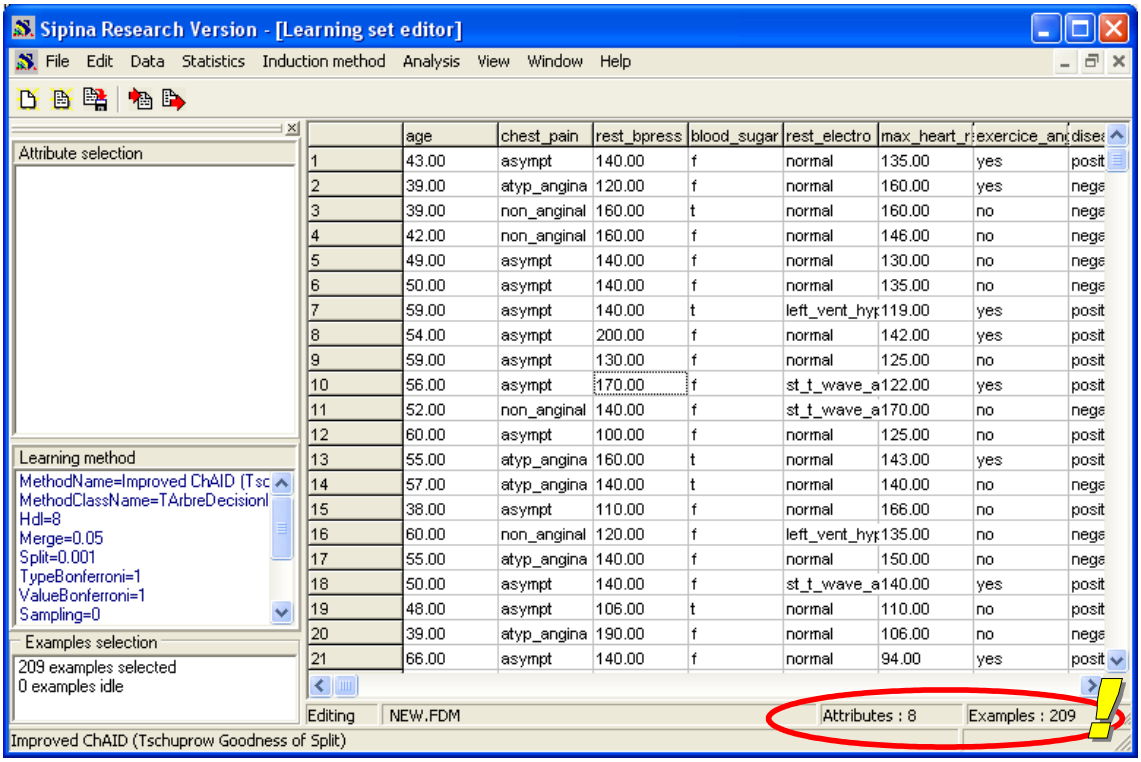


SIPINA est automatiquement démarré, les données transférées via le presse-papier. L'échantillon doit comporter 209 individus et 8 variables.

Le transfert des données est rapide. Par exemple, pour une base de 50000 observations et 15 variables, le traitement prend quelques secondes. Aucun fichier temporaire n'est copié sur le disque dur. Si nous souhaitons conserver une copie des données au format SIPINA (fichier *.FDM),

[1] http://eric.univ-lyon2.fr/~ricco/dataset/heart_disease_male.xls
[2] http://tutoriels-data-mining.blogspot.com/2008/03/connexion-excel-sipina.html

nous ferons FILE / SAVE AS. Le format FDM est binaire, SIPINA est le seul à pouvoir le manipuler. Les temps d'accès sont optimisés.



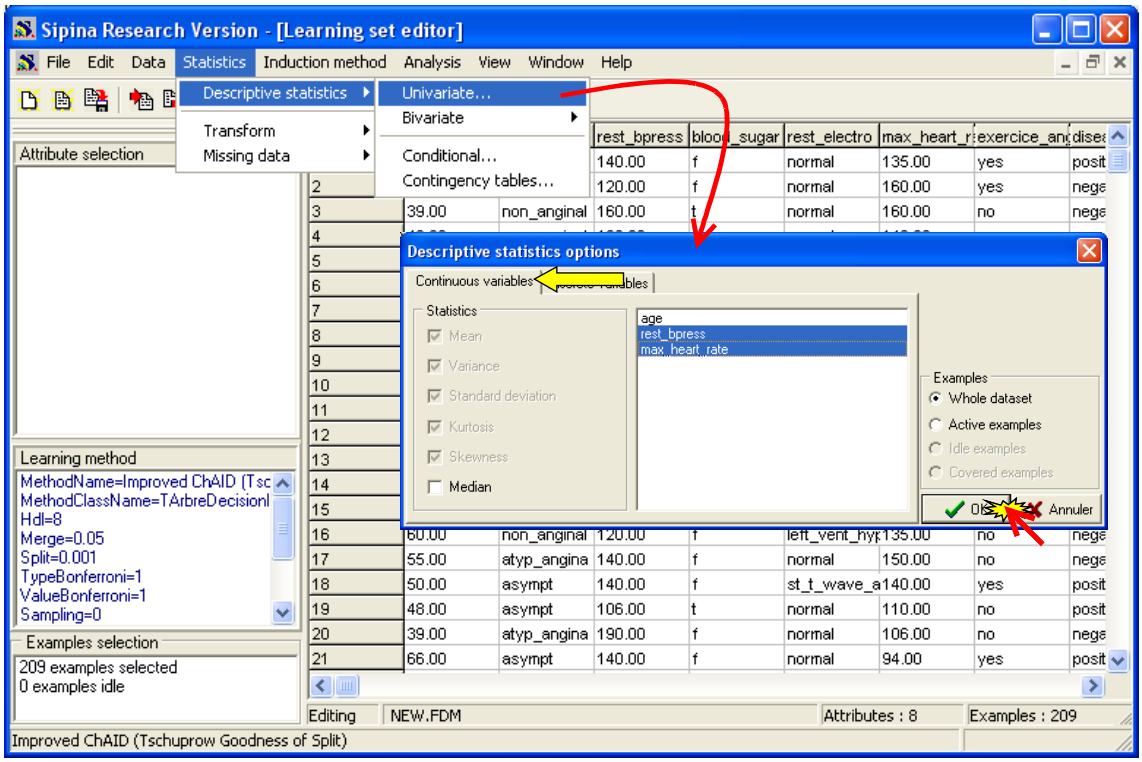
Lors du transfert, SIPINA type automatiquement les variables en utilisant la règle suivante : si la colonne est composée de valeurs numériques, il considère qu'elle représente une variable quantitative ; dans le cas contraire, il l'encode en variable catégorielle, les modalités sont les valeurs différentes trouvées dans la colonne. La première ligne des données est toujours censée représenter le nom de la variable, qu'elle soit alphanumérique ou numérique.

3.2 Statistiques univariées

Les outils de description statistique sont regroupés dans le **menu STATISTICS** de SIPINA. Attention, ce menu **n'est visible que si la grille des données est activée**. Dans le cas où nous sommes en train de visualiser une grille de résultats, pour accéder aux données il nous faudra revenir à la fenêtre des données (menu WINDOW / LEARNING SET EDITOR). C'est un peu contraignant. Ce sont les mystères des menus contextuels dans les applications MDI de Windows. Il faut le savoir simplement.

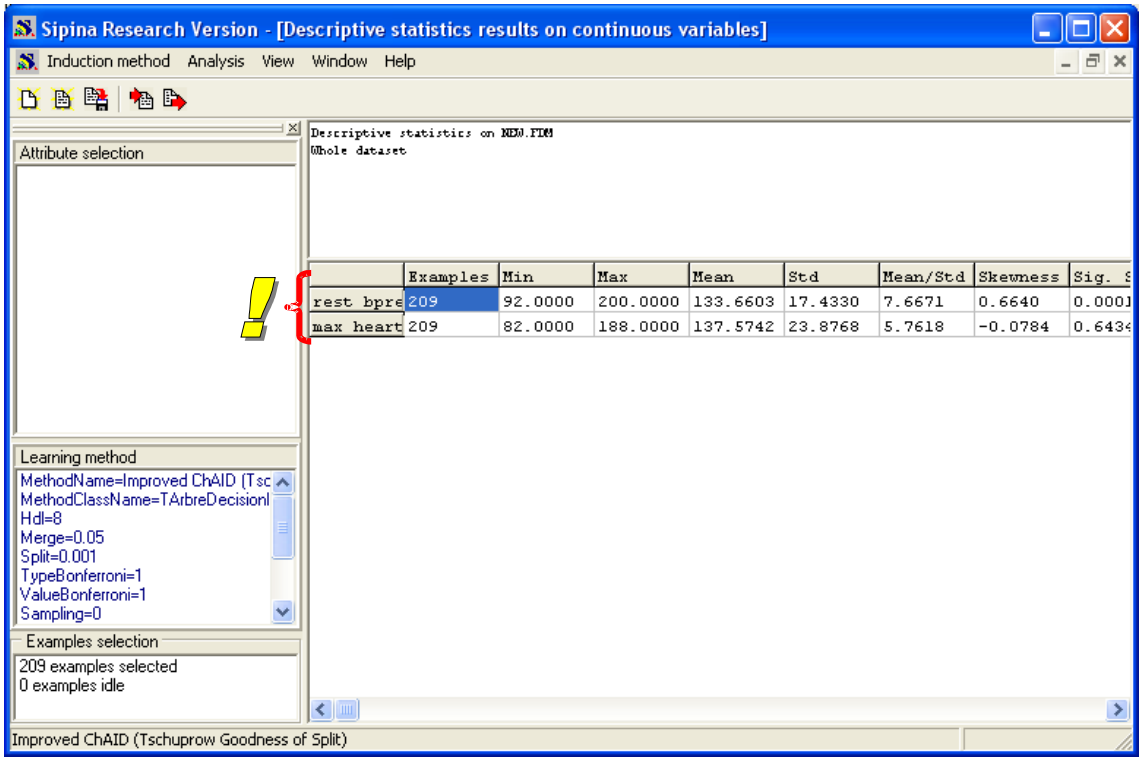
3.2.1 Variables continues

SIPINA calcule les indicateurs standards pour les variables numériques : moyenne, max, min, écart type, etc. Pour y accéder, nous cliquons sur le menu STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE. Une boîte de dialogue apparaît, nous activons l'onglet adéquat (CONTINUOUS VARIABLES), puis nous sélectionnons les variables à analyser, REST_BPRESS et MAX_HEART_RATE par exemple.



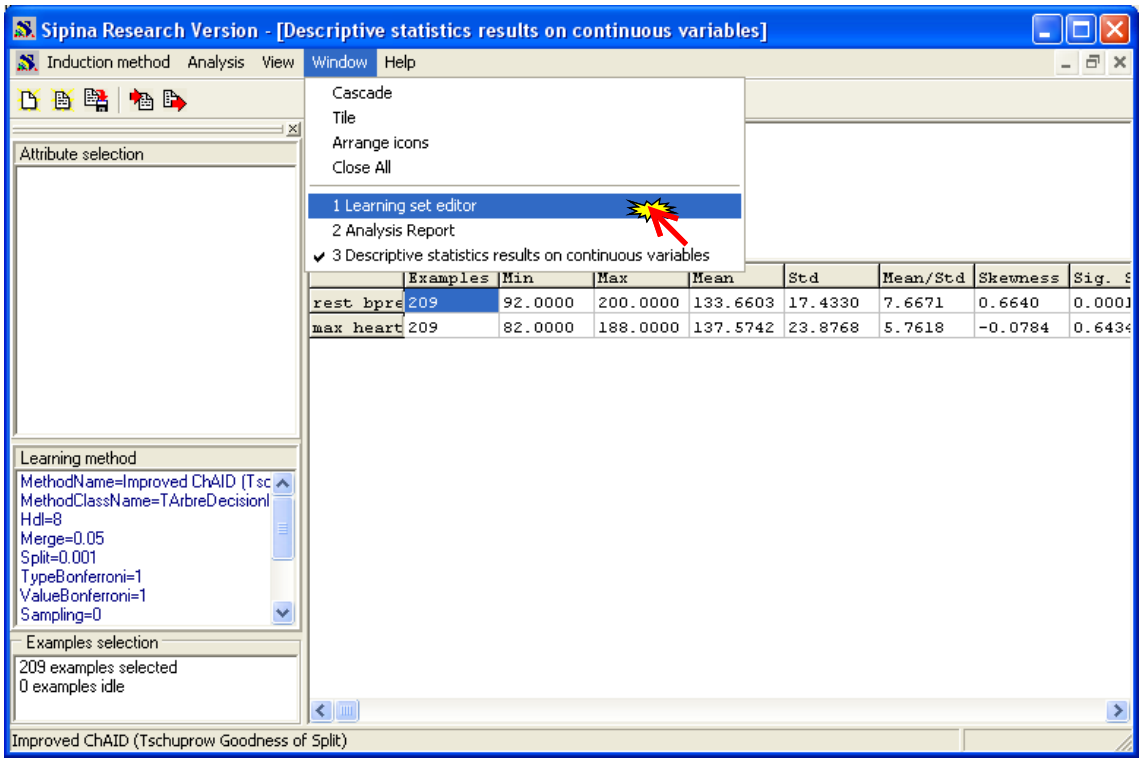
Nous remarquerons que nous pouvons restreindre les calculs aux individus actifs. Cela peut être utile si nous avons subdivisé les données en échantillon apprentissage et test dans le processus de fouille de données.

La fenêtre de résultats apparaît. Elle liste les variables dans une grille, chaque colonne correspondant à un indicateur statistique. Il est possible, via un menu contextuel, de copier le contenu de la grille de résultats dans un tableur, de modifier la précision de l'affichage, etc.

