

1 Subject

SIPINA proposes some descriptive statistics functionalities.

In itself, the information is not really exceptional; there is a large number of freeware which do that. It becomes more interesting when we combine these tools with the decision tree. The exploratory phase is improved. Indeed, every node of the tree corresponds to a subpopulation. The variables which do not appear in the tree are not necessarily irrelevant. Perhaps, some of them were hidden during the tree learning which selects the “best” variables. By computing contextual descriptive statistics, in connection with the each node, we better understand the prediction rules highlighted during the induction process.

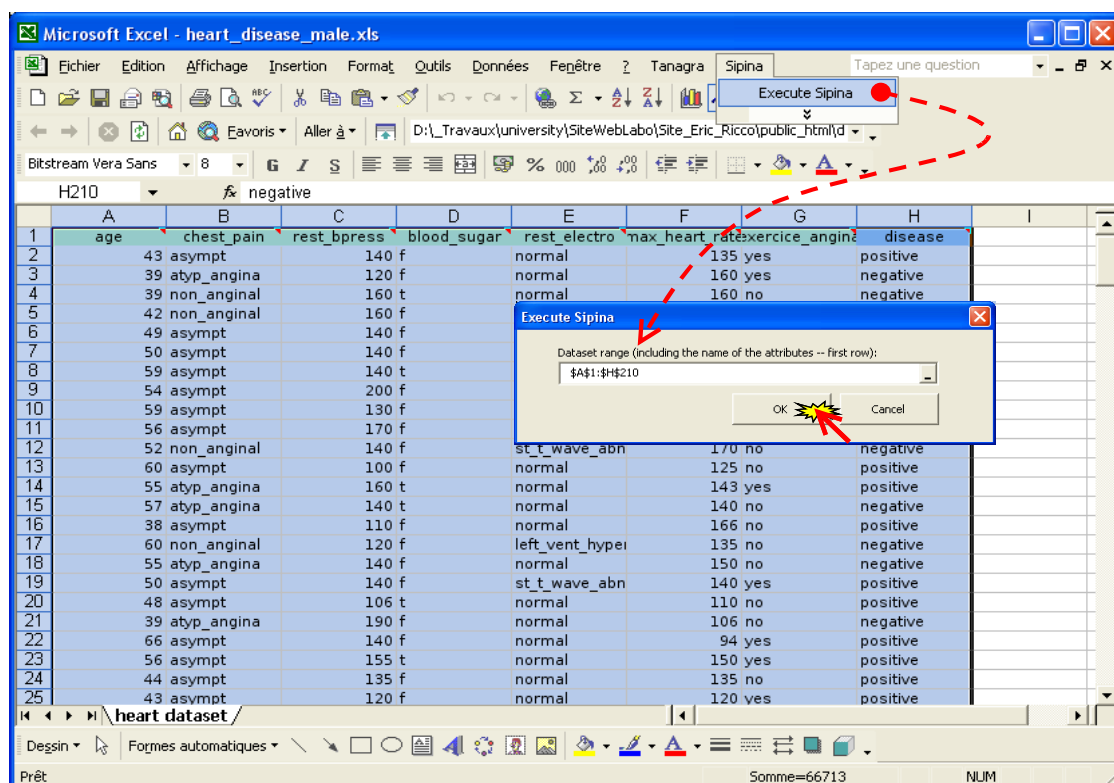
2 Dataset

We use the HEART_DISEASE_MALE.XLS¹ dataset. We want to predict the DISEASE from patient’s characteristics (AGE, SUGAR in the blood, etc.). There are 209 examples.

3 Descriptive statistics

3.1 Data importation

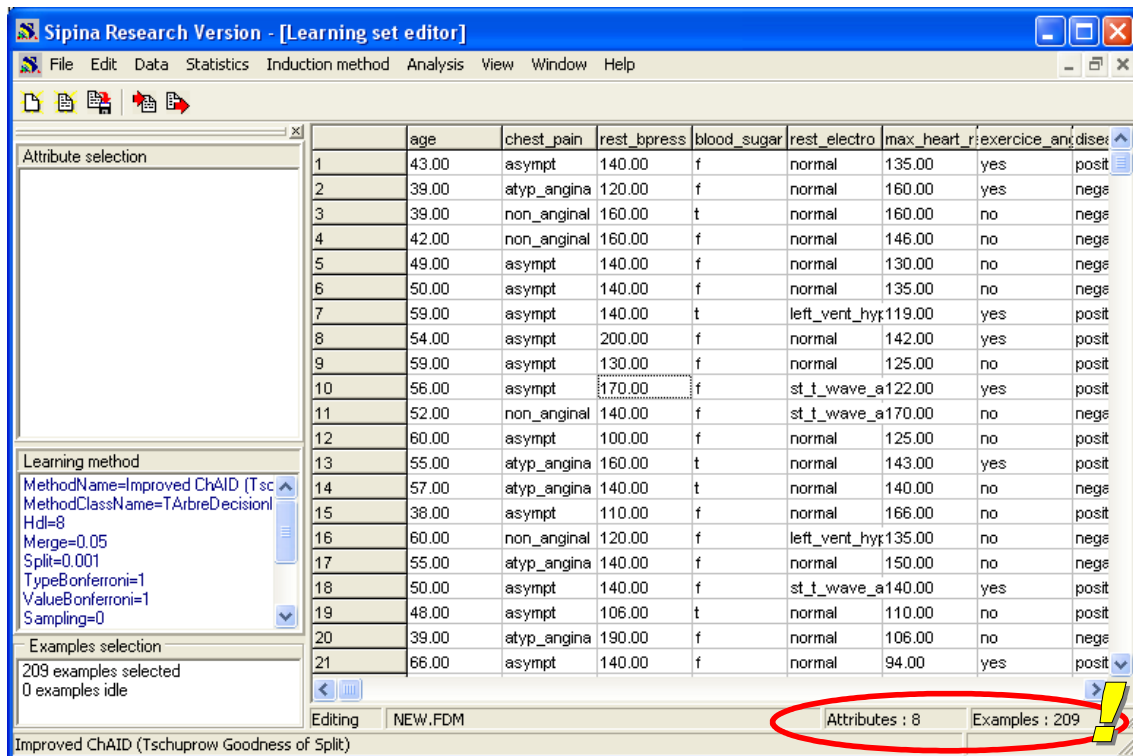
The easiest way to import the dataset is to download the file into the EXCEL spreadsheet (see http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_installation.htm for the installation of the SIPINA.XLA add-in). Then we select the cells and activate the SIPINA / EXECUTE SIPINA menu (see http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_processing.htm).



¹ http://eric.univ-lyon2.fr/~ricco/dataset/heart_disease_male.xls

SIPINA is automatically started. The data were transferred through the clipboard. The data file contains 209 individuals and 8 variables.

Note: We can save the dataset in the SIPINA binary file format (*.FDM) by clicking the FILE /SAVE AS menu. The format is useful when we handle a large dataset. During the transfer, numeric columns are encoded as continuous attributes, the other ones as discrete attributes. The first row is always the variable names.



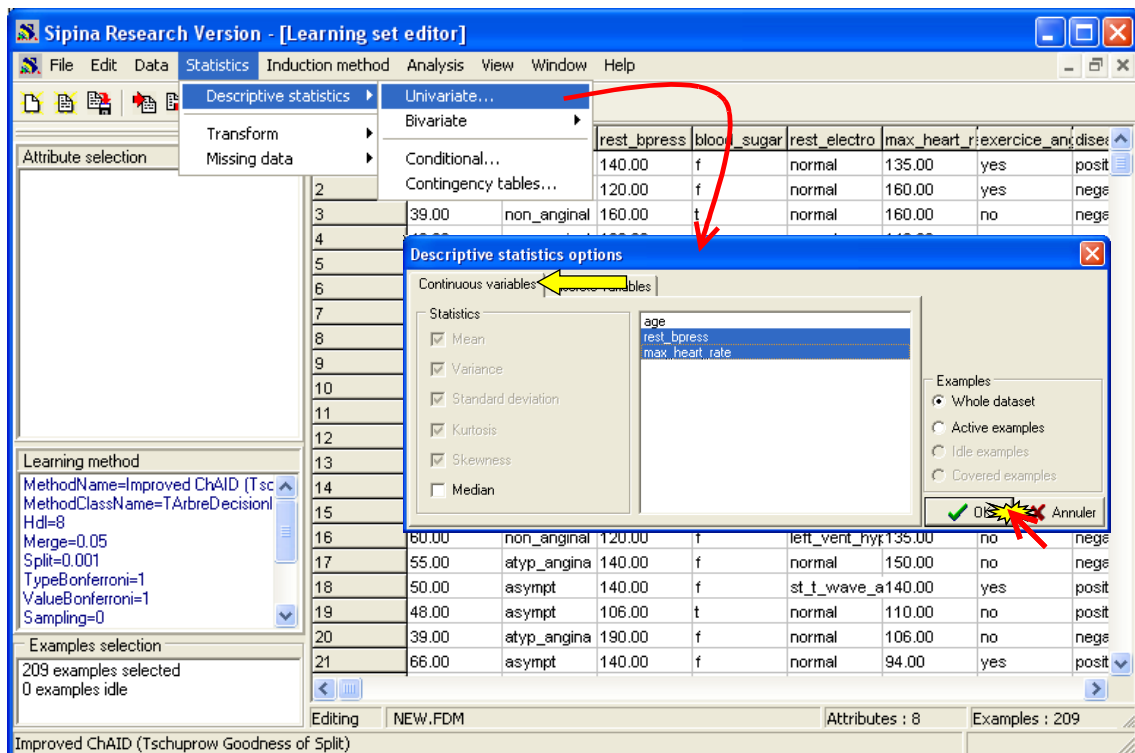
3.2 Univariate statistics

Descriptive statistics commands are available through the **STATISTICS** menu.

Note: This menu is only visible if the data grid is selected. In the other situation i.e. another window is selected, this menu is hidden. Among the various ways to select the data grid, we can use the WINDOW / LEARNING SET EDITOR menu.

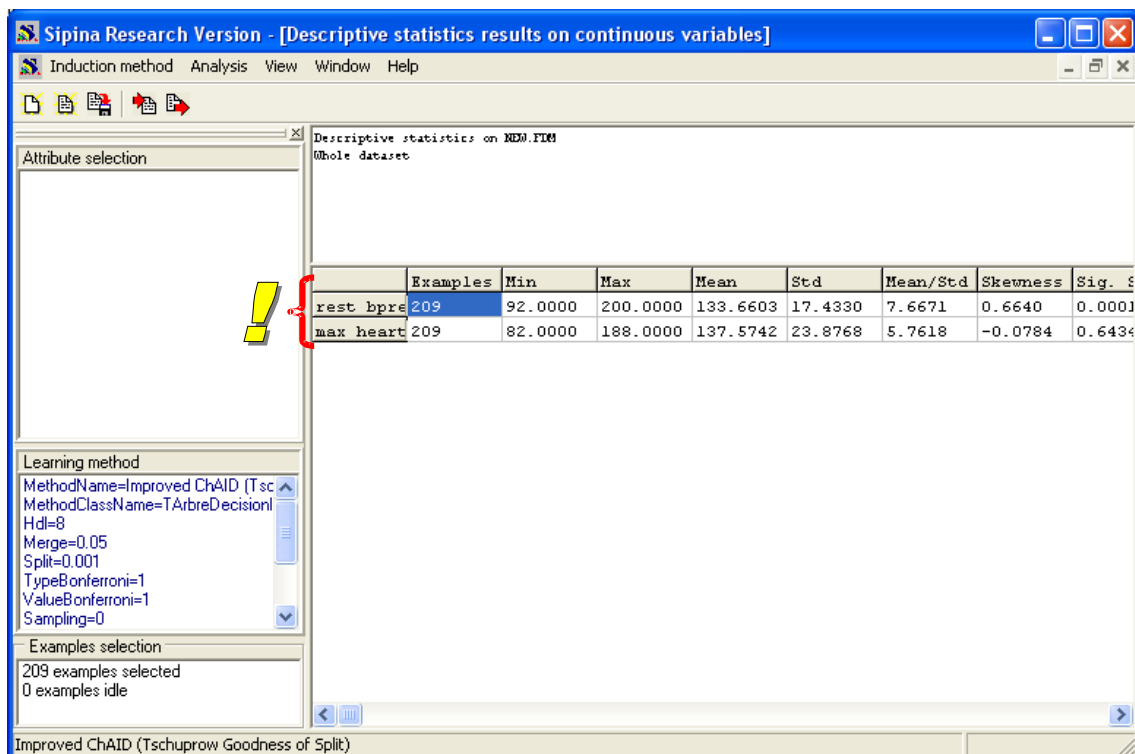
3.2.1 Continuous variables

We select the STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE menu in order to compute the descriptive statistics for continuous variables. In the dialog box which appears, we activate the CONTINUOUS VARIABLES tab. Then, we select the two following variables: REST_BPRESS and MAX_HEART_RATE.