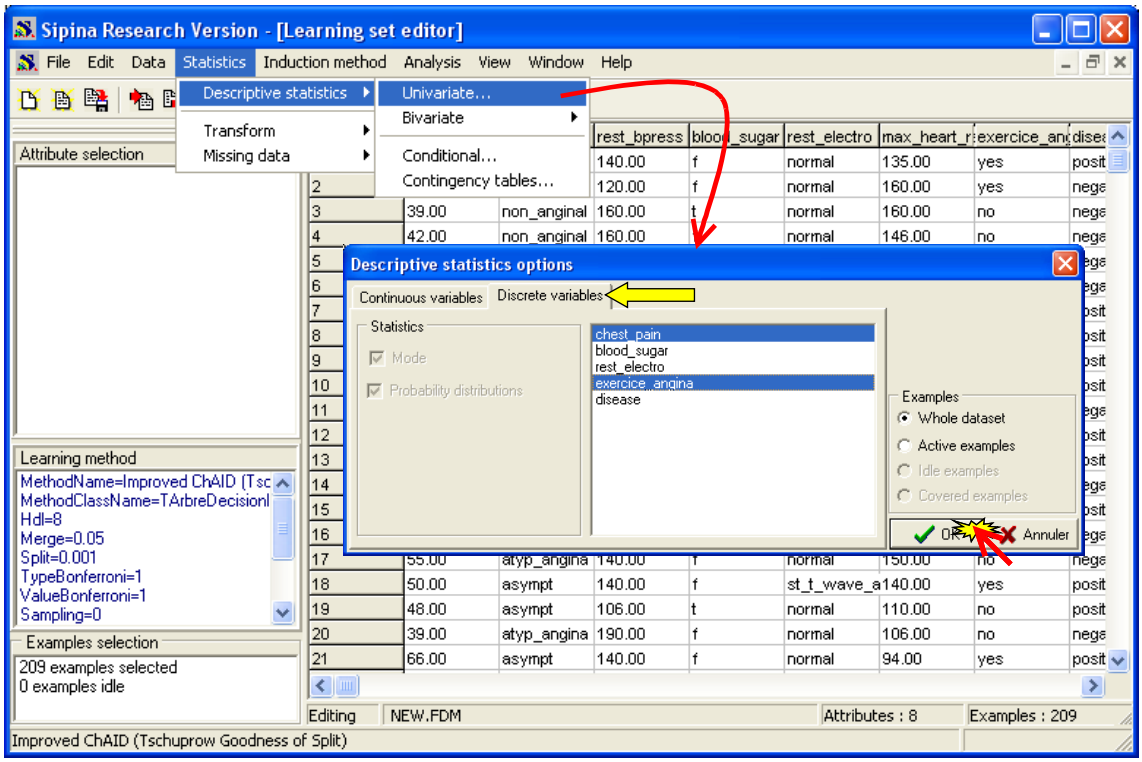


3.2.2 Variables catégorielles

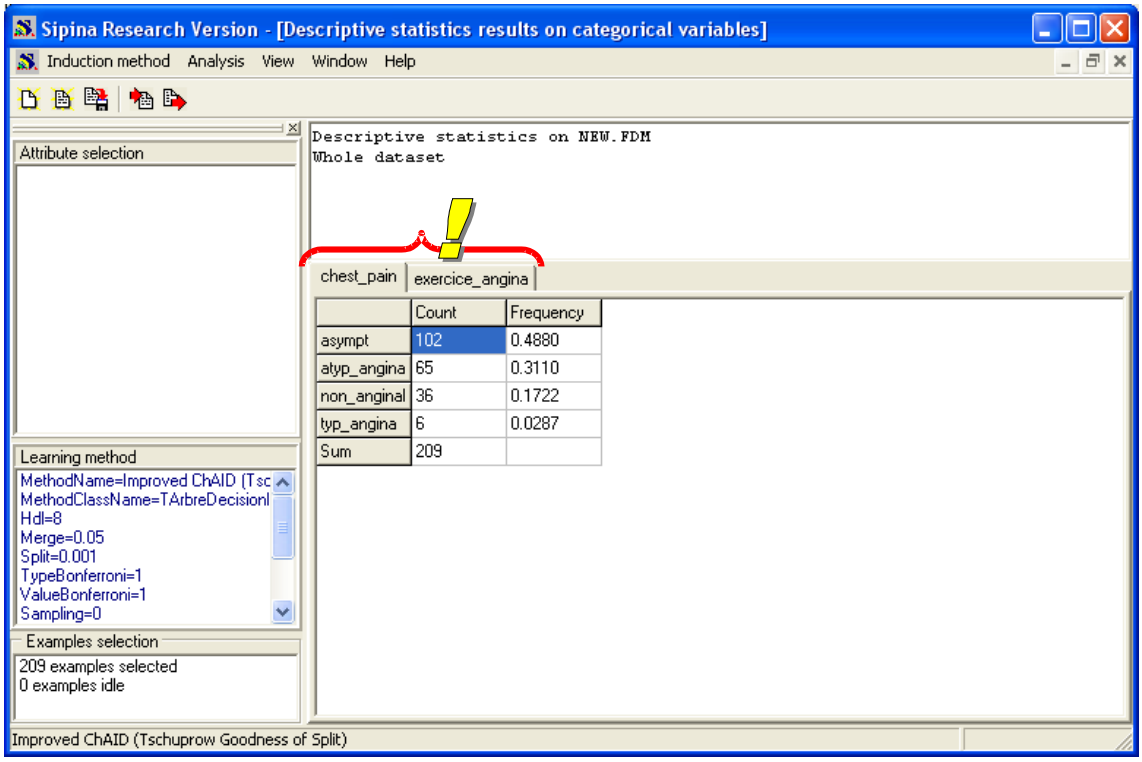
La démarche est la même concernant les variables catégorielles. Mais il faut auparavant revenir à la grille de données : nous activons l'item LEARNING SET EDITOR dans le menu WINDOW.



La fenêtre STATISTICS apparaît de nouveau, nous pouvons lancer les calculs en cliquant sur STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE. Nous optons pour l'onglet DISCRETE VARIABLES. Nous sélectionnons les variables CHEST_PAIN et EXERCICE_ANGINA.



Une nouvelle fenêtre de résultats est affichée. Nous y trouvons les distributions de fréquence des variables, répartis sur différents onglets.



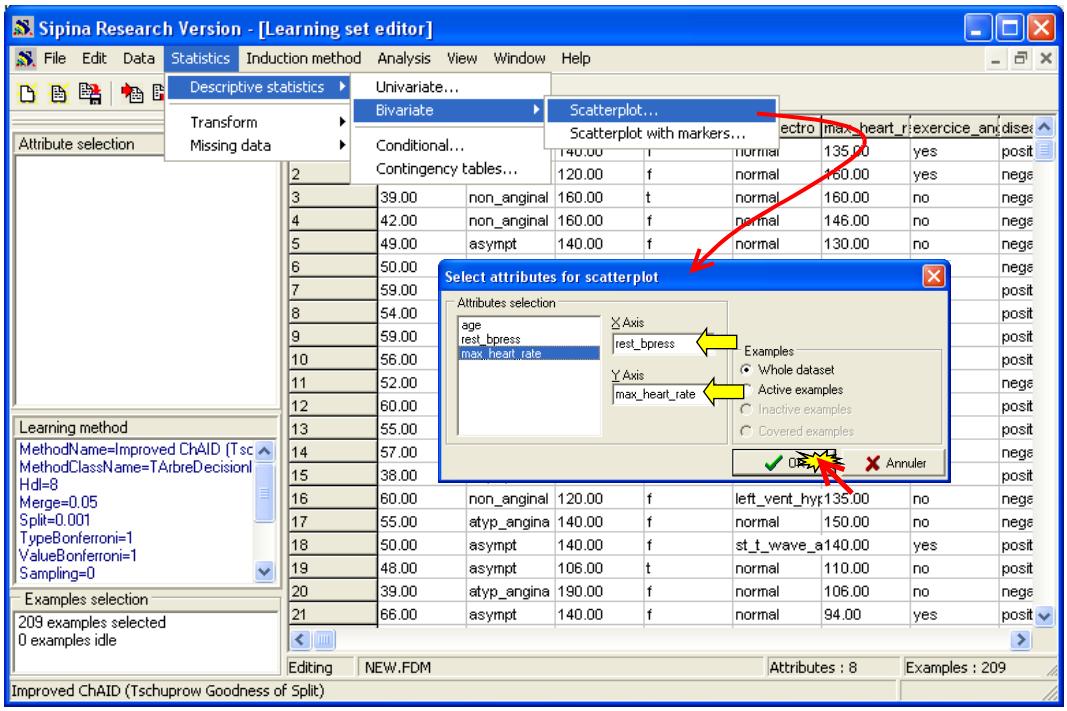
3.3 Statistiques bivariées

Nous pouvons construire des statistiques bivariées, en associant une variable continue à une variable catégorielle, deux variables catégorielles, ou deux variables continues. Bien entendu, il faut revenir à la fenêtre des données (WINDOW / LEARNING SET EDITOR) pour que le menu STATISTICS apparaisse.

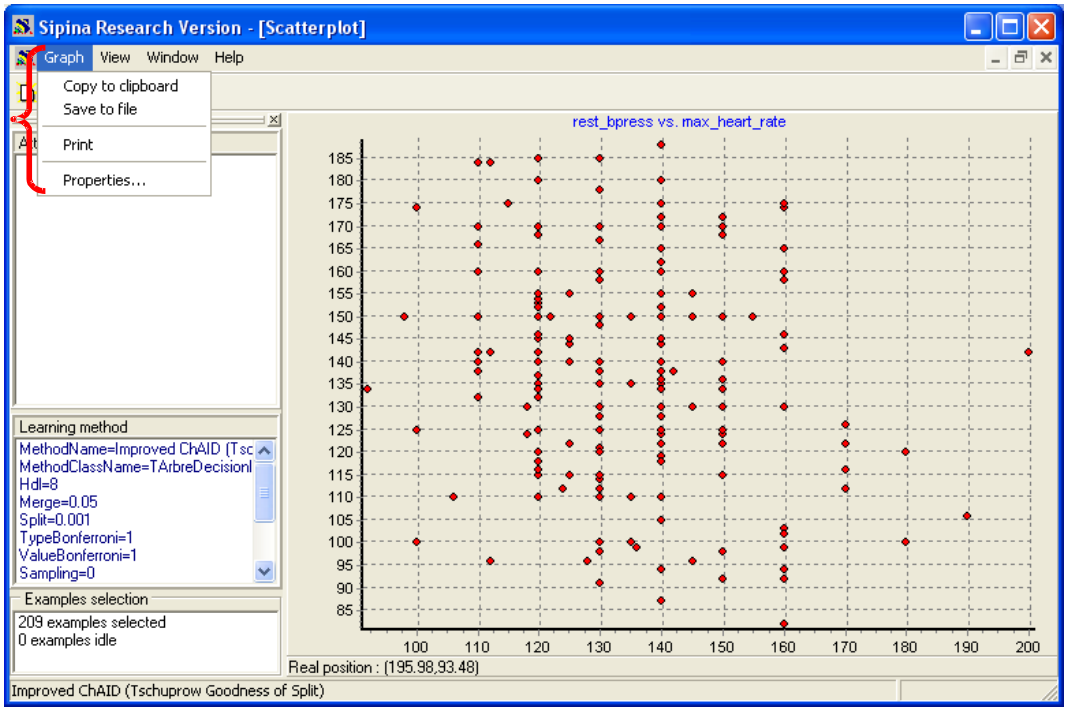
3.3.1 Deux variables continues : nuage de points

Représenter les points dans le plan est toujours riche d'enseignements, sur les variables (corrélations, etc.) et sur les individus (points atypiques, etc.).

Dans SIPINA, cette fonctionnalité est accessible dans le menu STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT. Une boîte de dialogue permet de choisir les variables à placer en abscisse et en ordonnée. Par glisser déposer, nous choisissons les variables REST_BPRESS et MAX_HEART_RATE.



La fenêtre de visualisation est créée. Dans le même temps, un nouveau menu GRAPH est maintenant disponible dans la barre de menu. Certaines options nous permettent de copier le graphique dans le presse-papier, l'imprimer, modifier la taille des points, etc.