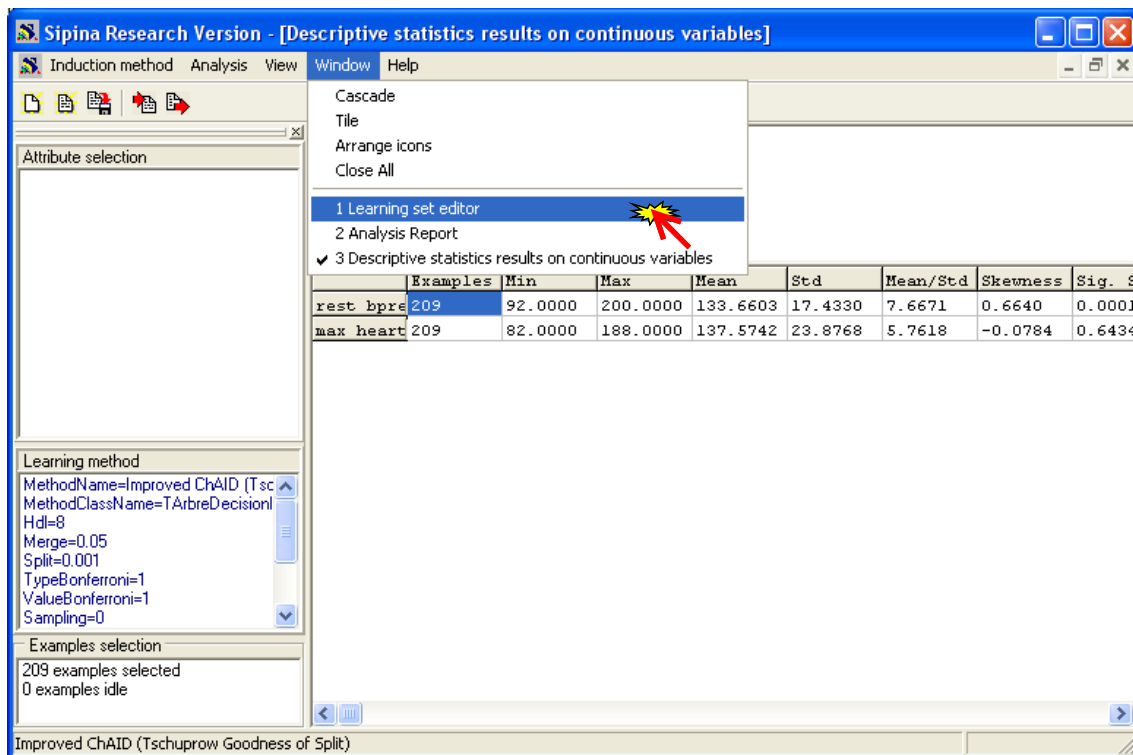
We note that the statistical indicators can be computed only on the active (selected) examples. This is useful for instance if we have partitioned the dataset into learning set and test set.

The results are displayed in a new window. Each column corresponds to a statistical indicator; each row to a variable. We can copy the values in the clipboard or modify the numerical precision using the contextual menu.

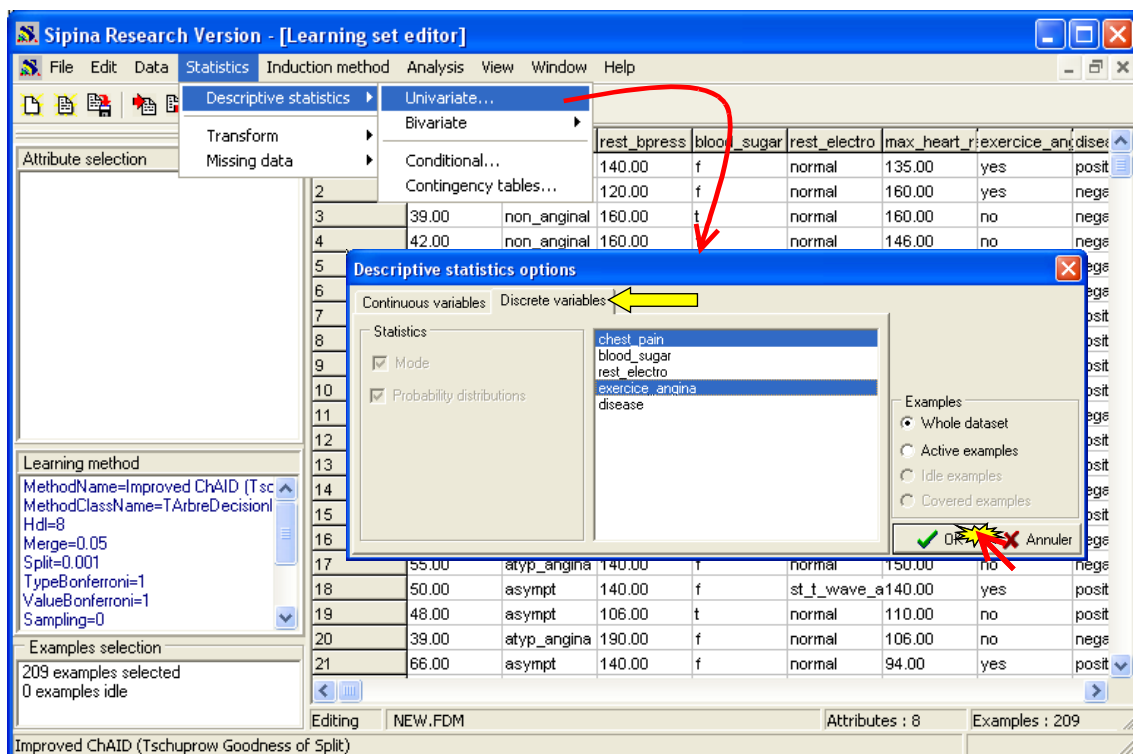


3.2.2 Discrete variables

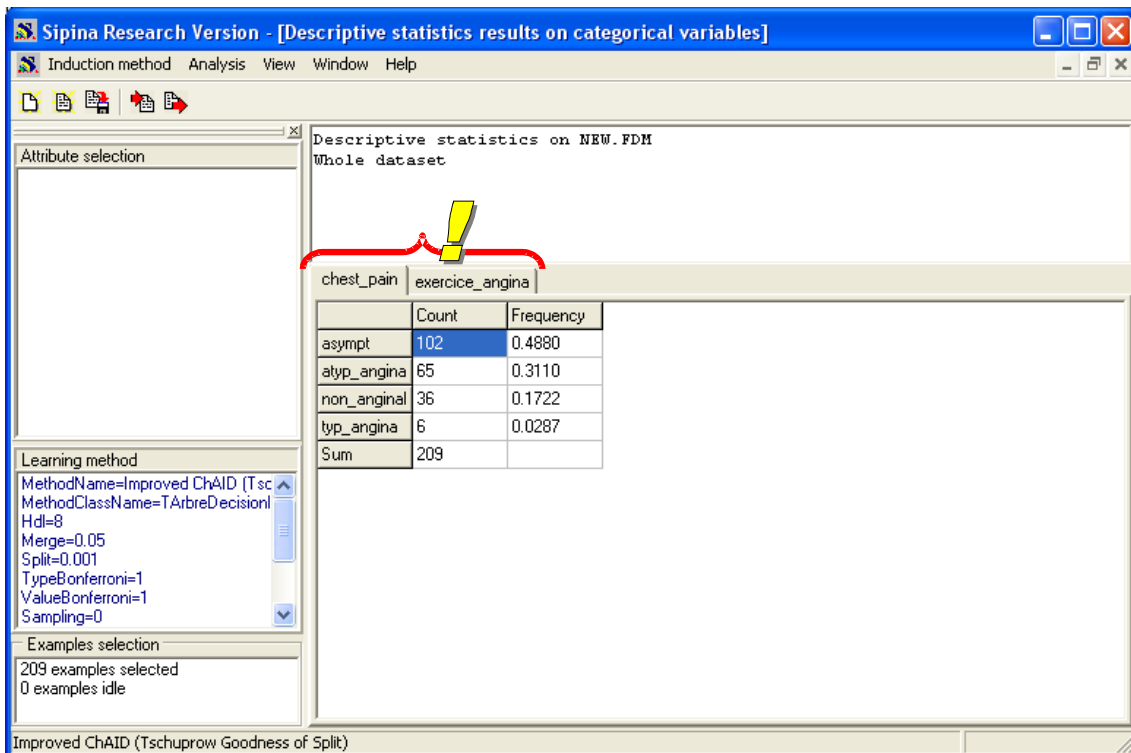
We follow the same way for the discrete variables. We activate before the data grid, by clicking the window or by clicking the WINDOW / LEARNING SET EDITOR menu.



Then, we select again the STATISTICS / DESCRIPTIVE STATISTICS / UNIVARIATE menu. In the dialog box, we choose the DISCRETE VARIABLES tab. We want to compute statistical indicators about CHEST_PAIN and EXERCICE_ANGINA.



The frequencies of values are displayed for each variable (one variable by tab).



3.3 Bivariate statistics

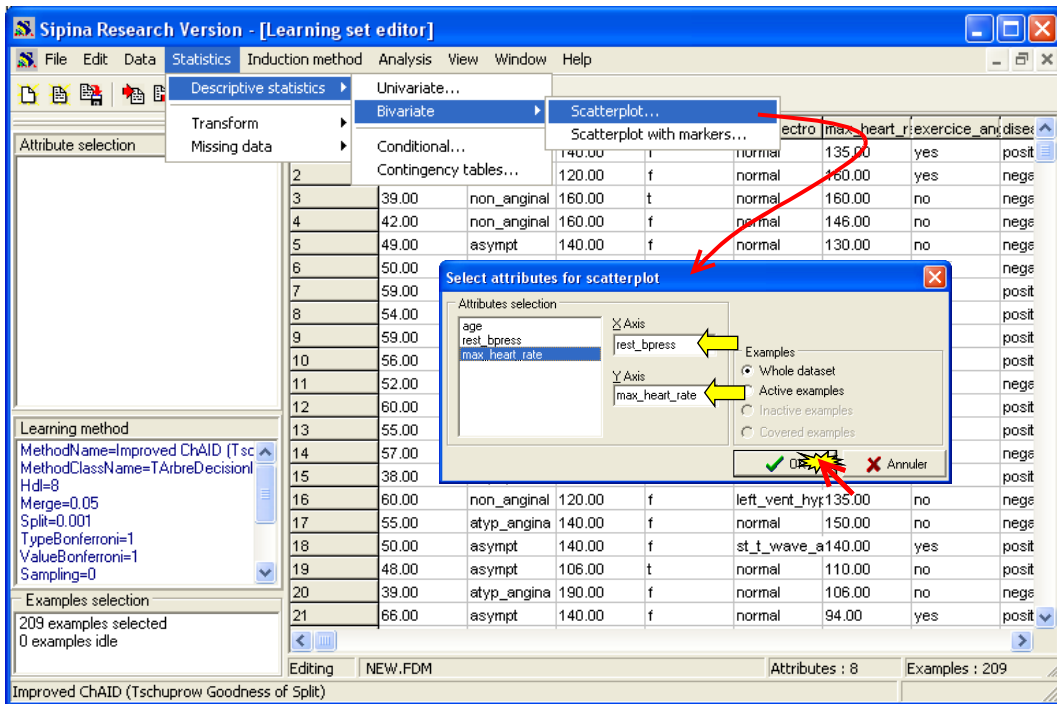
We can also compute bivariate statistics: combining two discrete variables (contingency table), two continuous variables (correlation) or mixed variables (comparison of populations).

We click on the WINDOW / LEARNING SET EDITOR menu in order to activate the data grid. The STATISTICS menu is now visible.

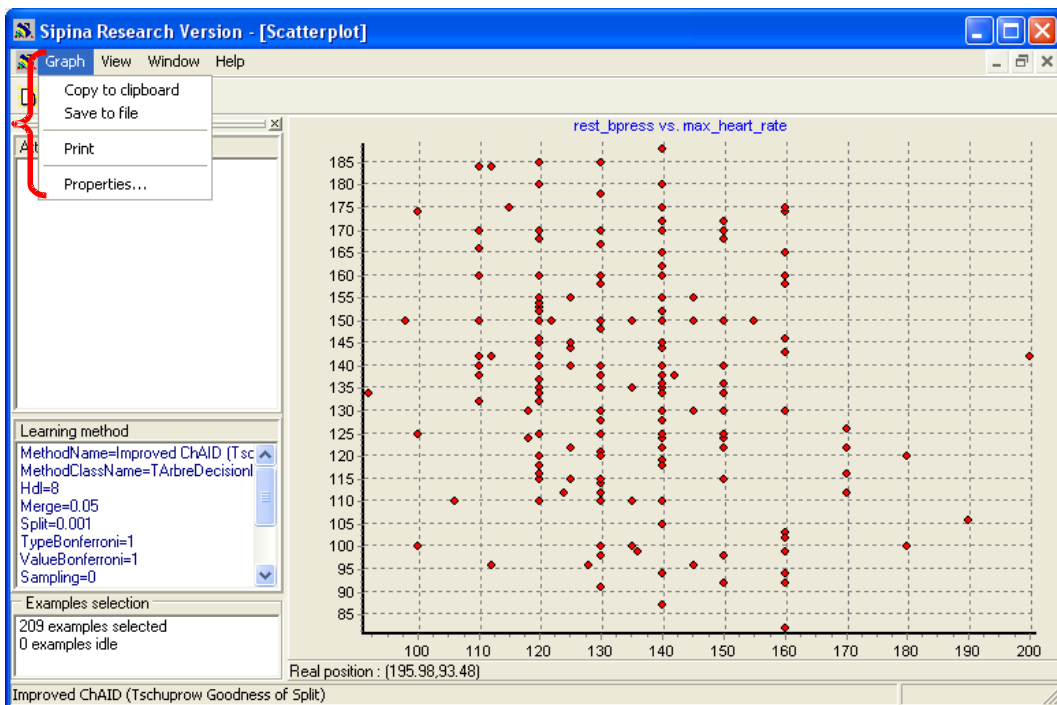
3.3.1 Two continuous variables: scatter plot

The scatter plot provides useful information about the relation between two variables (kind of association, outliers, etc.).

This functionality is available with the STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT menu. A dialog box enables to select the variable on the horizontal axis (X: REST_BPRESS) and the vertical axis (Y: MAX_HEART_RATE).



The display window is generated. In the same time, a new menu (GRAPH) is now available. Some options enable us to copy the graph in the clipboard, to print it, to modify the size of the points, etc.

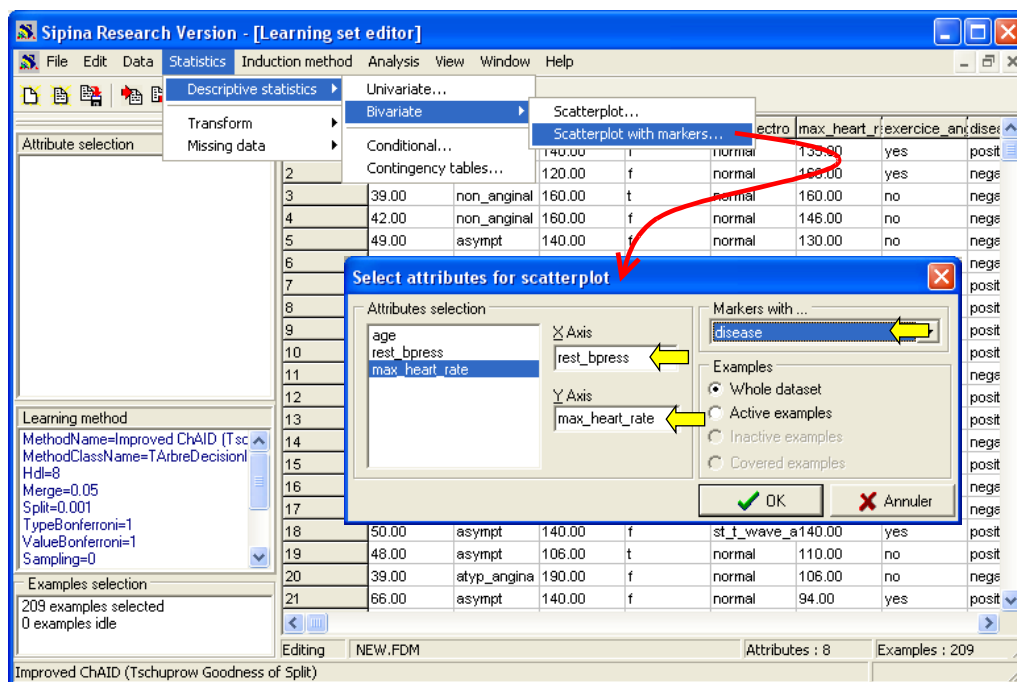


3.3.2 Conditional scatter plot

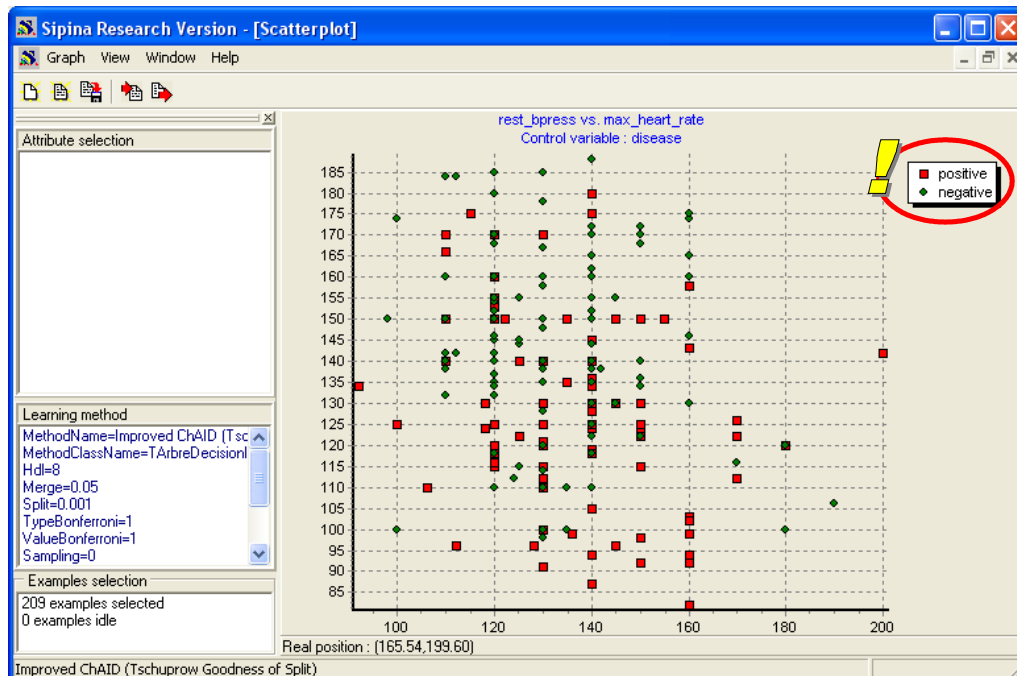
The scatter plot is all the more interesting when we can illustrate the relative situation of groups of individuals. In this case, we use a third variable in order to “colorize” the points. In our preceding example, we want to distinguish the people according the DISEASE.

We select again the data grid (WINDOW / LEARNING SET EDITOR menu). Then we click on the STATISTICS / DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT WITH MARKERS menu. In the

dialog box, we set REST_PRESS as horizontal axis, MAX_HEART_RATE as vertical axis, and DISEASE as marker.



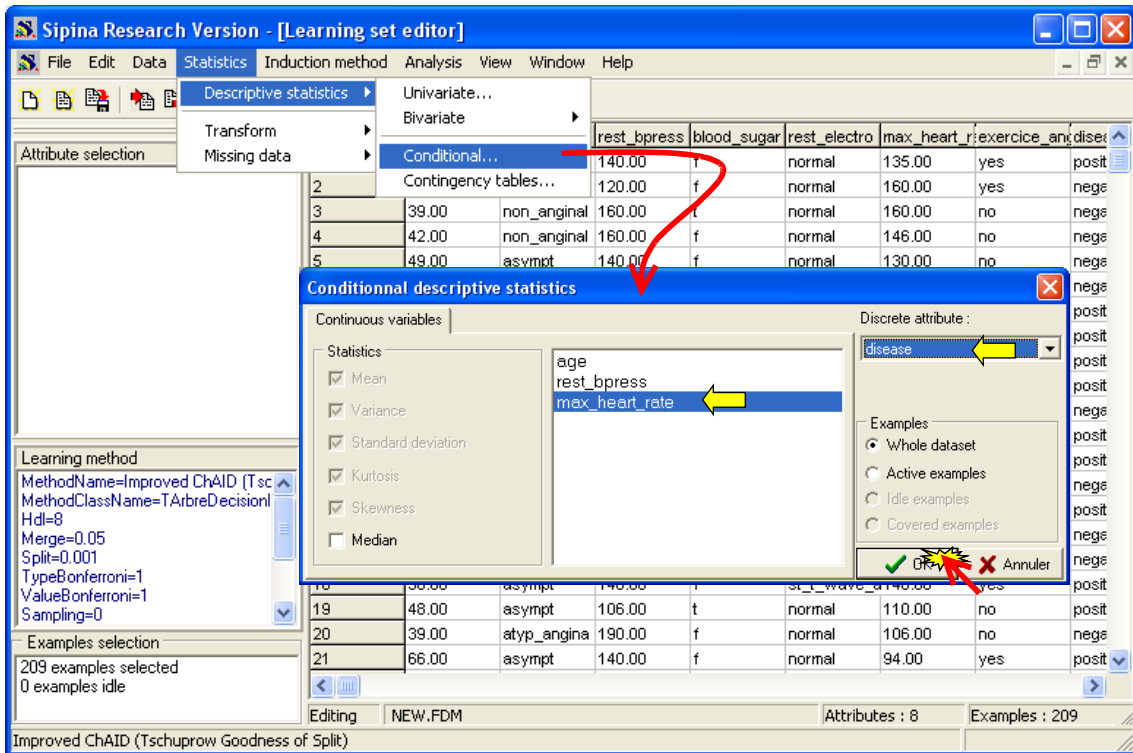
We obtain the same scatter plot than previously. The difference is we distinguish now the people DISEASE = YES.