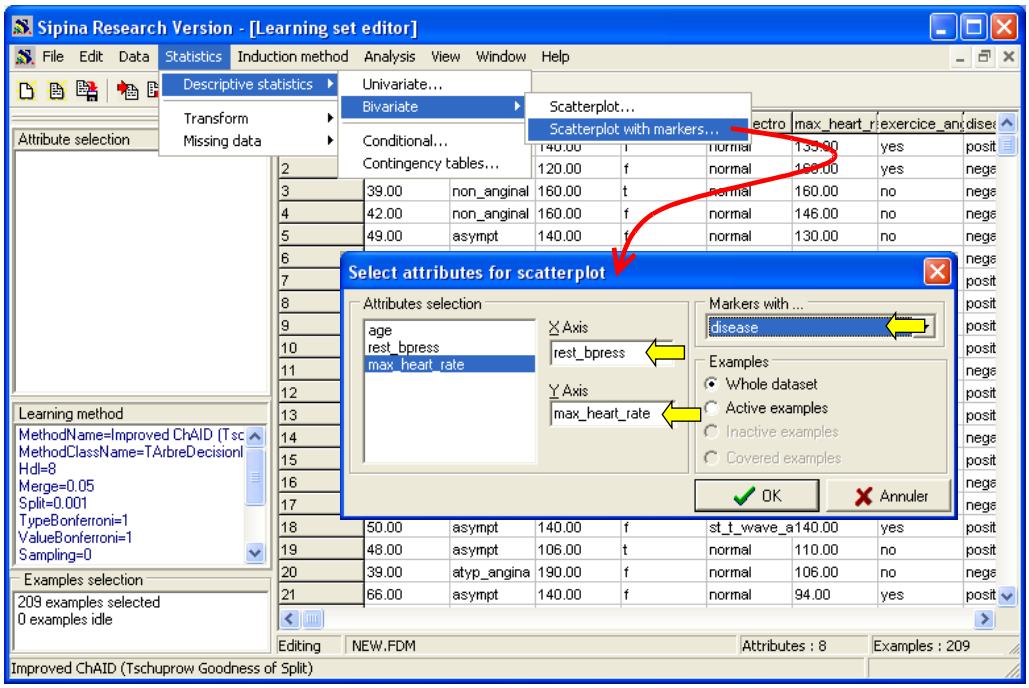


3.3.2 Deux variables continues : nuage de points conditionnel

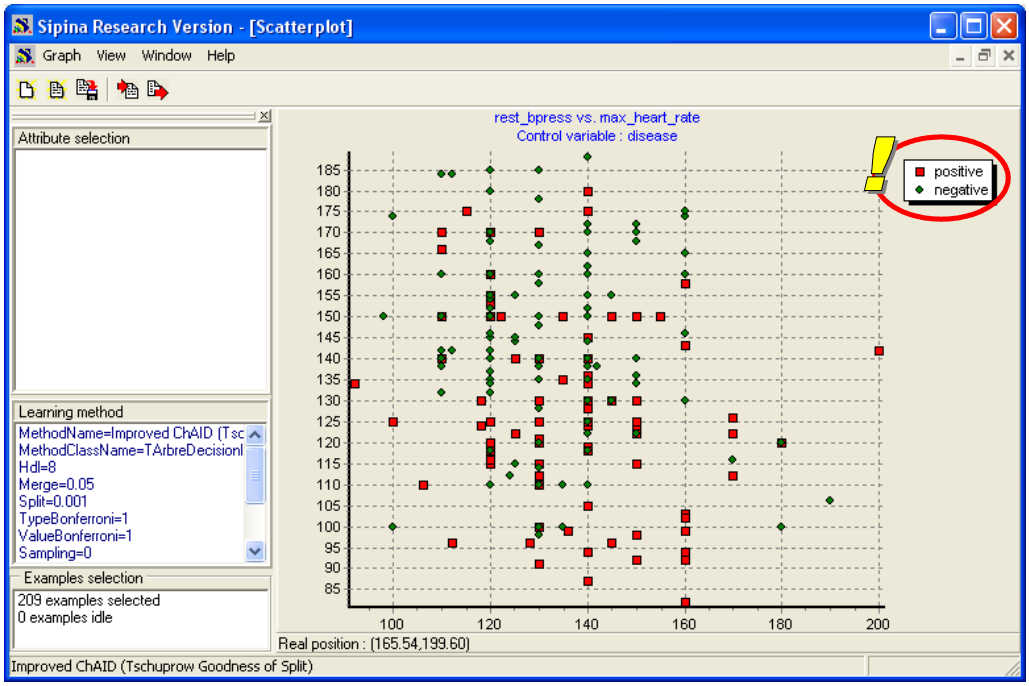
Un nuage de points est d'autant plus intéressant que l'on peut illustrer les points avec une troisième variable. Cela permet de mieux caractériser les positions relatives entre les individus vis à vis de cette tierce variable, voire d'en déduire des règles de classement. Dans notre cas, nous voulons situer les groupes d'individus selon l'occurrence de la maladie DISEASE. Pour ce faire, nous

activons la fenêtre des données (WINDOW / LEARNING SET EDITOR), puis le menu STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT WITH MARKERS[3].

Dans la boîte de paramétrage, nous remettons en abscisse REST_BPRESS, en ordonnée MAX_HEART_RATE, et en markers DISEASE .



Nous obtenons le même nuage de points que précédemment. A la différence que nous discernons maintenant les individus malades des individus sains.

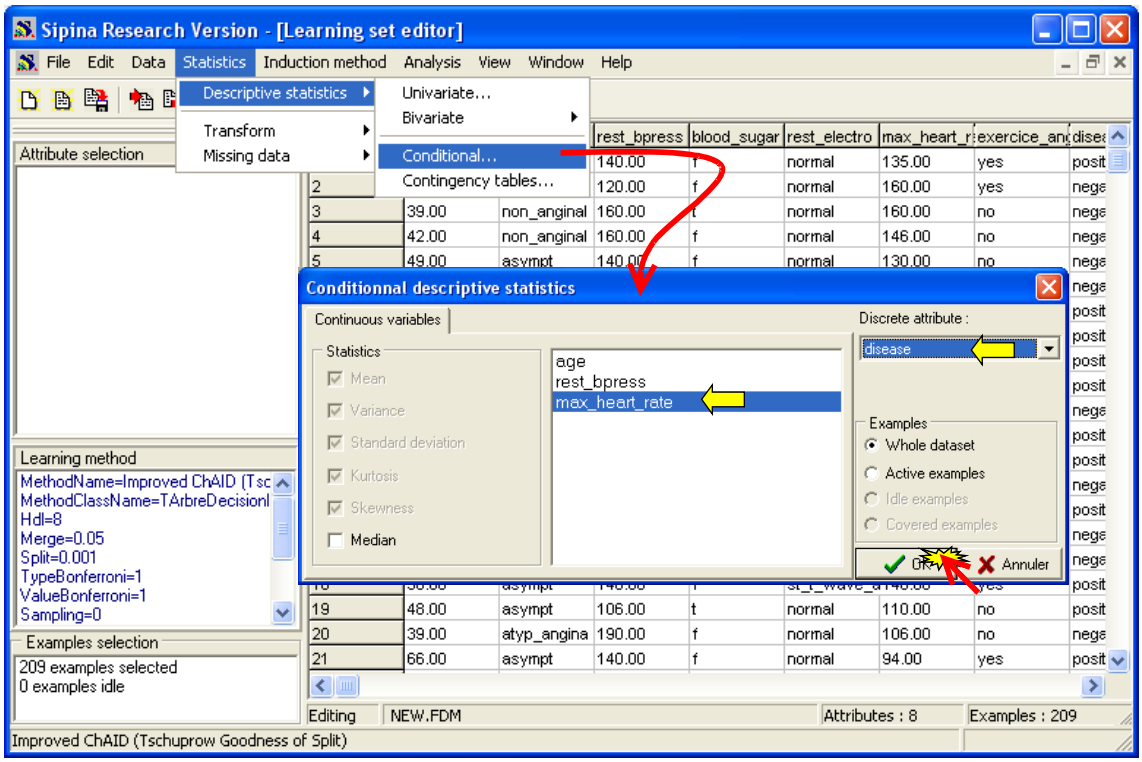


[3] Note : On comprendra maintenant pourquoi, lorsque j'ai fait les spécifications de TANAGRA, j'ai opté pour une simplification à outrance. Avoir à farfouiller constamment dans les menus de SIPINA, pour réaliser une opération somme toute banale, m'a dégoûté à jamais des menus imbriqués.

3.3.3 Variable continue vs. Variable discrète

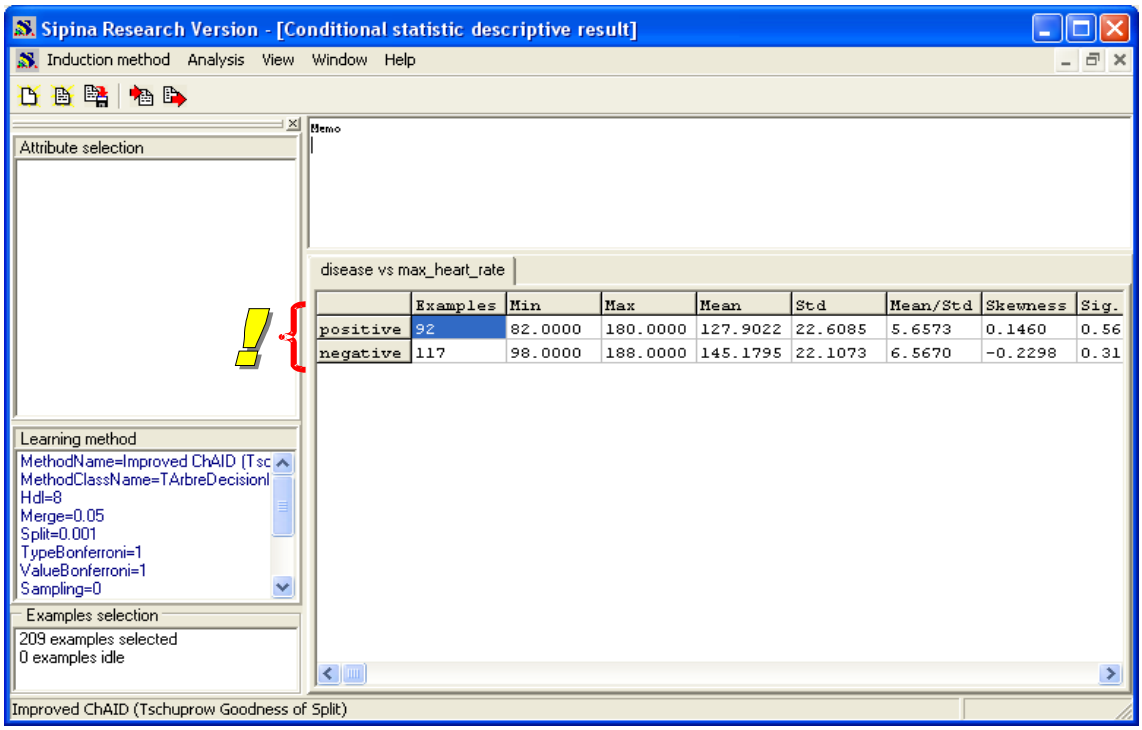
Autre outil de la statistique très simple et pourtant très riche d'enseignements, les distributions conditionnelles permettent de caractériser le comportement des variables continues dans les sous populations définies par une variable discrète. Nous voulons par exemple affiner l'analyse précédente en évaluant le comportement de MAX_HEART_RATE par rapport à DISEASE. Calculer les indicateurs standards (moyenne, etc.) conditionnellement aux valeurs prises par la variable caractérisante permet d'y répondre.

Revenons à la grille des données (WINDOWS / LEARNING SET EDITOR), puis activons le menu STATISTICS / DESCRIPTIVE STATISTICS / CONDITIONNAL. Dans la boîte de paramétrage, nous choisissons la variable MAX_HEART_RATE à caractériser, et la variable discrète caractérisante DISEASE.

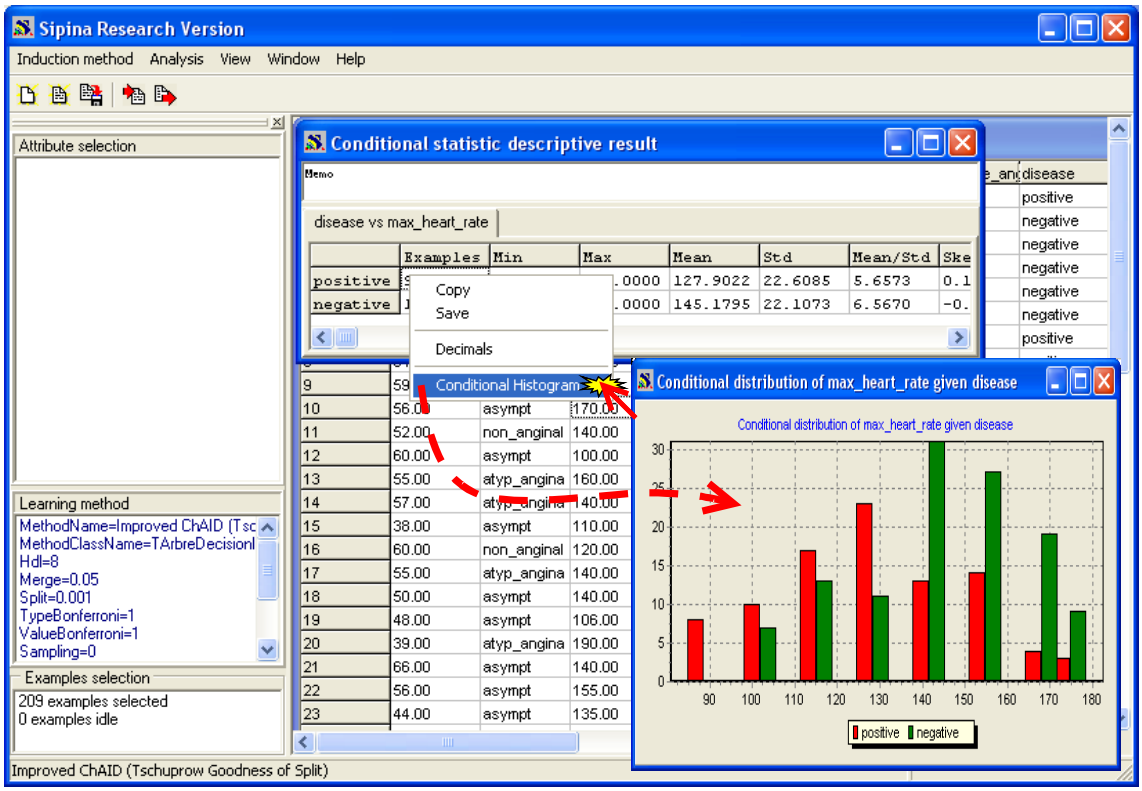


La grille de résultats nous donne les statistiques conditionnelles. Nous constatons par exemple que les individus DISEASE = POSITIVE (92 observations) présentent un MAX_HEART_RATE moyen (127.9022) plus faible que les DISEASE = NEGATIVE (117 observations), la moyenne dans ce dernier cas est de 145.1795. Le décalage des distributions se retrouve également au niveau des valeurs min et max conditionnelles.

En revanche, les dispersions dans les sous groupes sont identiques. L'écart type est quasiment le même chez les positifs et les négatifs.



En activant le menu contextuel de la grille (clic avec le bouton droit), nous avons accès à une option qui permet de construire les histogrammes conditionnels. L'information qui y est lue est plus riche que les simples ratios statistiques. Nous cliquons sur le menu CONDITIONNAL HISTOGRAM. Une nouvelle fenêtre[4] apparaît avec le graphique adéquat.

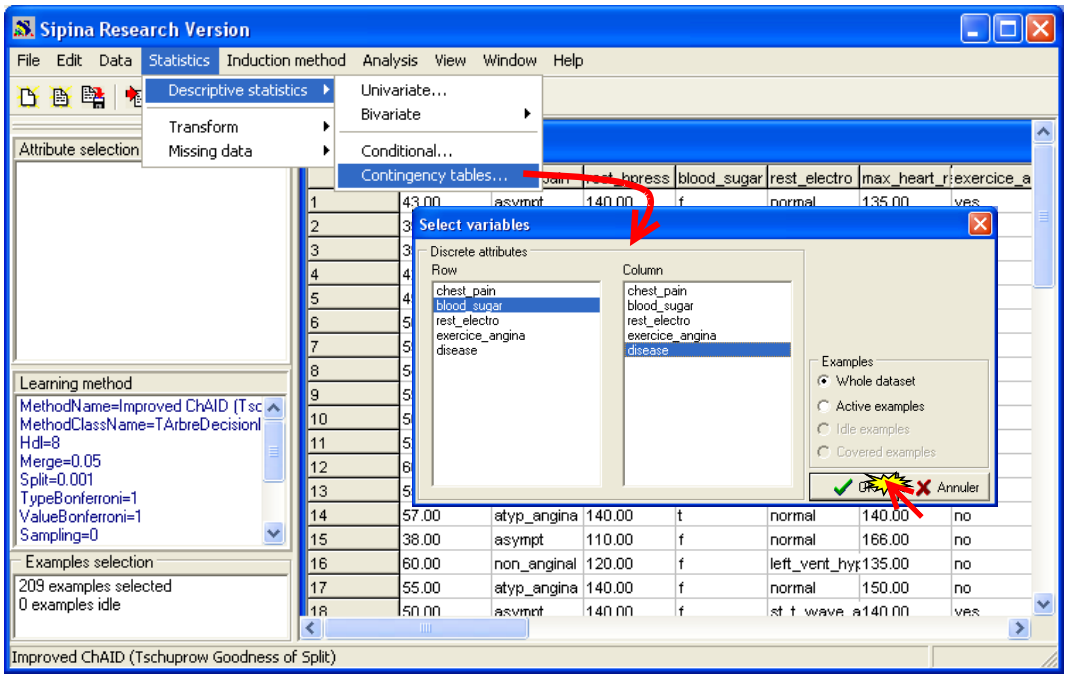


[4] Note : Pas besoin de chercher, ma religion s'appelait STATISTICA à l'époque. Plus il y avait de fenêtres dans mon application, plus j'avais l'impression de faire des choses exaltantes.

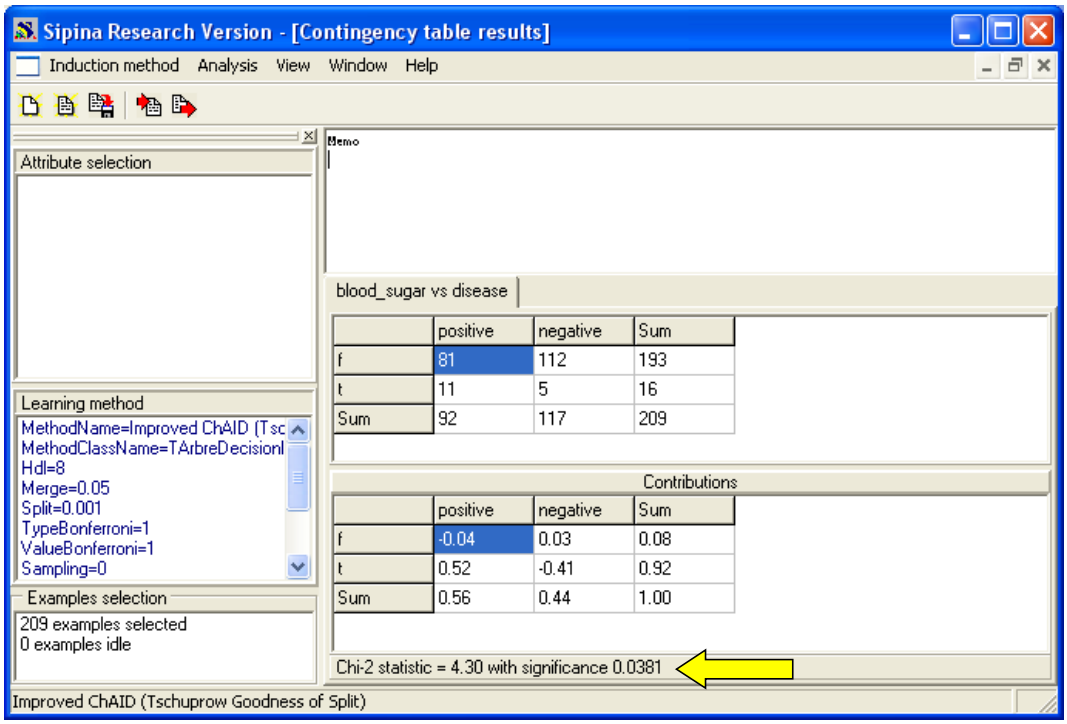
3.3.4 Deux variables discrètes : tableaux de contingence

Enfin, dernière fonctionnalité dans notre menu à tiroirs, le croisement entre 2 variables discrètes. L'idée est d'élaborer un tableau de contingence et de réaliser le test du KHI-2 d'indépendance.

Nous revenons toujours à la grille de données. Nous activons le menu STATISTICS / DESCRIPTIVE STATISTICS / CONTINGENCY TABLES. Nous voulons croiser dans cet exemple les variables BLOOD_SUGAR et DISEASE. Nous effectuons le paramétrage adéquat dans la boîte de dialogue.



Nous obtenons le tableau des effectifs, le tableau des contributions au KHI-2, la statistique du Khi-2 et la probabilité critique du test d'indépendance.

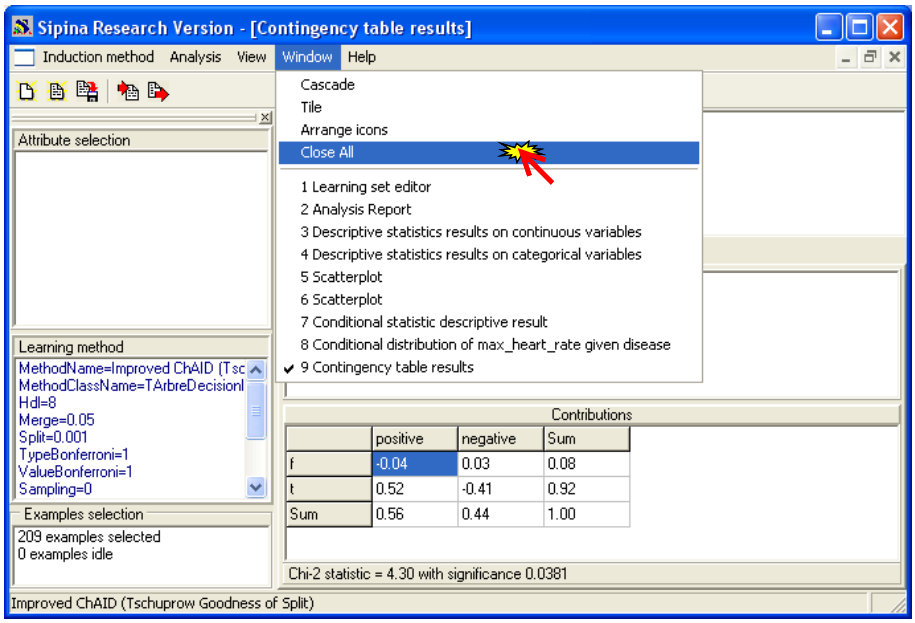


4 Statistiques descriptives associées à un nœud de l'arbre

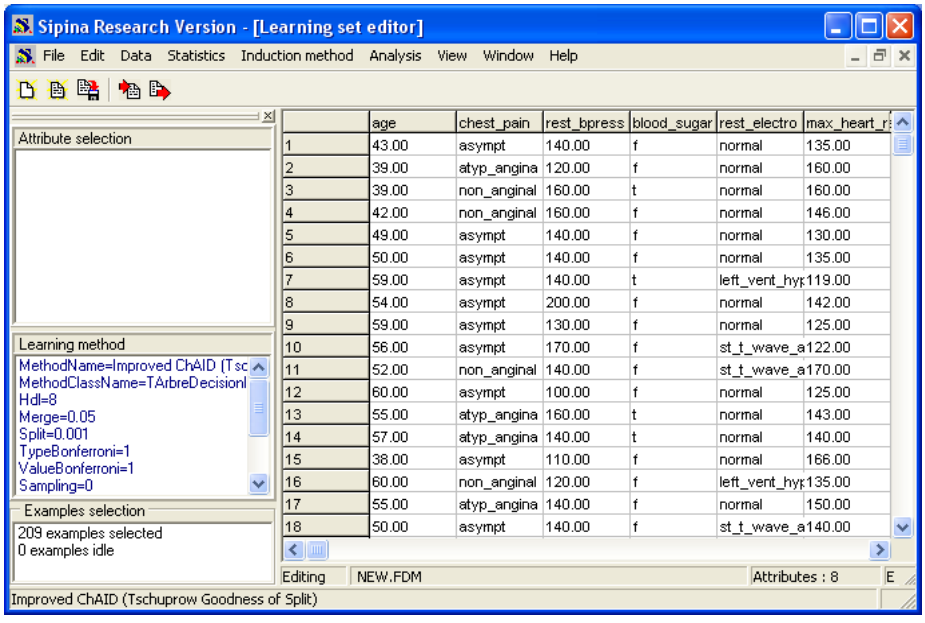
Les statistiques décrites jusqu'à présent sont relativement simples. Elles sont accessibles dans tout logiciel de statistique. La particularité de SIPINA est que nous allons pouvoir réaliser, relativement simplement, les mêmes calculs sur les différents nœuds d'un arbre de décision. Cette fonctionnalité permet d'affiner les résultats lors de l'exploration interactive.

4.1 Construction d'un arbre de décision

Dans un premier temps, il nous faut bien sûr construire un arbre de décision : DISEASE est la variable à prédire. Avant de lancer une analyse, il convient de vider toutes les fenêtres flottantes précédemment construites en cliquant sur le menu WINDOW / CLOSE ALL.

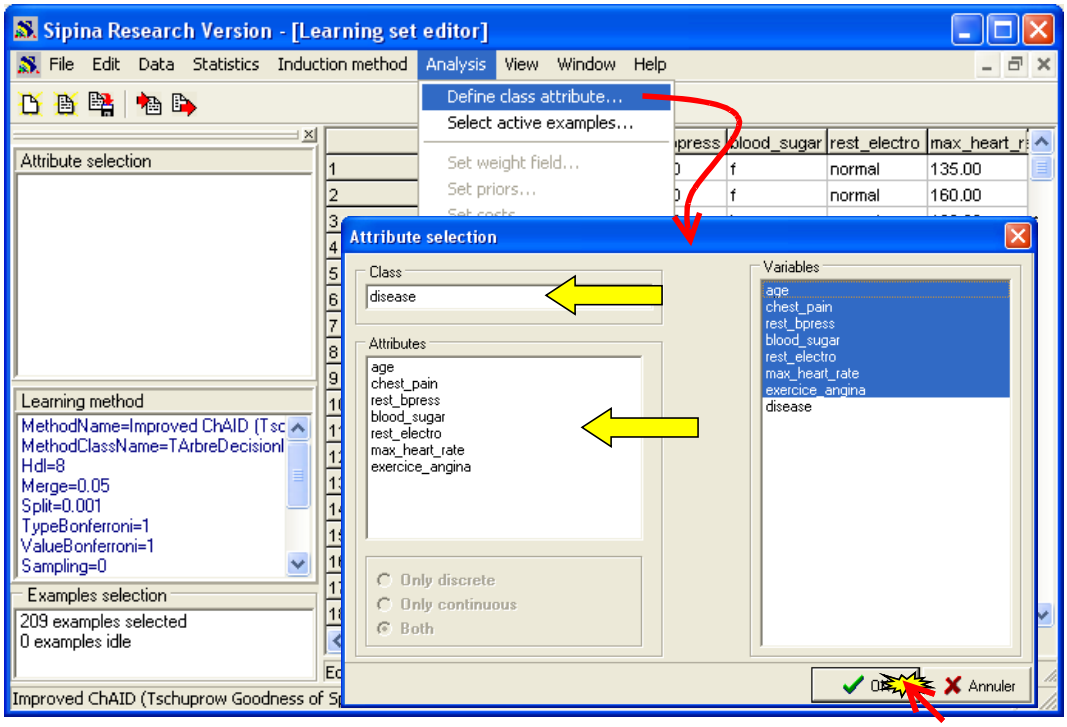


Seules les fenêtres principales (Données, Explorateur de projet) sont conservées.