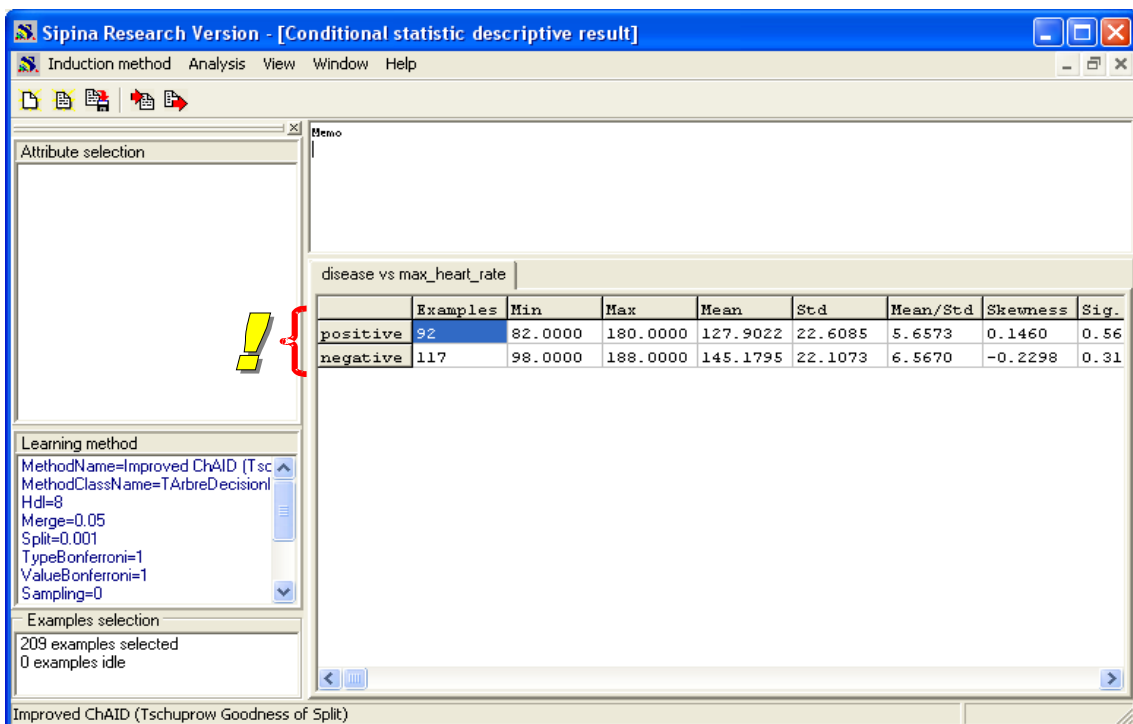
3.3.3 Continuous variable vs. Discrete variable

This functionality enables, among other, to compare the characteristics of subpopulations. Especially, it allows comparing the conditional distribution of a continuous attribute according the value of a discrete variable. For our dataset, we want to study the distribution of MAX_HEART_RATE for each subpopulation corresponding to the DISEASE values.

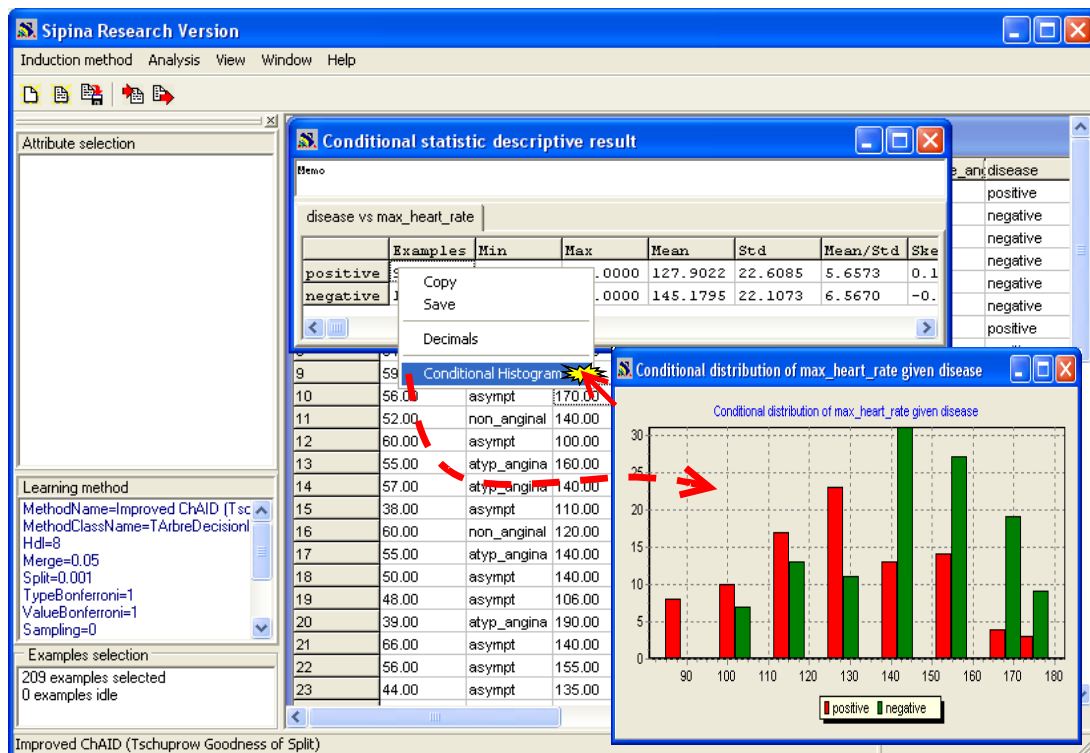
We select the data grid (WINDOWS / LEARNING SET EDITOR menu). Then, we click on the STATISTICS / DESCRIPTIVE STATISTICS / CONDITIONNAL menu. In the dialog box, we select MAX_HEART_RATE and DISEASE.



The result grid gives the descriptive statistics. We obtain the mean, the standard deviation, etc. We observe for our dataset that the average of MAX_HEART_RATE is lower for the people with DISEASE = POSITIVE (YES).

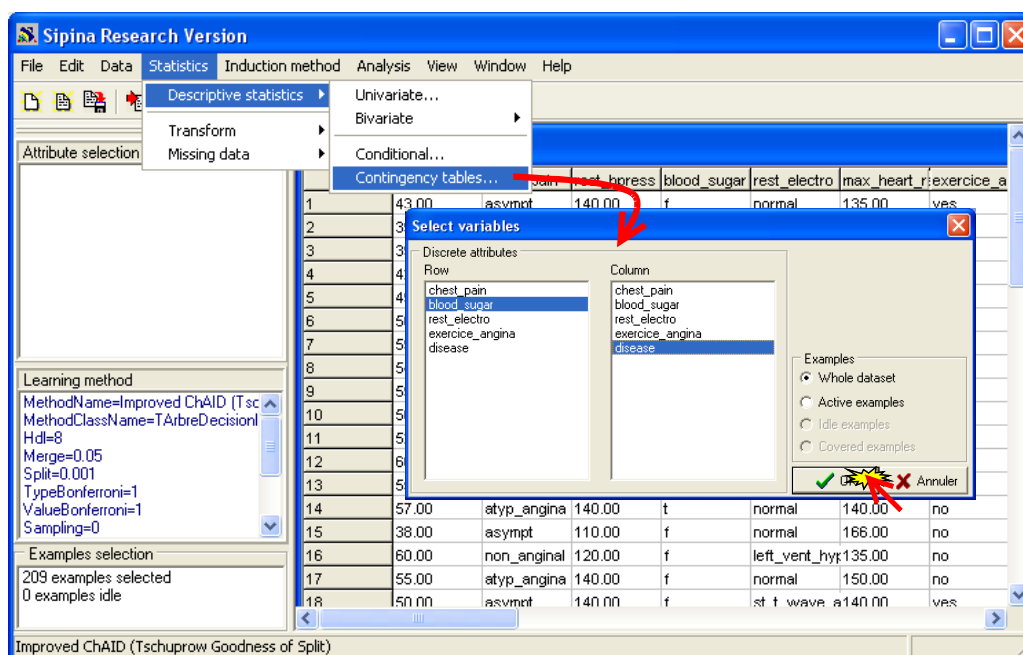


Using the contextual menu, we can display the conditional histogram (CONDITIONAL HISTOGRAM menu) in the new output window.

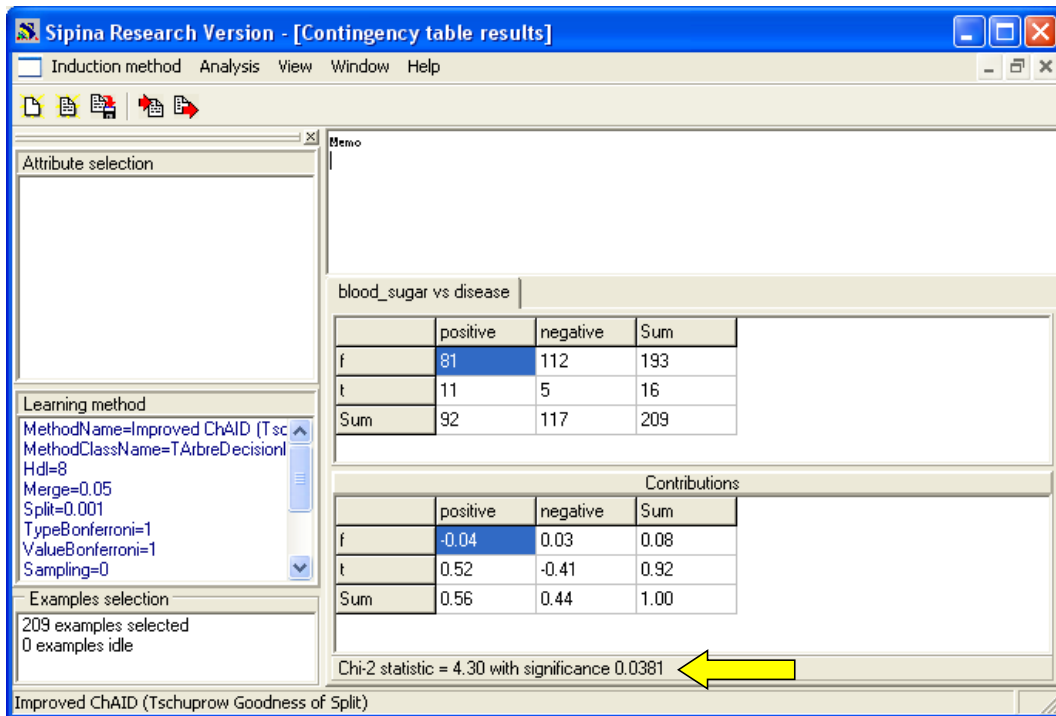


3.3.4 Two discrete variables: contingency table

This functionality enables to measure the association between two discrete variables through the contingency table. The chi-square statistic for independence test is computed. We activate the menu STATISTICS / DESCRIPTIVE STATISTICS / CONTINGENCY TABLES menu. In the dialog box, we select BLOOD_SUGAR and DISEASE.



We obtain the contingency table, the chi-square statistic, and the p-value of the test. We have also the contribution of each cell to the chi-square statistic.