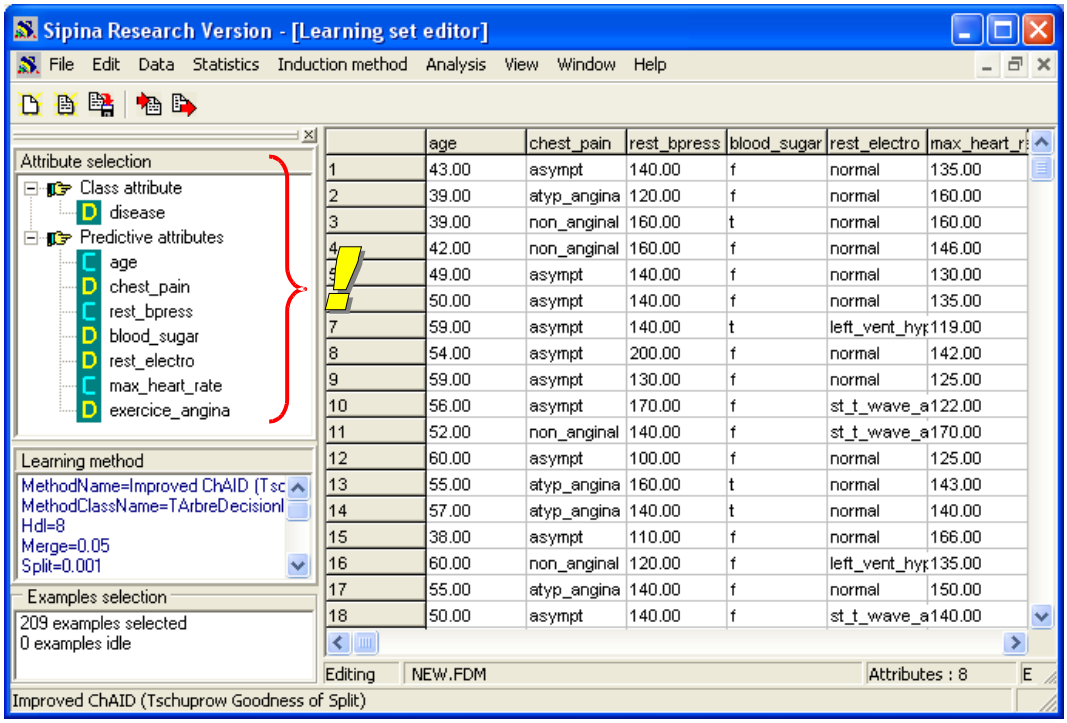


4.1.1 Choix des variables de l'étude

Pour définir le rôle des variables dans l'étude, nous actionnons le menu ANALYSIS / DEFINE CLASS ATTRIBUTE. Dans la boîte de paramétrage, par glisser déposer, nous plaçons DISEASE en CLASS, les autres en ATTRIBUTES.



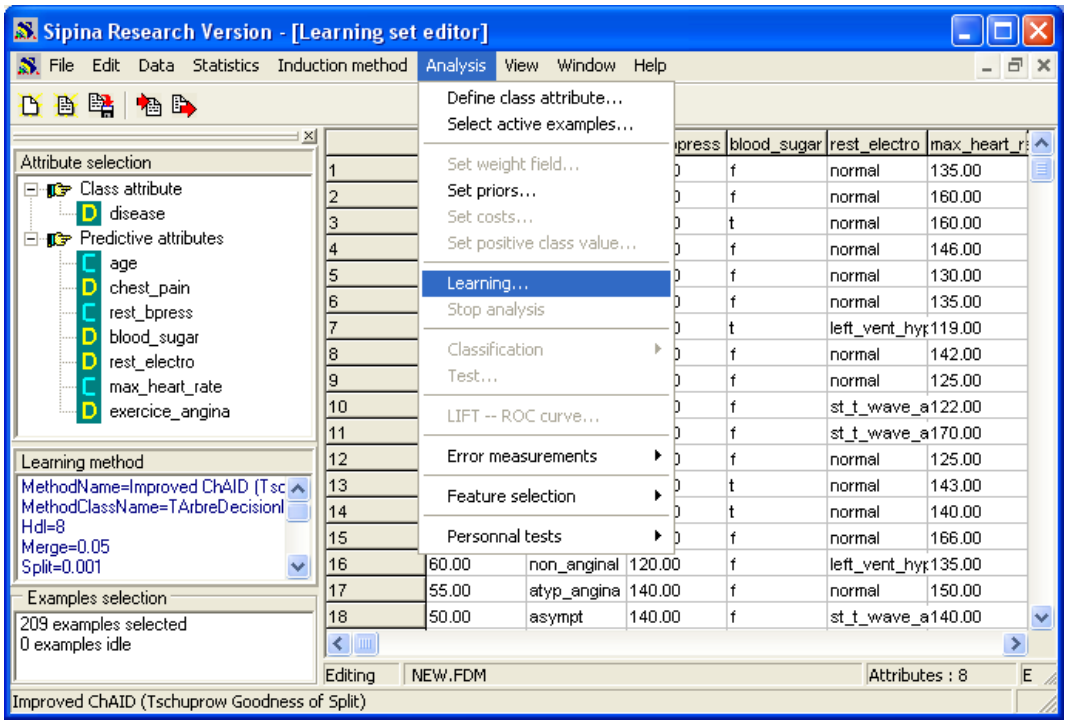
La sélection est retranscrite dans la partie haute de l'explorateur. Le type des variables (D : discrète, C ; continue) est indiqué.



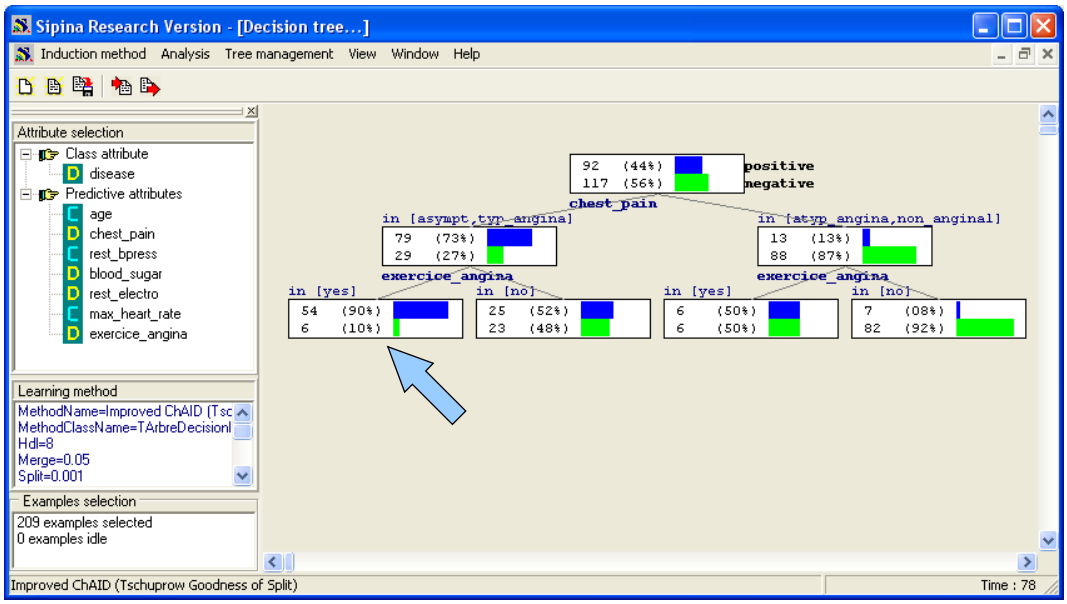
	age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_r
1	43.00	asympt	140.00	f	normal	135.00
2	39.00	atyp_angina	120.00	f	normal	160.00
3	39.00	non_anginal	160.00	t	normal	160.00
4	42.00	non_anginal	160.00	f	normal	146.00
5	49.00	asympt	140.00	f	normal	130.00
6	50.00	asympt	140.00	f	normal	135.00
7	59.00	asympt	140.00	t	left_vent_hyp	119.00
8	54.00	asympt	200.00	f	normal	142.00
9	59.00	asympt	130.00	f	normal	125.00
10	56.00	asympt	170.00	f	st_t_wave_a	122.00
11	52.00	non_anginal	140.00	f	st_t_wave_a	170.00
12	60.00	asympt	100.00	f	normal	125.00
13	55.00	atyp_angina	160.00	t	normal	143.00
14	57.00	atyp_angina	140.00	t	normal	140.00
15	38.00	asympt	110.00	f	normal	166.00
16	60.00	non_anginal	120.00	f	left_vent_hyp	135.00
17	55.00	atyp_angina	140.00	f	normal	150.00
18	50.00	asympt	140.00	f	st_t_wave_a	140.00

4.1.2 Induction de l'arbre

Pour construire l'arbre de décision sur les observations disponibles, il faut activer le menu ANALYSIS / LEARNING.



L'arbre est automatiquement affiché. Notre base étant de taille réduite, le calcul est très rapide. La distribution de DISEASE est affichée dans les nœuds de l'arbre. Dans certaines feuilles, la décision est assez tranchée c.-à-d. nous observons une forte proportion d'une des modalités de DISEASE.



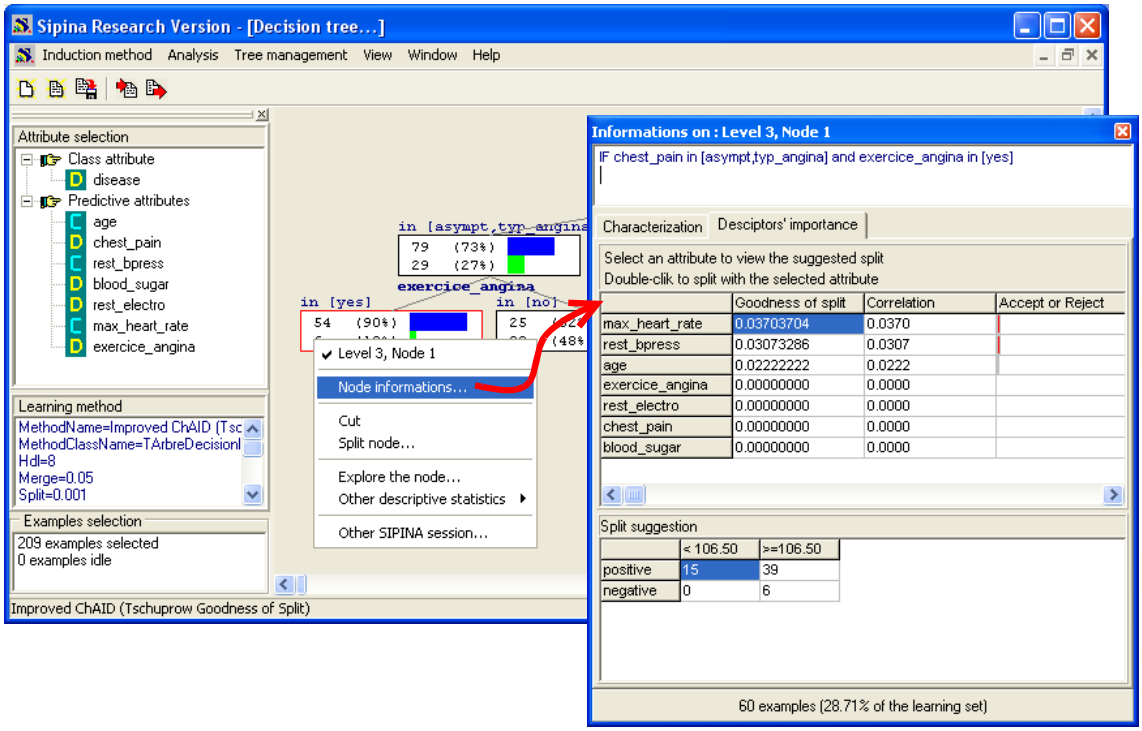
Prenons la feuille à gauche, correspondant à la règle « **SI** CHEST_PAIN = (ASYMPT ou TYP_ANGINA) **ET** EXERCICE_ANGINA = YES **ALORS** DISEASE = YES ». Nous avons isolé un groupe de personnes malades, parfaitement décrit par la règle. Mais nous ne savons pas ce qu'il en est des autres variables. Est-ce qu'elles ne sont pas pertinentes dans cette étude ? Certaines peut-être. Plus

vraisemblablement, leur rôle est masqué par les variables qui ont été introduites dans l'arbre. Les statistiques descriptives nous permettront de répondre à cette interrogation.

4.2 Exploration d'un sommet – Comparaisons sommaires

Pour l'ensemble des variables intégrées dans l'étude, SIPINA propose directement des statistiques comparatives entre le nœud sélectionné, correspondant au sous groupe d'observations que l'on veut étudier, et la racine de l'arbre, correspondant à la totalité de notre échantillon.

Pour obtenir ces statistiques, nous sélectionnons la feuille à gauche sur le dernier niveau de l'arbre, ses bords sont surlignés en rouge. Par un clic avec le bouton droit de la souris, le menu contextuel apparaît, nous activons l'item NODE INFORMATIONS. Une fenêtre flottante apparaît alors. Elle donne nombre d'indications sur le sommet : les segmentations candidates, les effectifs, les statistiques descriptives, etc.



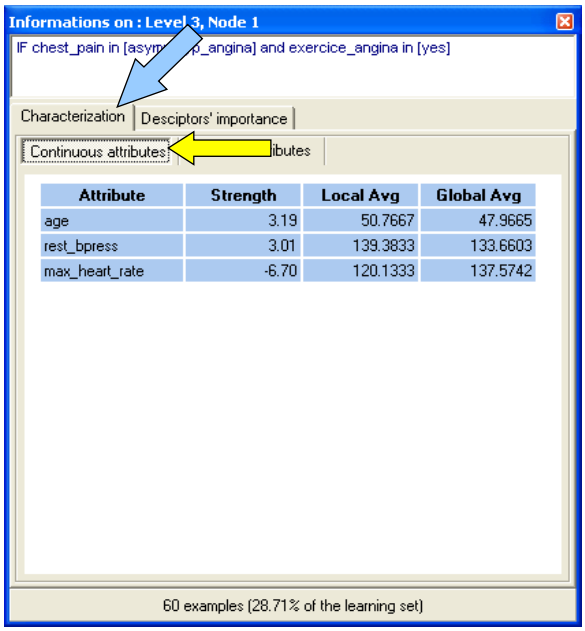
Pour obtenir les statistiques descriptives, nous cliquons sur l'onglet CHARACTERIZATION. Nous observons 2 sous onglets dans la fenêtre.

CONTINUOUS ATTRIBUTES. Pour chaque variable, nous observons la moyenne calculée sur la totalité de l'échantillon (Global Avg) et la moyenne pour le groupe d'observations correspondant au sommet étudié (Local Avg). Dans notre exemple, l'âge moyen dans le fichier est 47.9665. Pour les individus du sommet, il est de 50.7667. En moyenne, ces individus, qui sont pour une grande majorité des personnes malades (90% DISEASE = YES), sont plus âgés que la moyenne.