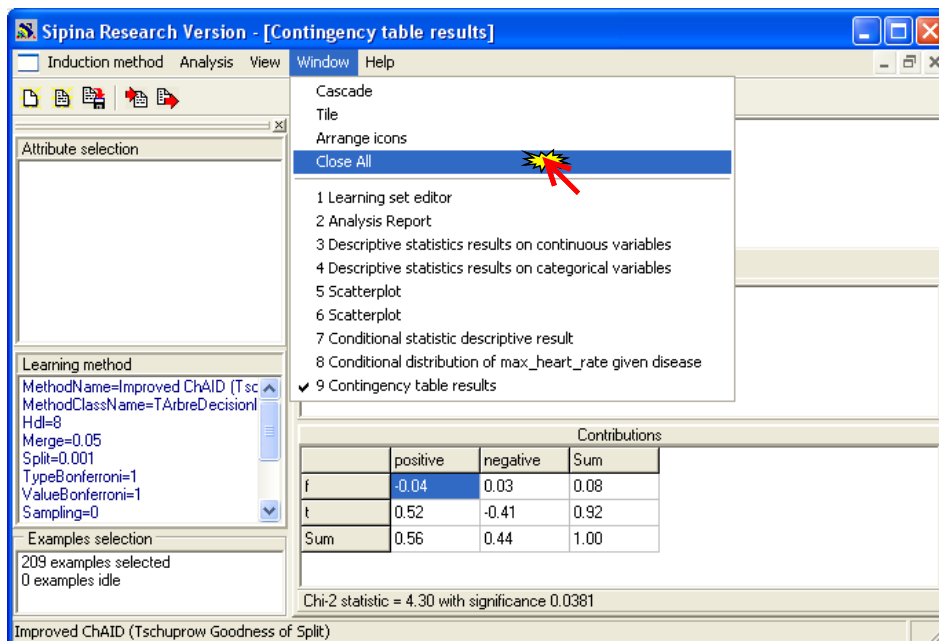


4 Descriptive statistics for a subpopulation

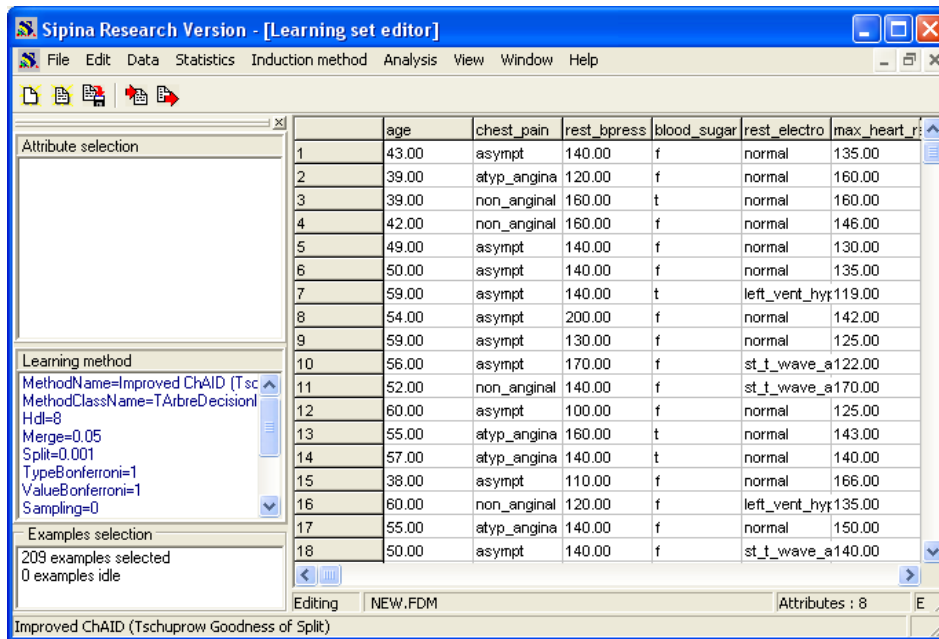
Each node of a classification tree corresponds to a subsample of the dataset. It will be very interesting to compare the characteristics of these groups using descriptive statistics. This functionality is very useful when we want to build interactively the tree.

4.1 Interactive tree induction

First of all, we must close all the windows in relation with the previous analysis. We click on the WINDOW / CLOSE ALL menu.

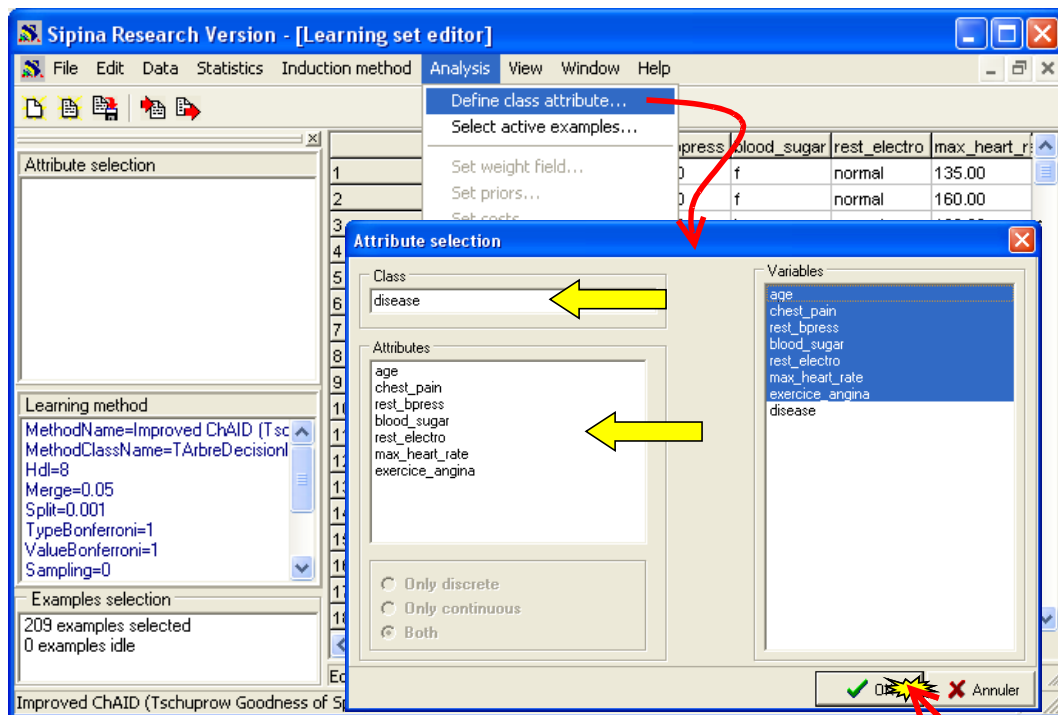


Only the main window, the data grid and the project explorer are visible.

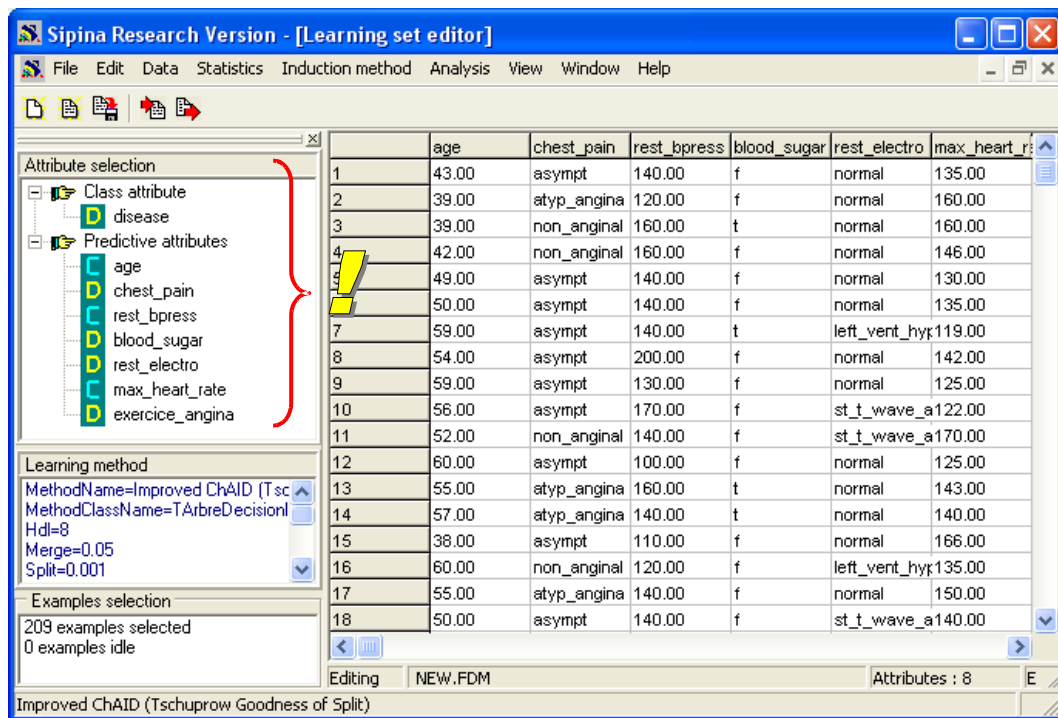


4.1.1 Selecting the variables of the analysis

In order to defining the target and the input attributes, we select the ANALYSIS / DEFINE CLASS ATTRIBUTE menu. In the dialog box, we set DISEASE as CLASS (TARGET), the others as ATTRIBUTES (INPUT).

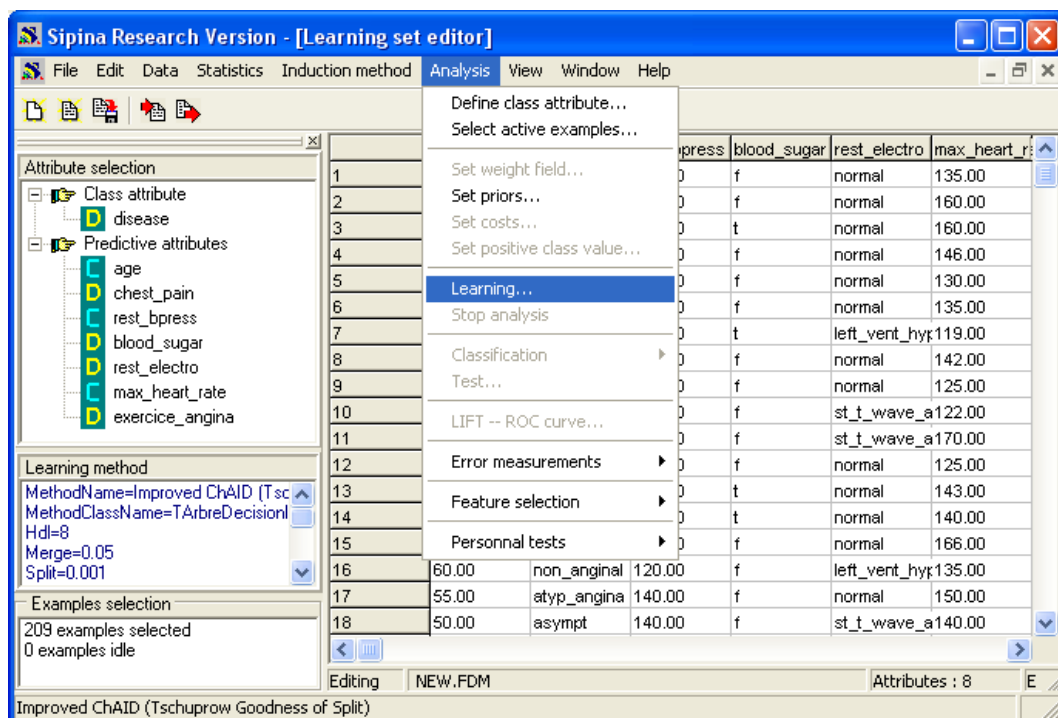


The selection is now visible in the top part of the project explorer. The letter "C" pinpoints a continuous attribute, "D" a discrete variable. The target attribute must be discrete.

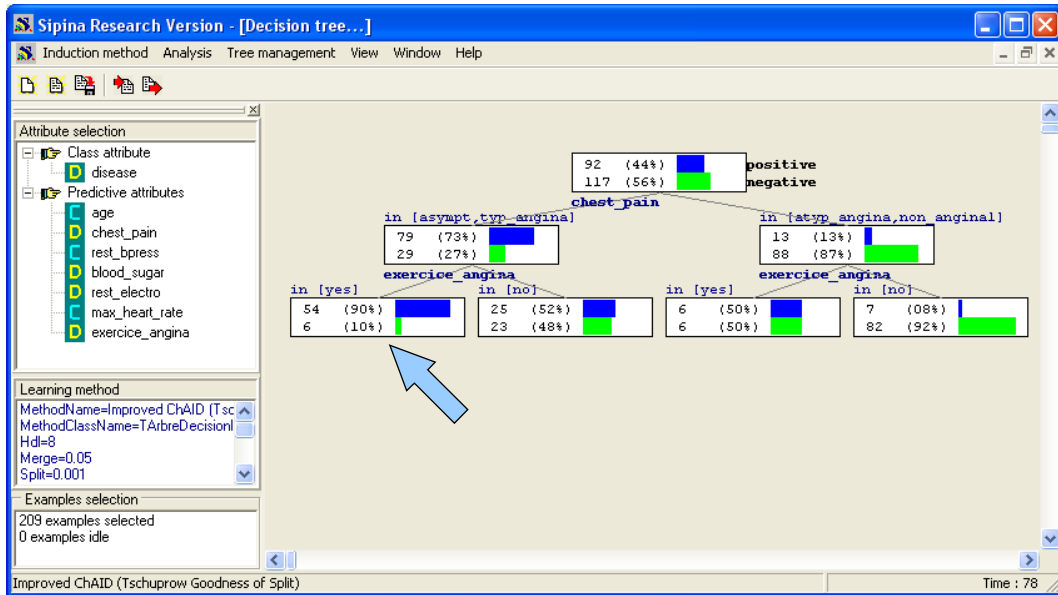


4.1.2 Tree induction

We click on the ANALYSIS / LEARNING menu. The learning phase is finalized.