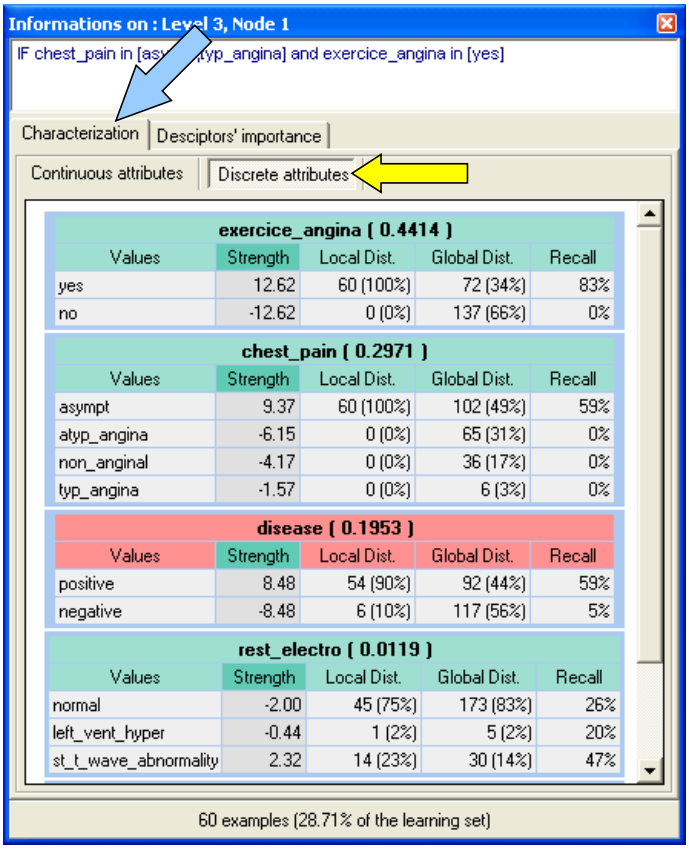
Les variables sont exprimées dans des unités différentes. Nous ne pouvons pas comparer les différences d'une variable à l'autre. Pour rendre possible cette opération, l'indicateur STRENGTH est proposé. Il s'agit de la « valeur test » décrite dans l'ouvrage de Lebart et al. (2000)[5]. Il

[5] L. Lebart, A. Morineau, M. Piron, « Statistique exploratoire multidimensionnelle », Dunod, 2000 ; pages 181 à 184. Les formules des valeurs test sont également décrites dans R. Rakotomalala, « Arbres de classification», http://eric.univ-lyon2.fr/~ricco/cours/slides/arbres_de_classification.pdf, page 11.

correspond grosso modo à la statistique du test de comparaison de moyennes. Sauf qu'il ne s'agit pas d'un vrai test ici, les échantillons ne sont pas indépendants. Il ne faut pas prendre au pied de la lettre les valeurs, et surtout pas s'attacher au caractère significatif des différences en les comparant avec les seuils fournis par les lois de probabilités associées. Il s'agit avant tout d'un outil destiné à hiérarchiser les écarts.



DISCRETE ATTRIBUTES. Nous activons le second sous onglet pour obtenir les statistiques comparatives des variables discrètes. Nous comparons les distributions de fréquence dans ce cas.



Les variables CHEST_PAIN et EXERCICE_ANGINA interviennent déjà dans la construction du groupe, il n'est pas étonnant qu'elles apparaissent en premier. De même, DISEASE est la variable à prédire, on cherche à produire un groupe pur selon cette variable, les écarts sont le fruit de l'induction.

Le cas de REST_ELECTRO est autrement plus intéressant. Il n'intervient jamais dans la constitution du groupe. Pourtant, nous observons une surreprésentation de la modalité « ST_T_WAVE_ABNORMALITY ». Dans l'échantillon global, 14% des individus présentent cette caractéristique (GLOBAL DIST.). La proportion passe à 23% (LOCAL DIST.) sur le sommet. Enfin, 47% (RECALL) des individus « REST_ELECTRO = ST_T_WAVE_ABNORMALITY » se retrouvent dans ce groupe.

L'indicateur STRENGTH correspond à la statistique du test de comparaison de proportions. Les réserves émises ci-dessus restent d'actualité.

Un second indicateur (J-MEASURE) est affiché dans l'en-tête de chaque tableau intermédiaire. Il quantifie les écarts entre les distributions de fréquences sur la racine et sur le sommet c.-à-d. entre les colonnes GLOBAL DIST et LOCAL DIST. La valeur en elle même n'est pas très intéressante. Il n'est pas question non plus de la comparer à un hypothétique seuil. La J-MEASURE sert surtout à hiérarchiser les variables dans notre contexte : plus les distributions sont dissemblables, plus la J-MEASURE prend une valeur élevée.

4.3 Exploration d'un sommet – Statistiques descriptives

Ces premières statistiques comparatives sont faciles d'accès. En un coup d'œil, nous visualisons les divergences entre les sommets. En revanche, elles sont essentiellement univariées.

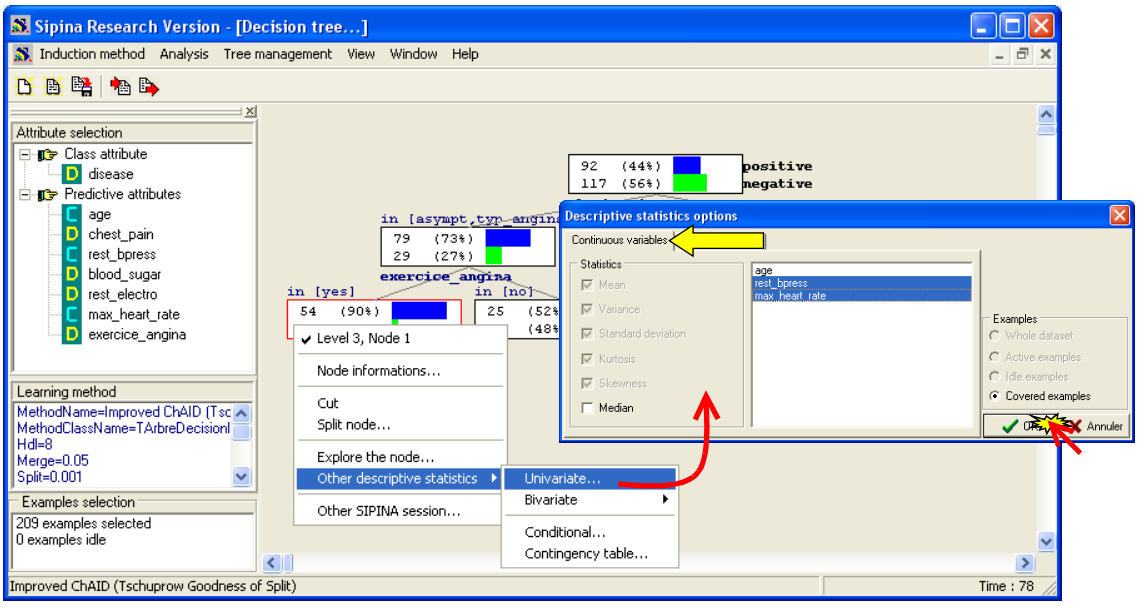
SIPINA nous permet de calculer toutes les statistiques univariées et bivariées présentées précédemment (sections 3.2 et 3.3) sur n'importe quel sommet de l'arbre de décision. L'architecture des menus, le paramétrage et la présentation des résultats sont les mêmes. A la différence que le calcul est restreint aux individus présents sur le sommet.

Reprenons les mêmes analyses (sections 3.2 et 3.3), voyons ce qu'il en est pour les individus correspondant à la description « CHEST_PAIN = (ASYMPT ou TYP_ANGINA) **ET** EXERCICE_ANGINA = YES ».

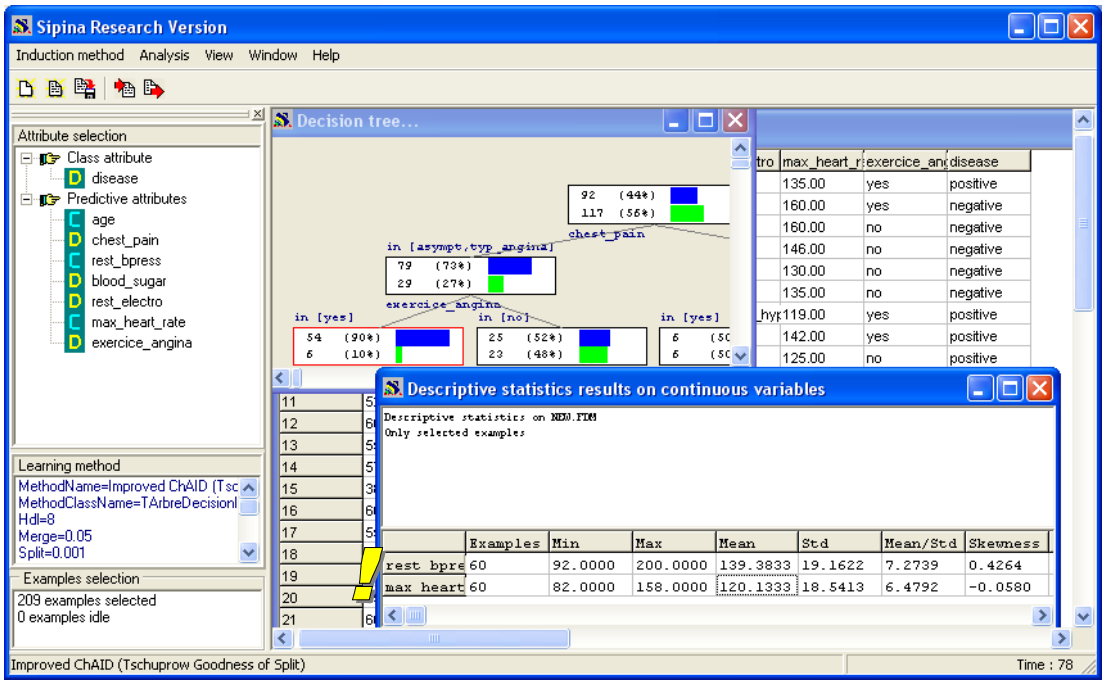
4.3.1 Statistiques univariées

Par rapport à l'outil précédent (section 4.2), l'intérêt n'est pas réellement déterminant concernant les statistiques univariées. Au lieu de les obtenir directement, il nous faut paramétrer les opérations. De plus, nous ne disposons pas de l'indicateur « valeur test » qui permet de comparer l'importance de l'écart entre les variables. Le principal intérêt de cette partie est de montrer l'aspect générique de la procédure. Nous disposons également d'autres indicateurs que la simple moyenne qui est un peu réductrice dans certains cas.

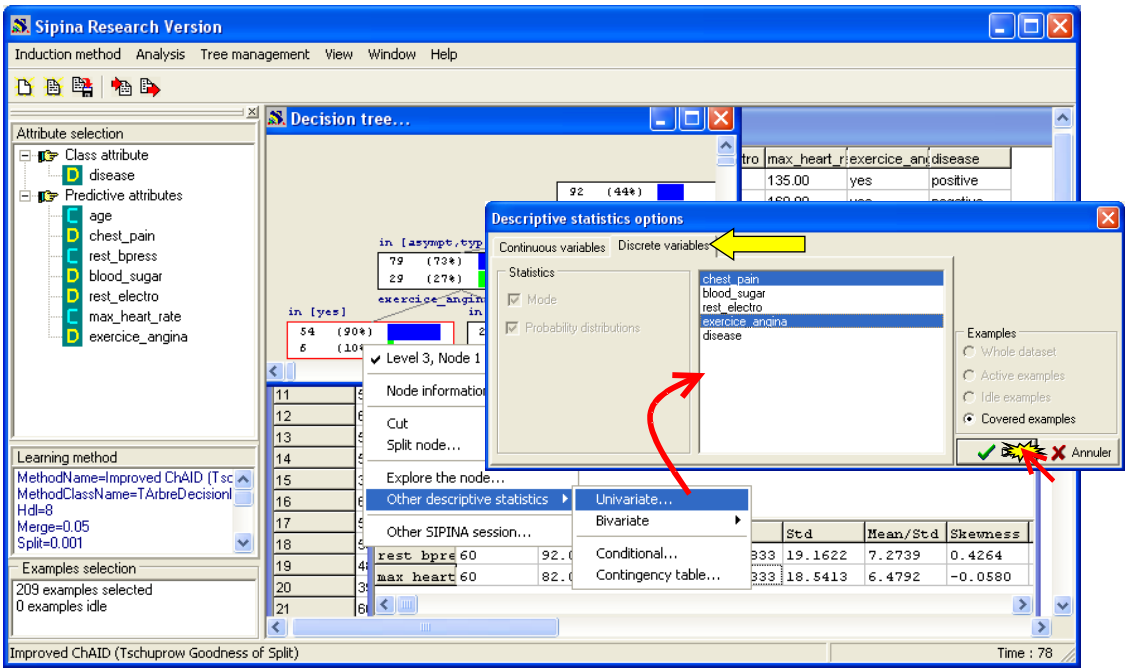
Variables continues. Pour accéder aux statistiques relatives à un sommet, nous activons le menu contextuel (clic avec le bouton droit), puis nous cliquons sur l'item OTHER DESCRIPTIVE STATISTICS / UNIVARIATE. La boîte de paramétrage apparaît. Nous noterons que l'option « COVERED EXAMPLES » est cochée d'office : seuls les 60 individus correspondant au sommet participent aux calculs. Nous sélectionnons les variables REST_BPRESS et MAX_HEART_RATE puis nous validons.