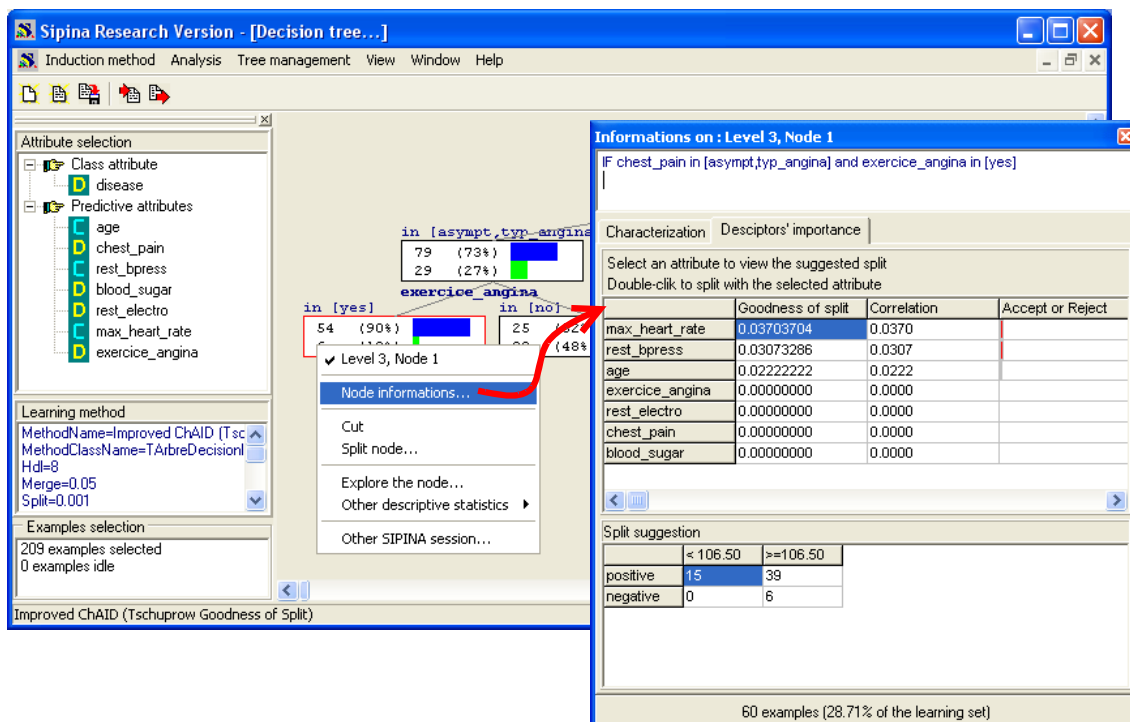
The tree is now displayed. The distribution of the values of DISEASE is available on each node.



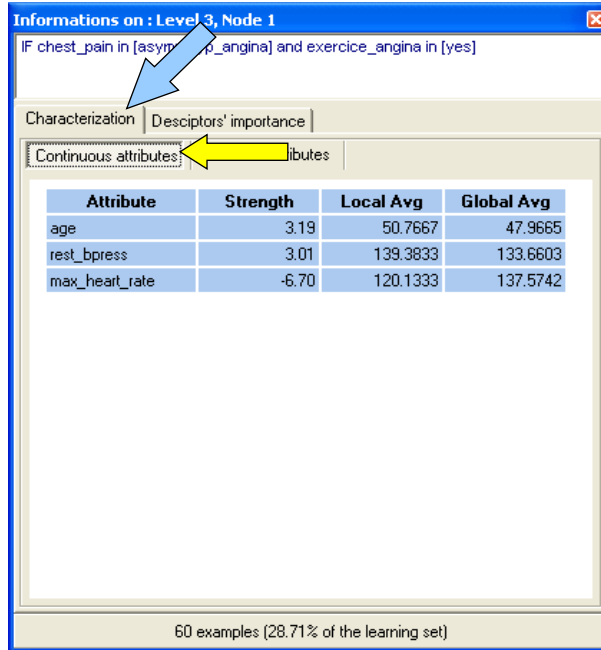
We want to analysis the leaf at the last level of the tree. The corresponding prediction rule is « **IF** CHEST_PAIN = (ASYMPT OR TYP_ANGINA) **AND** EXERCICE_ANGINA = YES **THEN** DISEASE = YES ». This group is defined by two variables. But, what about the other variables of the dataset? Are they really irrelevant for the characterization of this subpopulation? Computing descriptive statistics enables to answer to this question.

4.2 Node exploration - Elementary statistics

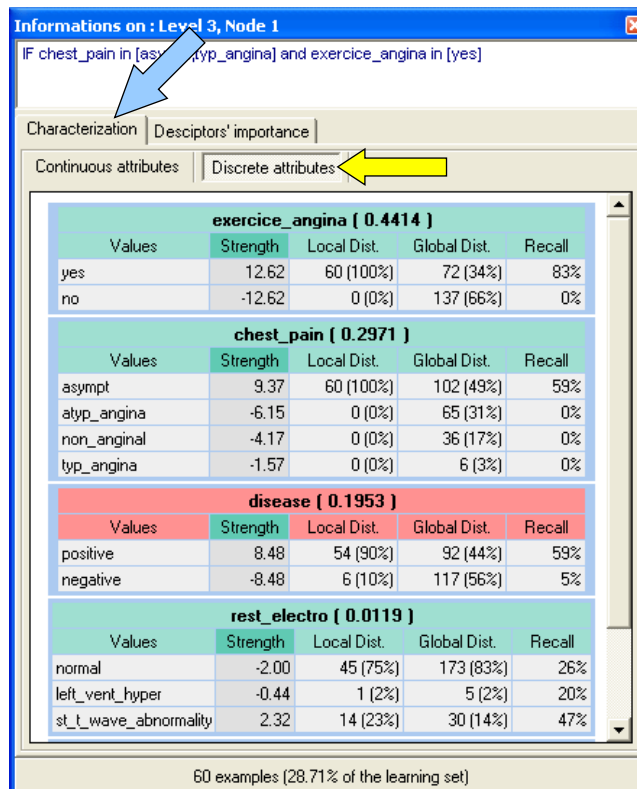
Each node of the tree matches to a subsample. The root node constitutes the whole dataset. In order to obtain descriptive statistics of the observations on a node, we select the node. Then we activate the contextual menu (right click). We select the NODE INFORMATIONS menu item. A new window appears. We observe the goodness-of-split for each predictive variable, the number of examples, some descriptive statistics, etc. **We select the CHARACTERIZATION tab.**



CONTINUOUS ATTRIBUTES. For each continuous attribute, we compare the local average (i.e. the mean of the variable for the subsample) and the global average (i.e. the mean of the variable for the whole dataset). In order to characterize the importance of the deviation, we compute also the t-test statistic (STRENGTH indicator) for a comparison of mean. It is not really a test because the samples are not independent. But it enables to order the variables according the importance of the difference. Indeed, the variables are not measured in the same unit and/or scale, the STRENGTH indicator can be understood as a normalized deviation. In this example, the mean of “age” for the whole dataset is 47.9. For the subgroup corresponding to the node, it is 50.76.



DISCRETE ATTRIBUTES. We compute a statistical indicator for the comparison of proportion here.



The variable REST_ELECTRO is really interesting. It is not visible in the tree. So it seems irrelevant. But when we compare the proportions, we observe that there is an over representation of the value ST_T_WAVE_ABNORMALITY for this subgroup. In the whole dataset, 14% of the examples have this characteristic. They are 23% for the examples related to the node.

The RECALL indicator says that 47% of the examples "REST_ELECTRO = ST_T_WAVE_ABNORMALITY" are located on this subgroup.

An additional indicator is used (J-MEASURE) in order to organize the variables. It has not really a valuable interpretation in our context.

4.3 Node exploration – Descriptive statistics

These comparative descriptive statistics are directly available. But they are mainly univariate. If we want to deeply analyze a subpopulation, it is (maybe) useful to compute the detailed descriptive statistics (univariate or bivariate) which were outlined previously (see section 3.2 and section 3.3).

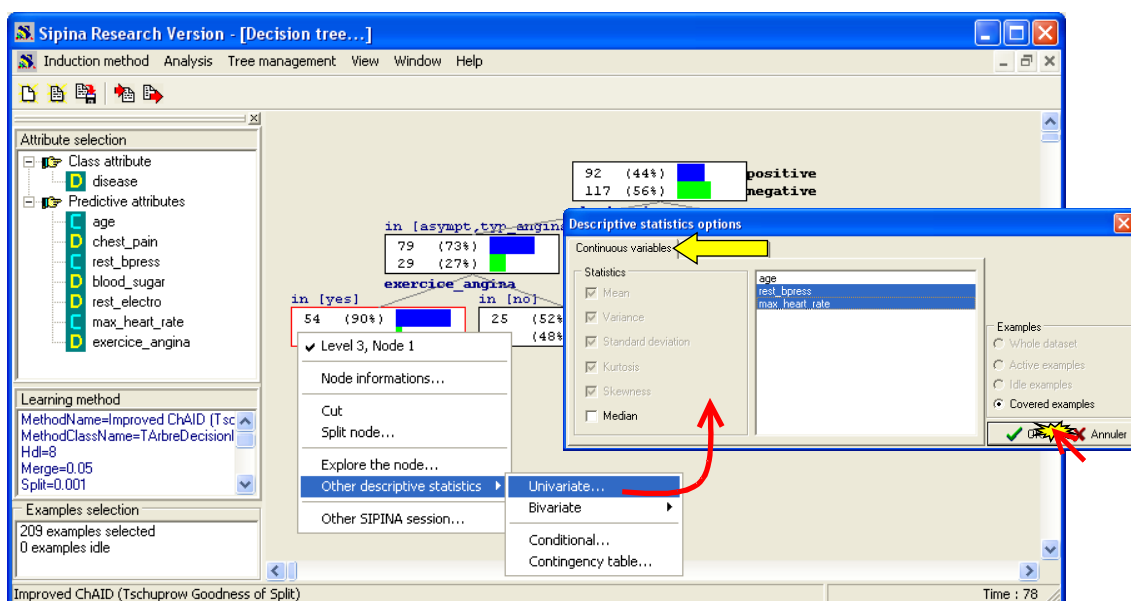
SIPINA enables to compute the previous descriptive statistics on each node. Of course, the computation is restricted to the covered examples i.e. the subpopulation highlighted by the node.

Let us repeat the same analysis than previously (see section 3.2 and section 3.3). But the calculations are now restricted to the sample corresponding to the rule "CHEST_PAIN = (ASYMPT OR TYP_ANGINA) AND EXERCICE_ANGINA = YES".

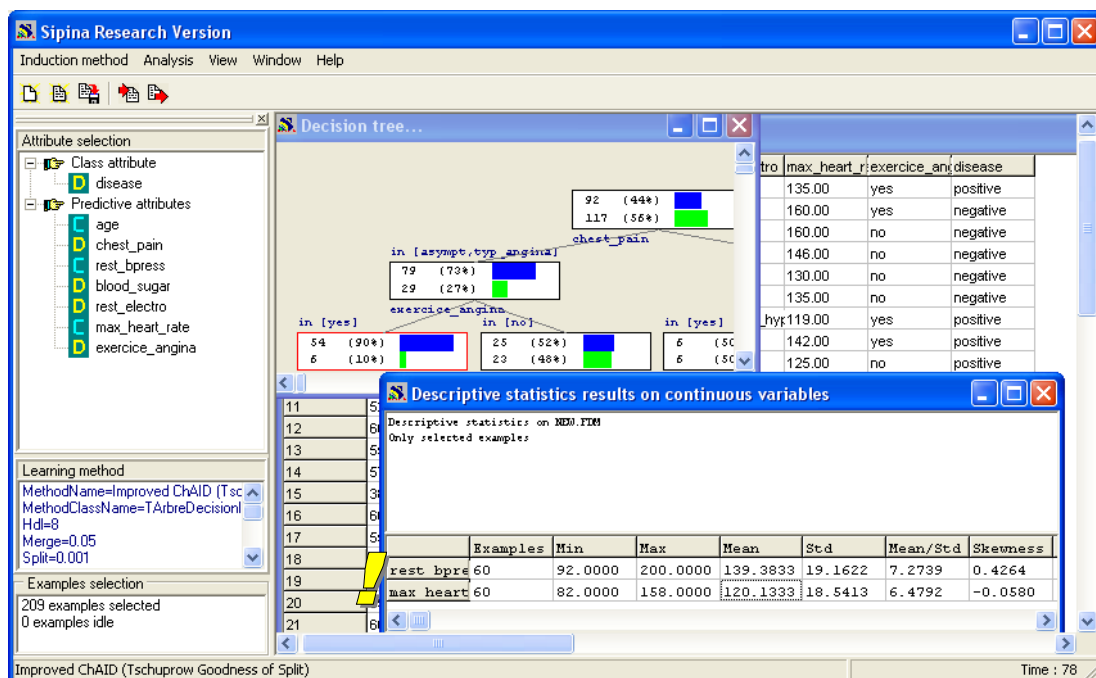
4.3.1 Univariate statistics

Compared with the preceding tool (section 4.2), this functionality is not really useful for the univariate statistics. We obtain the same results.

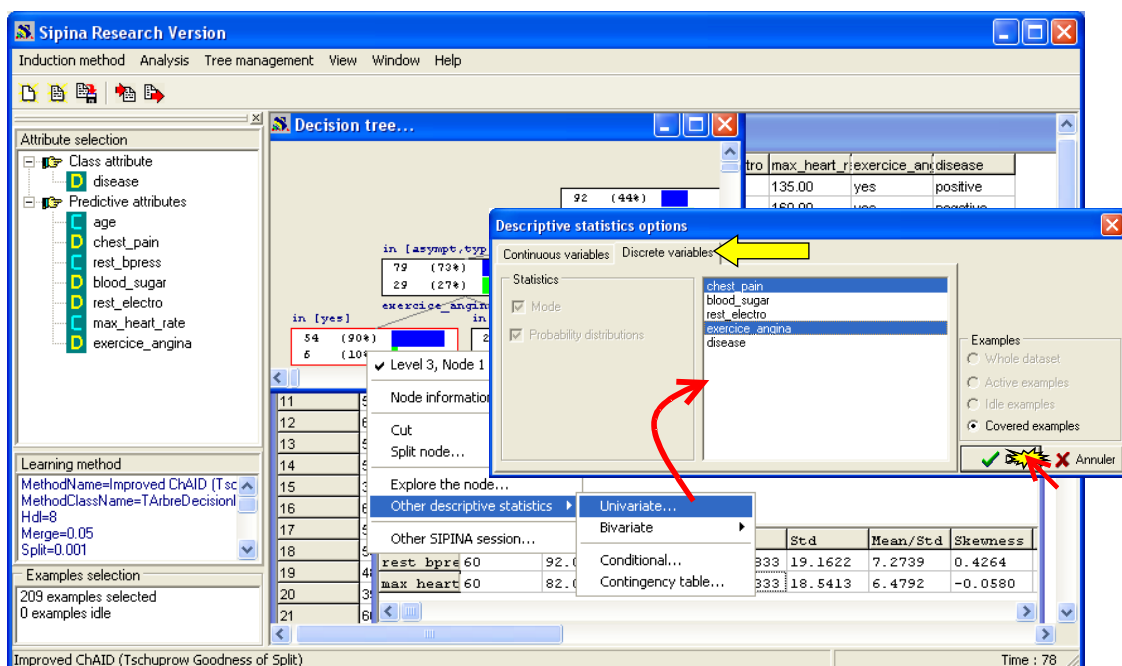
Continuous variables. In order to obtain the descriptive statistics related to a node. We select first the node. Then we activate the contextual menu (right click). We select the OTHER DESCRIPTIVE STATISTICS / UNIVARIATE menu item. In the dialog box, we observe that "COVERED EXAMPLES" option is activated. Only the 60 examples related to the node are used for the statistical computation.



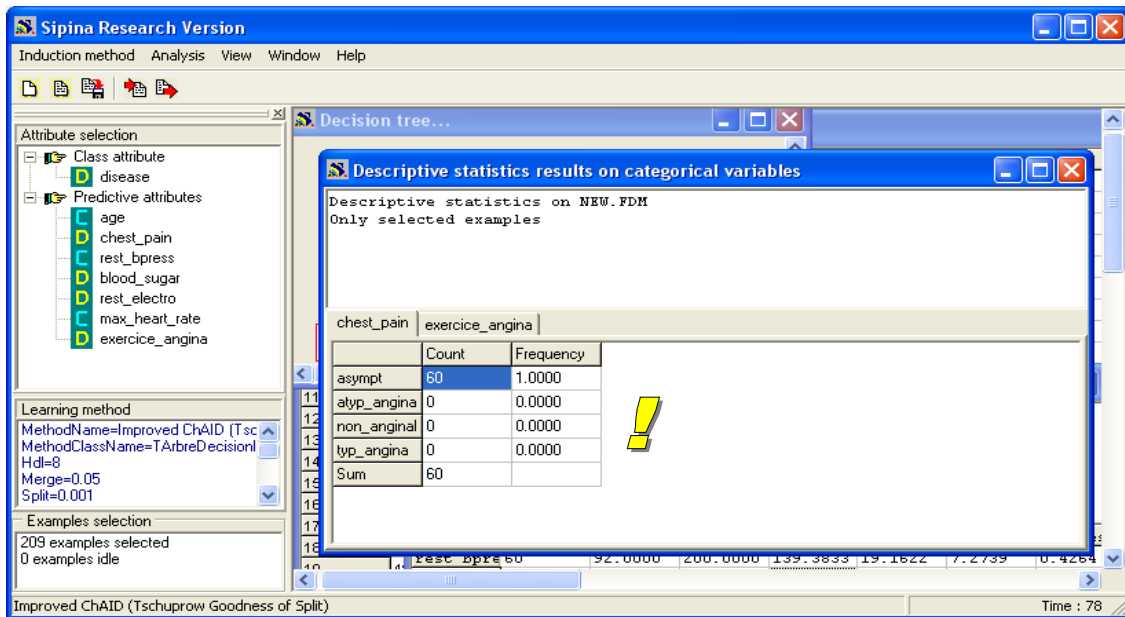
In the result window, we obtain all the descriptive indicators for each variable. We can compare these values with those computed for the whole dataset (see section 3.2.1).



Discrete variables. We follow the same approach for the discrete variables. We activate the OTHER DESCRIPTIVE STATISTICS / UNIVARIATE menu item in the contextual menu. We select the DISCRETE VARIABLE tab of course.



We obtain the distribution of values for each variable. We can compare these results with those obtained for the whole dataset (see section 3.2.2).

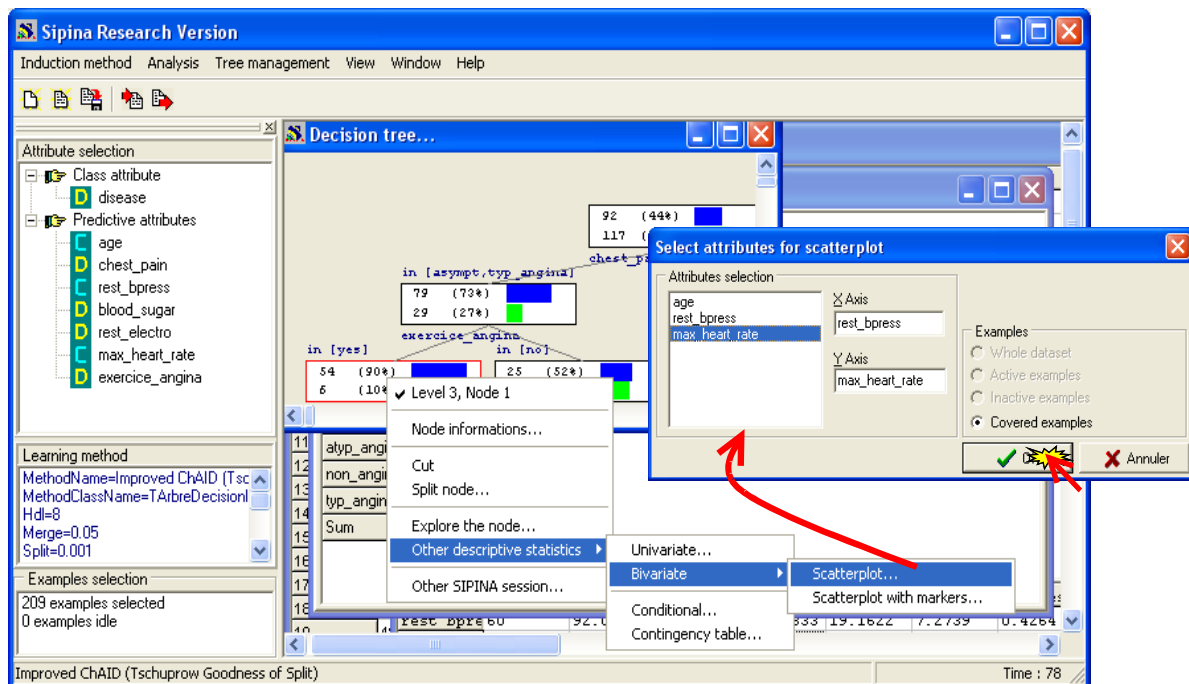


In this case, because CHEST_PAIN and EXERCICE_ANGINA are involved in the tree, only the values used in the path appear in the distribution.

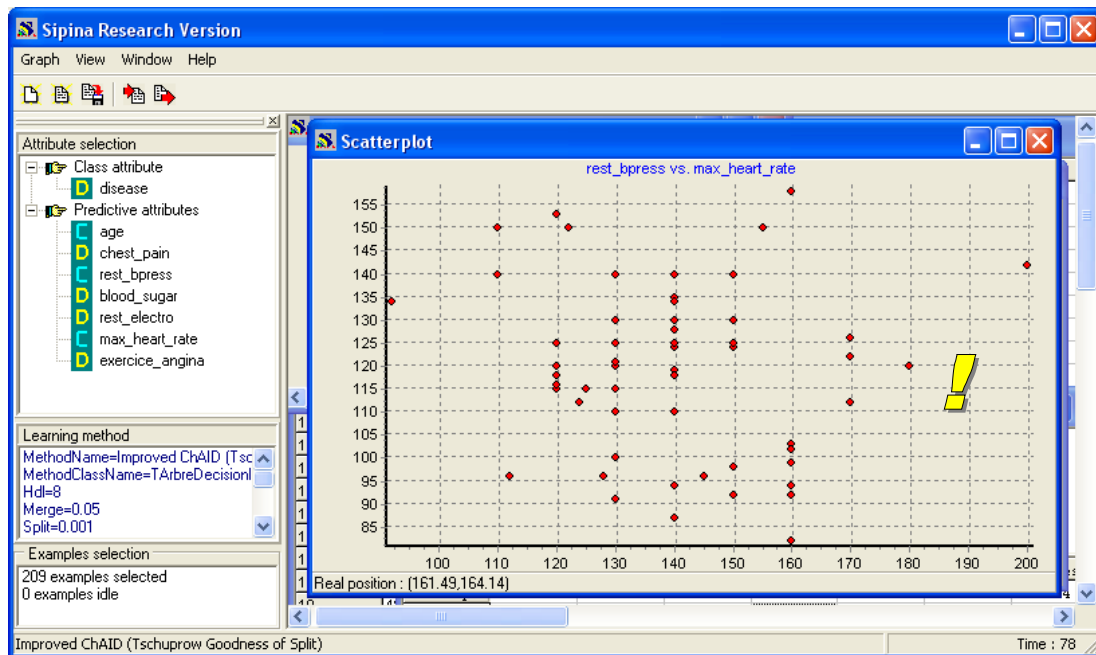
4.3.2 Bivariate statistics

This tool is useful for the bivariate statistics. We can analyze the association between some variables in each subgroups related to the nodes of the tree. Then we can carefully characterize each subpopulation. This functionality is essential when the interpretation of the results is at least as significant as the accuracy of the rules.