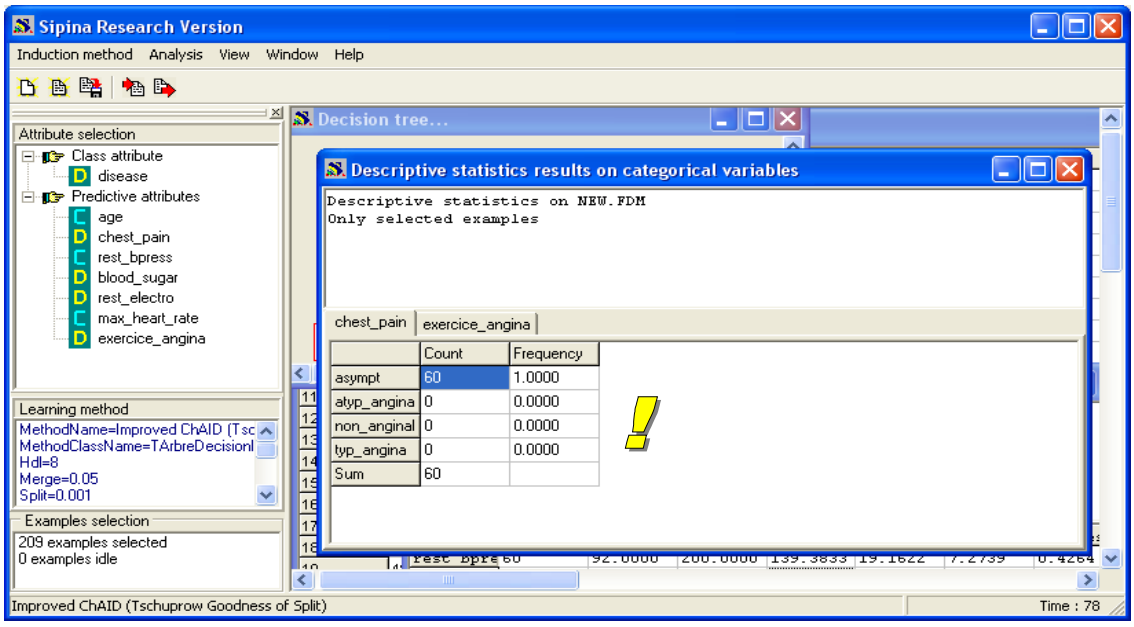
Dans la grille de résultats, nous obtenons la moyenne pour chaque variable, mais aussi tous les autres indicateurs statistiques tels que le minimum, le maximum, l'écart type, etc. Toutes ces valeurs prennent leur sens lorsque nous les comparons aux statistiques calculées sur la globalité de l'échantillon (cf. section 3.2.1).



Variables discrètes. La démarche est la même pour les variables discrètes. De nouveau, nous activons le menu contextuel, l'option OTHER DESCRIPTIVE STATISTICS / UNIVARIATE fait apparaître la boîte de paramétrage. Nous allons sur le second onglet DISCRETE VARIABLE, puis nous choisissons les variables CHEST_PAIN et EXERCICE_ANGINA.



Nous disposons des distributions de fréquence. Ici également, les résultats prennent tout leur sens lorsque nous les comparons avec les répartitions calculées sur la totalité de l'échantillon (section 3.2.2).



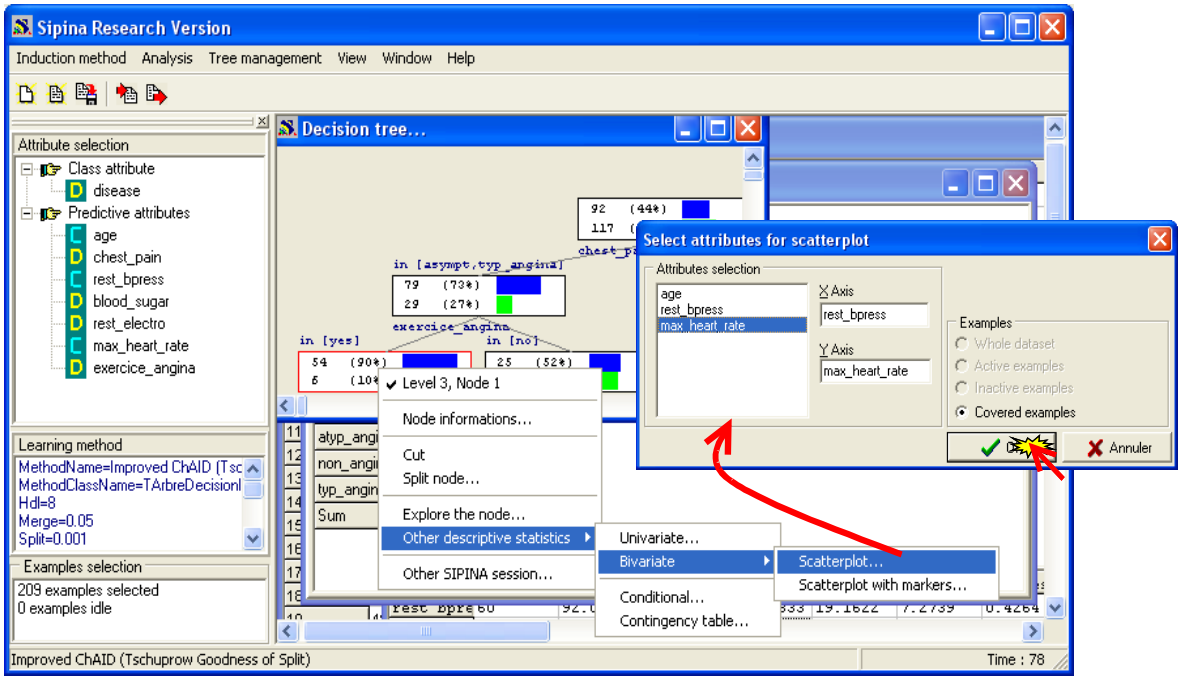
Dans notre exemple, CHEST_PAIN et EXERCICE_ANGINA ayant participé à la constitution des groupes, seules les modalités apparaissant dans la règle sont représentées dans les distributions.

4.3.2 Statistiques bivariées

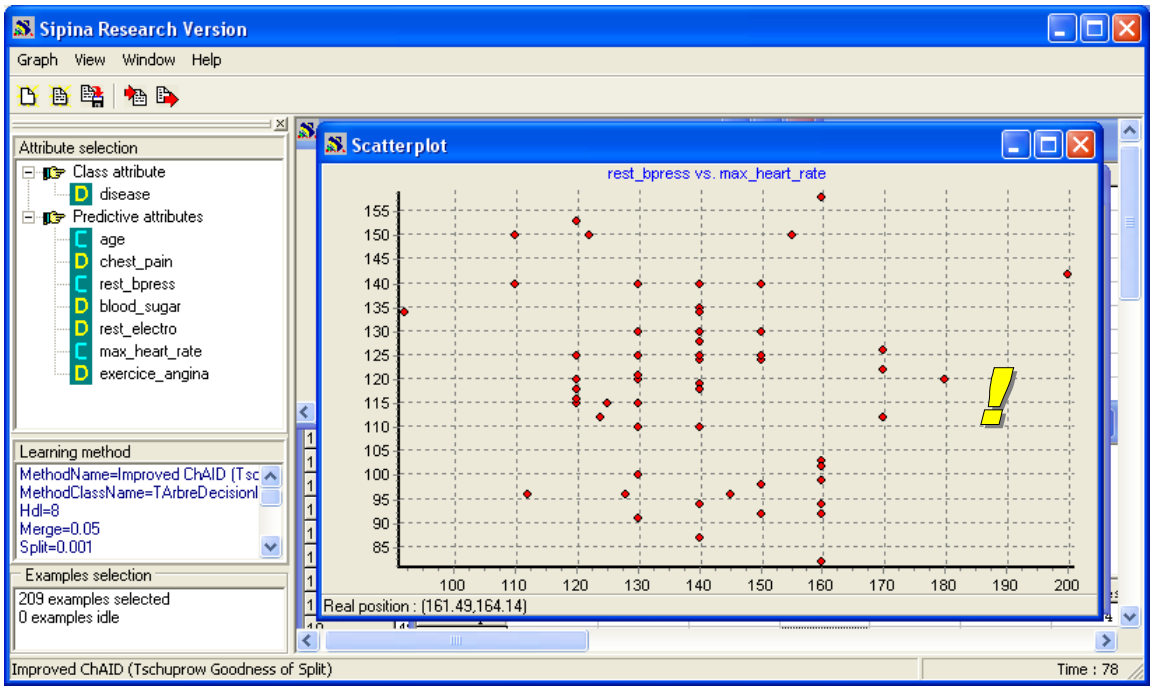
Le véritable intérêt des outils de statistiques descriptives dans SIPINA se révèle lorsque nous voulons calculer les statistiques bivariées. Nous pouvons construire les statistiques usuelles pour les observations circonscrites au sommet de l'arbre. Cette fonctionnalité est pour ainsi dire inexistante dans les logiciels issus de la recherche. Elle est peu mise en valeur dans les logiciels commerciaux. Pourtant, elle est loin d'être anecdotique dans l'exploration des données. Surtout

lorsque l'interprétation des résultats est au moins aussi importante que la « simple » recherche d'un modèle de prédiction performant.

Deux variables continues, nuage de points. Nous voulons croiser les REST_BPRESS et MAX_HEART_RATE (cf. section 3.3.1). Dans le menu contextuel du sommet, nous sélectionnons l'item OTHER DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT. Dans la boîte de paramétrage, nous plaçons en abscisse REST_BPRESS, en ordonnée MAX_HEART_RATE.

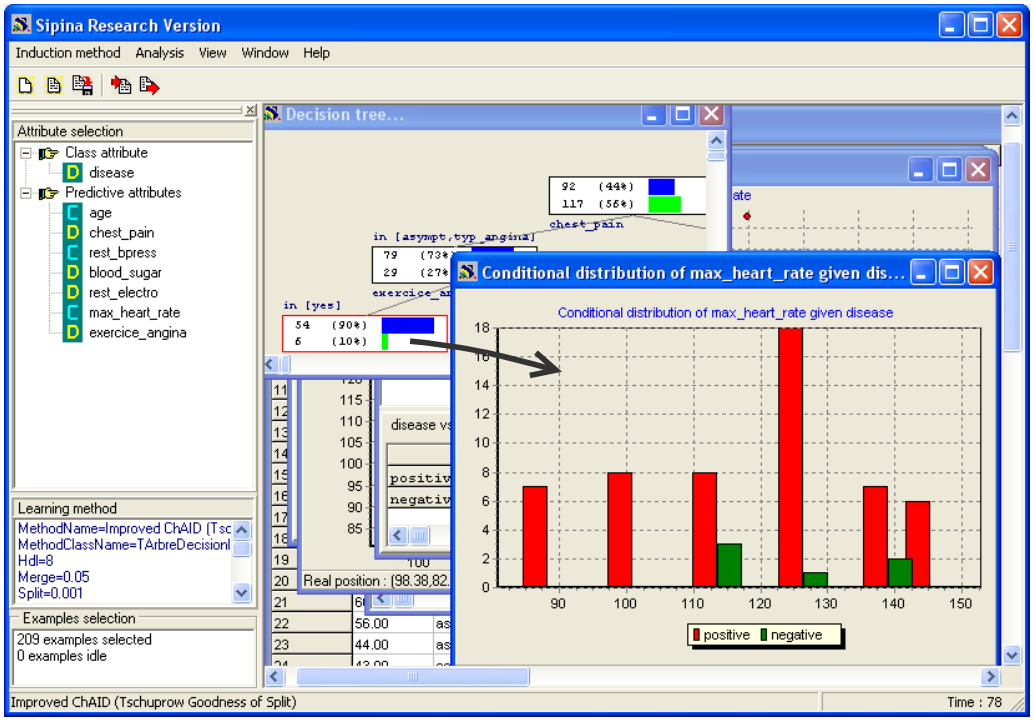


Le nuage de points est restreint aux 60 observations du sommet. Il est quelque peu différent du premier nuage de points intégrant l'ensemble de l'échantillon (cf. section 3.3.1). La relation négative entre les deux variables n'est plus de mise dans cette sous population.



Autres indicateurs. De la même manière, il nous est possible de retrouver les outils de description mis en avant plus haut (sections 3.3.2, 3.3.3 et 3.3.4).

Dans le cas de la distribution conditionnelle, MAX_HEART_RATE selon DISEASE, par rapport à la totalité de l'échantillon (section 3.3.3), nous notons que la personnes malades (DISEASE = POSITIVE) se situent essentiellement sur les valeurs faibles de MAX_HEART_RATE dans cette sous population.



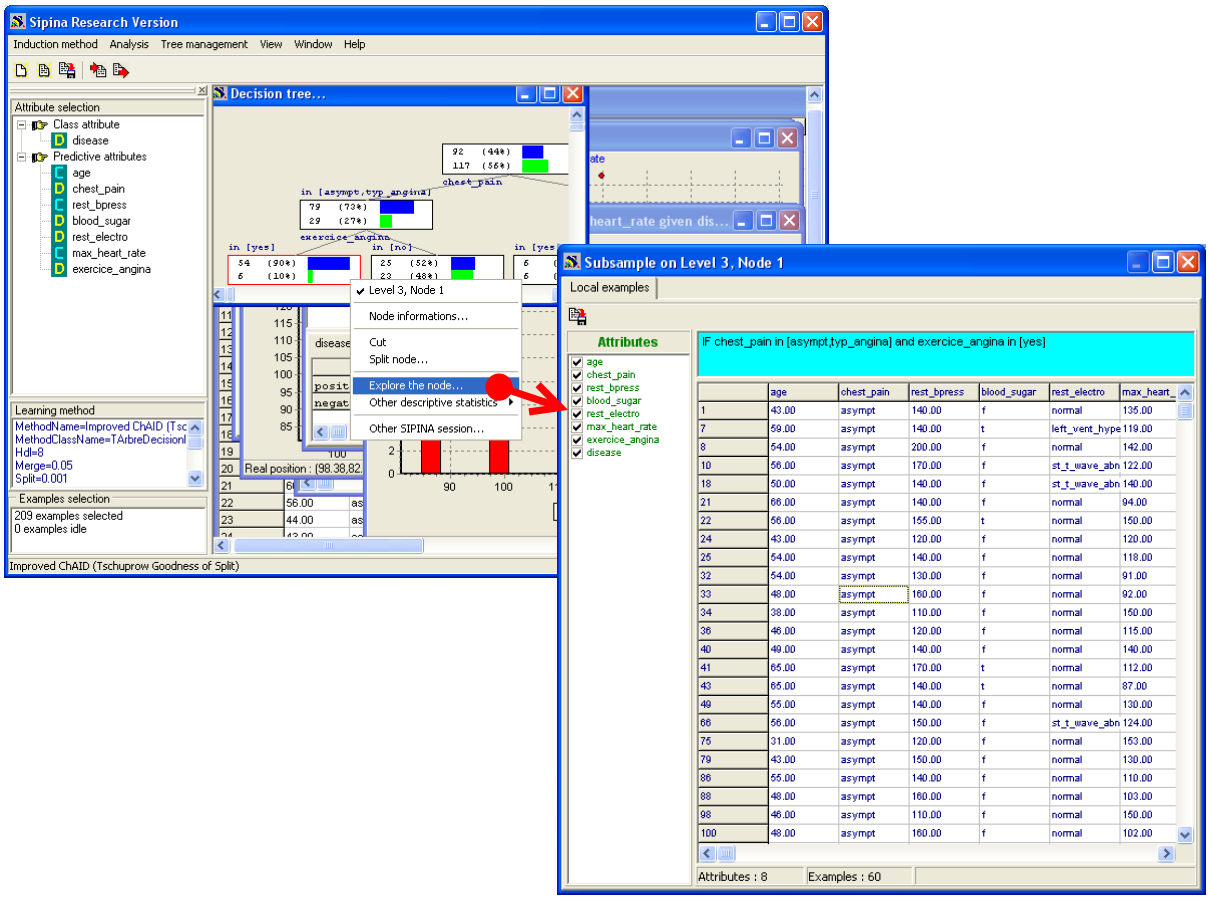
Remarque : En vérité, il est possible d'effectuer ces opérations sur n'importe quel logiciel de statistique. Il faut simplement définir le filtre correspondant à la règle de désignation d'un sommet. Puis, sur la vue ainsi élaborée, réaliser les différents calculs ci-dessus. Le principal intérêt de SIPINA est d'automatiser ces opérations qui, lorsqu'elles sont répétitives, peuvent s'avérer rapidement fastidieuses. Ce raccourci affranchit l'utilisateur de manipulations annexes (requête, extraction de fichiers intermédiaires, paramétrage, etc.), il permet une meilleure interaction lors de l'exploration de l'arbre de décision.

5 Tableaux de données intermédiaires

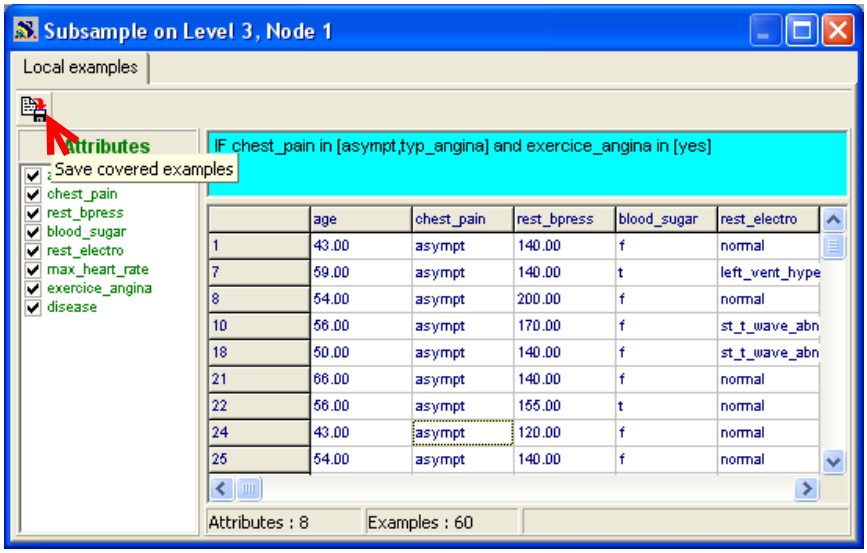
Lorsque nous désirons affiner les résultats, il peut être nécessaire de revenir sur les données, notamment pour analyser de manière approfondie les sous populations décrites par les sommets de l'arbre. Lorsque les observations sont identifiables (numéro de sécurité sociale, etc.), on peut même vouloir revenir sur les cas individuels.

Nous pouvons obtenir le détail des observations situées sur un sommet dans SIPINA. Il suffit pour cela, en veillant à sélectionner le sommet adéquat, d'activer le menu contextuel EXPLORE THE NODE. Le tableau de données est alors affiché dans une nouvelle fenêtre.

La fenêtre est modale c.-à-d. il n'est pas possible de revenir sur l'arbre sans fermer la fenêtre des observations locales.



Nous pouvons sauvegarder les observations dans un nouveau fichier (format propriétaire .FDM) si nous souhaitons étudier cette sous population dans un contexte différent.



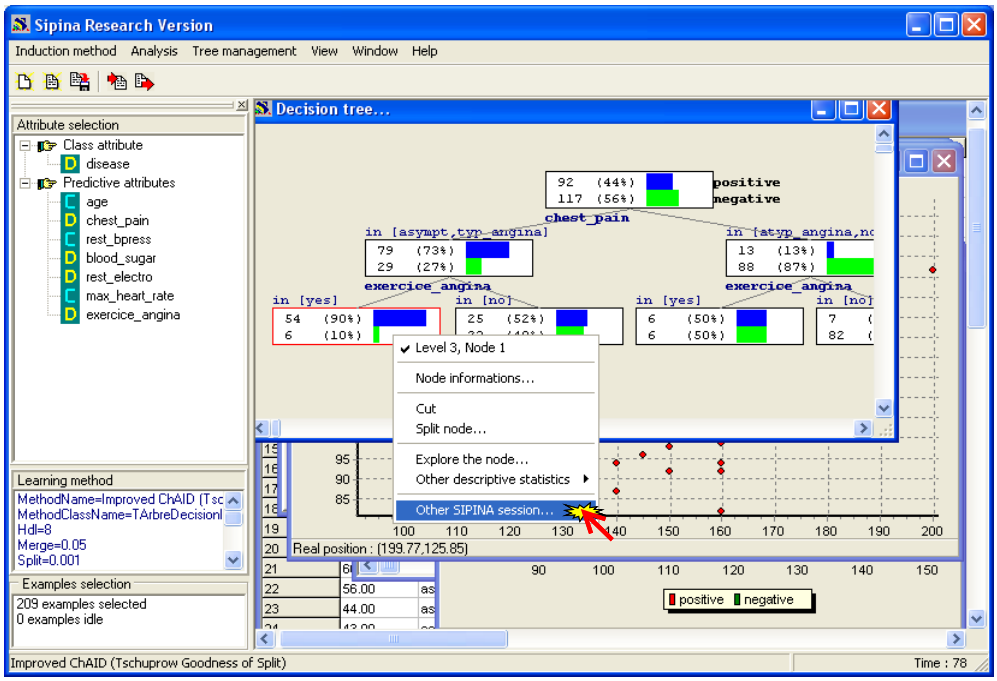
Le fichier est nommé automatiquement avec l'identifiant du sommet. Il est placé dans le même répertoire que le fichier originel[6].

[6] Note : Notre idée était de réaliser des suites de logiciels (analyse factorielle, analyse non supervisée, etc.) qui pouvaient communiquer entre eux via des fichiers partagés. Dans ce cadre, pouvoir sauver simplement des données correspondant à des sous populations constituait une option intéressante.

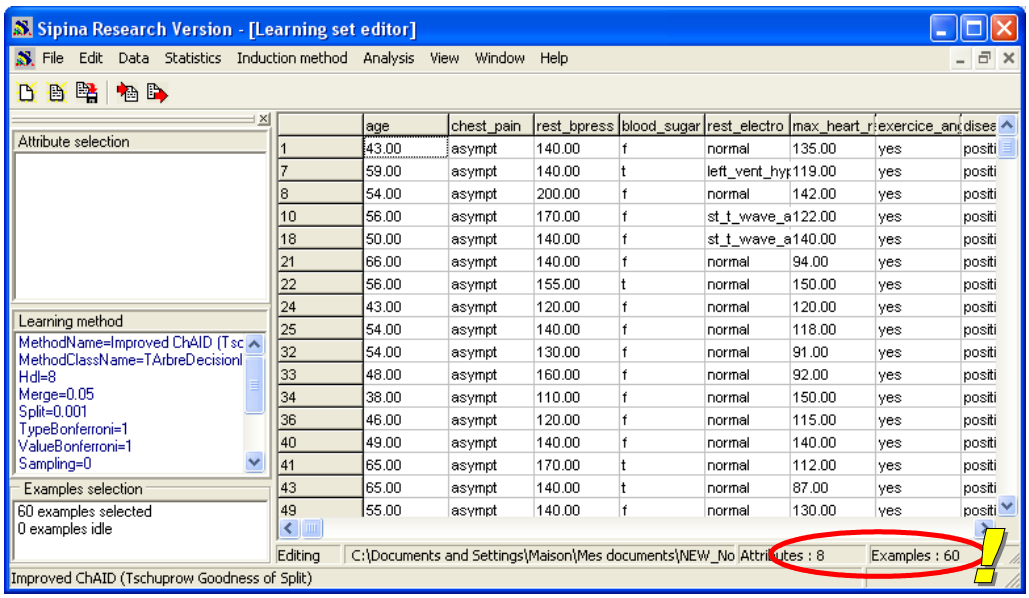
6 Nouvelle session d'analyse pour une sous population

Dans certains cas, il peut s'avérer utile de réaliser une analyse supervisée intermédiaire sur une sous population. Trois étapes sont nécessaires : sauver le fichier associé à un sommet, lancer manuellement SIPINA, charger le fichier. Ces opérations ont été réunies dans un seul item du menu contextuel.

Toujours à partir du sommet analysé, nous activons le menu contextuel OTHER SIPINA SESSION.

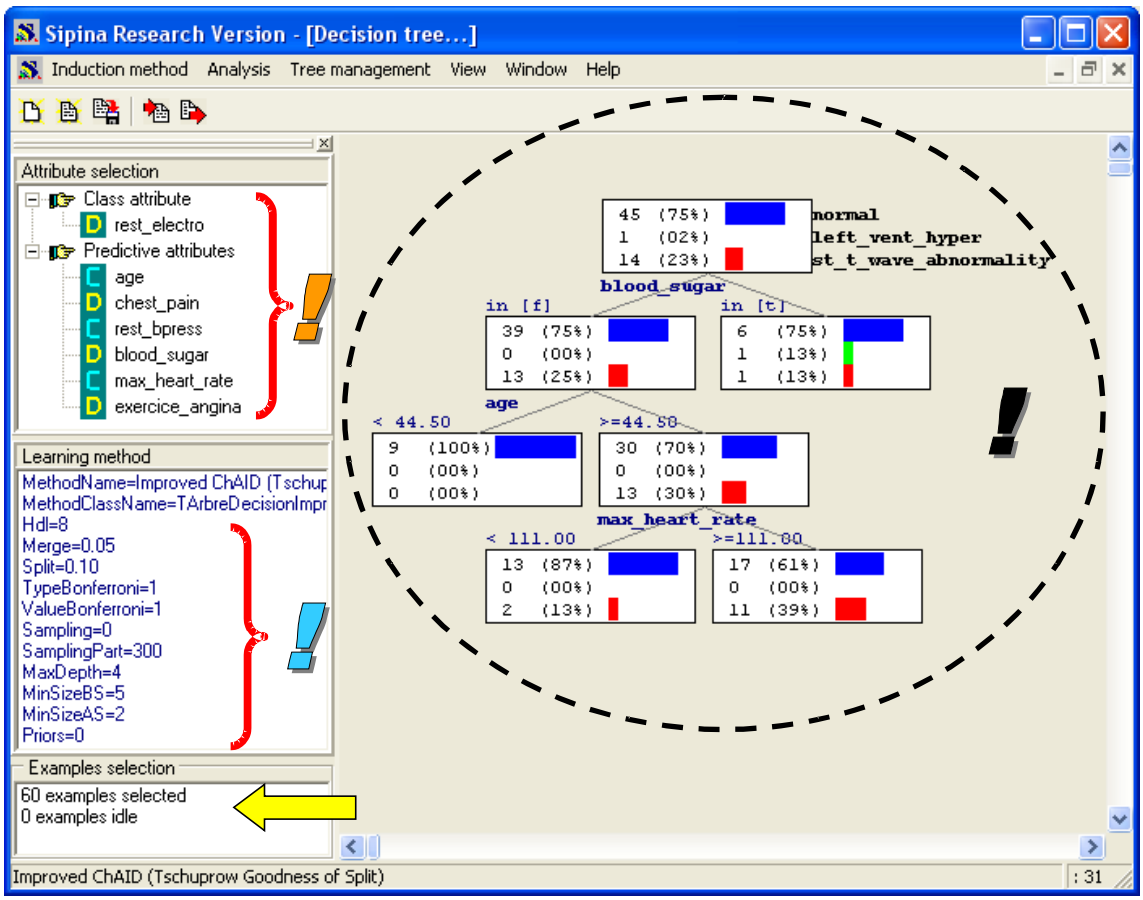


Une session de travail est proposée c.-à-d. une nouvelle application SIPINA est démarrée, les données sont automatique chargées. Nous y retrouvons nos 8 variables pour 60 observations. Nous pouvons initier une nouvelle analyse.



Par exemple, nous voulons expliquer la variable REST_ELECTRO à partir des autres descripteurs. La démarche est la même : choisir l'algorithme d'apprentissage, définir ses paramètres, désigner les variables de l'étude, etc.

Nous présentons ici un exemple de résultats avec cette nouvelle configuration. Nous avons modifié les paramètres de la méthode IMPROVED CHAID.



Les possibilités d'exploration sont infinies…

7 Conclusion

Dans ce didacticiel, nous avons voulu montrer les outils de statistiques descriptives de SIPINA. En tant que telles, ces fonctionnalités ne sont pas vraiment extraordinaires. Couplées avec l'exploration interactive d'un arbre de décision, elles s'avèrent très fécondes.