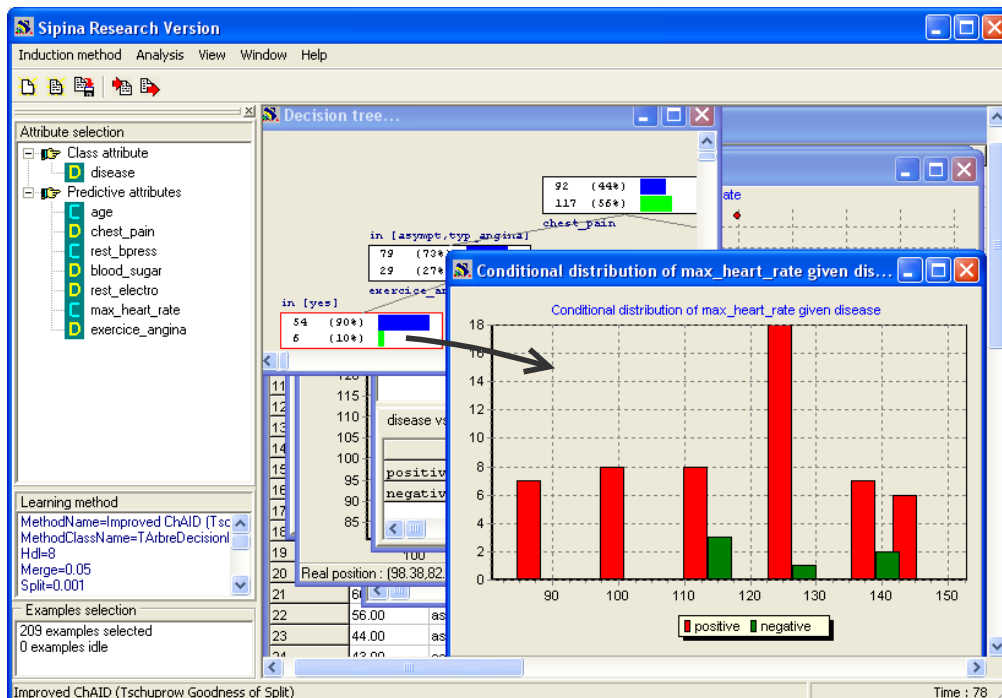
Scatter plot. We want to repeat the previous analysis (see section 3.3.1). In the contextual menu, we select the OTHER DESCRIPTIVE STATISTICS / BIVARIATE / SCATTERPLOT option.



The scatter plot is restricted to the 60 examples related to the node. It is different from the previous result (see section 3.3.1). The variables are not any more correlated in this configuration.



Other statistical indicators. In the same way, it is possible to find the tools highlighted previously (see sections 3.3.2, 3.3.3 et 3.3.4). E.g. MAX_HEART_RATE distribution according the DISEASE (see section 3.3.3 for the result on the whole dataset).



Note: Actually, it is possible to make these operations on any statistical software. It is simply necessary to make a query and to compute the statistical indicators on the subpopulation. The main interest of SIPINA is to automate all intermediate operations which, when they are repetitive, can end up quickly boring. This shortcut is very useful in practice.

5 Subsample related to a node

When we wish to refine the results, it can be necessary to go back on the data, notably to analyze in a deepened way the subpopulations described by the nodes of the tree. When the observations are recognizable (each case is associated to a label), we can even distinguish each individual.

In order to obtain the detailed description of the dataset, we select the node; we activate the EXPLORE THE NODE option in the contextual menu. The subsample is displayed in a new window.

The screenshot shows the Sipina Research Version interface. On the left, the 'Attribute selection' panel lists 'Class attribute' (disease) and 'Predictive attributes' (age, chest_pain, rest_bpress, blood_sugar, rest_electro, max_heart_rate, exercice_angina). The 'Learning method' section shows 'Improved CHAID (Tschuprow Goodness of Split)'. The main window displays a decision tree with a contextual menu open over a node, with 'Explore the node...' selected. A new window titled 'Subsample on Level 3, Node 1' is open, showing a table of local examples. The table has 60 rows and 8 columns: age, chest_pain, rest_bpress, blood_sugar, rest_electro, max_heart_rate, and disease. The filter condition is 'IF chest_pain in [asympt,typ_angina] and exercice_angina in [yes]'. A red arrow points to the 'Explore the node...' option in the menu.

	age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_rate	disease
1	43.00	asympt	140.00	f	normal	135.00	
7	59.00	asympt	140.00	t	left_vent_hype	119.00	
8	54.00	asympt	200.00	f	normal	142.00	
10	56.00	asympt	170.00	f	st_t_wave_abn	122.00	
18	50.00	asympt	140.00	f	st_t_wave_abn	140.00	
21	66.00	asympt	140.00	f	normal	94.00	
22	56.00	asympt	155.00	t	normal	150.00	
24	43.00	asympt	120.00	f	normal	120.00	
25	54.00	asympt	140.00	f	normal	118.00	
32	54.00	asympt	130.00	f	normal	91.00	
33	48.00	asympt	160.00	f	normal	92.00	
34	38.00	asympt	110.00	f	normal	150.00	
36	46.00	asympt	120.00	f	normal	115.00	
40	49.00	asympt	140.00	f	normal	140.00	
41	65.00	asympt	170.00	t	normal	112.00	
43	65.00	asympt	140.00	t	normal	87.00	
49	55.00	asympt	140.00	f	normal	130.00	
66	56.00	asympt	150.00	f	st_t_wave_abn	124.00	
75	31.00	asympt	120.00	f	normal	153.00	
79	43.00	asympt	150.00	f	normal	130.00	
86	55.00	asympt	140.00	f	normal	110.00	
88	48.00	asympt	160.00	f	normal	103.00	
98	46.00	asympt	110.00	f	normal	150.00	
100	48.00	asympt	160.00	f	normal	102.00	

The subsample can be saved in a new file (*.fdm file format).

This screenshot shows the 'Subsample on Level 3, Node 1' window with the 'Save covered examples' checkbox checked. The table of local examples is the same as in the previous screenshot. A red arrow points to the 'Save covered examples' checkbox.

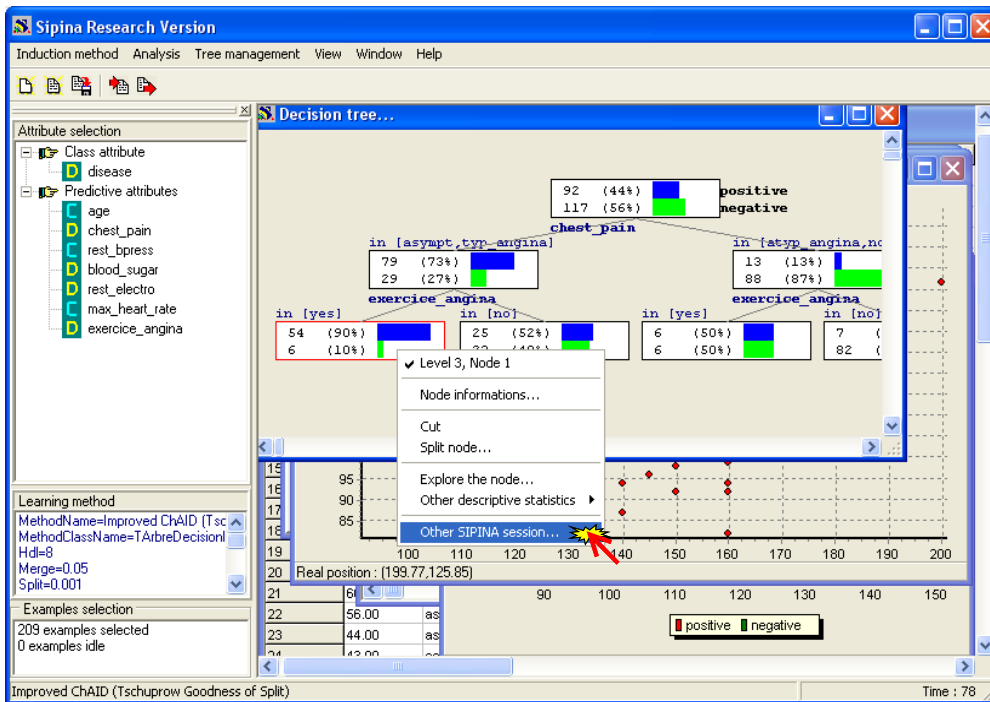
	age	chest_pain	rest_bpress	blood_sugar	rest_electro	max_heart_rate	disease
1	43.00	asympt	140.00	f	normal		
7	59.00	asympt	140.00	t	left_vent_hype		
8	54.00	asympt	200.00	f	normal		
10	56.00	asympt	170.00	f	st_t_wave_abn		
18	50.00	asympt	140.00	f	st_t_wave_abn		
21	66.00	asympt	140.00	f	normal		
22	56.00	asympt	155.00	t	normal		
24	43.00	asympt	120.00	f	normal		
25	54.00	asympt	140.00	f	normal		

The data file is automatically named with the identifier of the node. It is placed in the same directory as the source dataset.

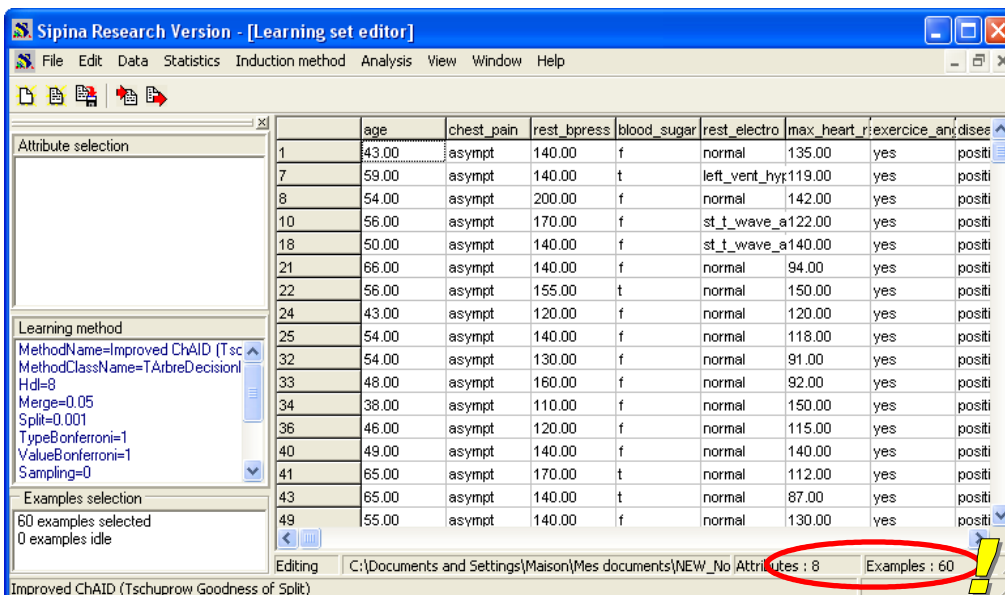
6 A new analysis of a subpopulation

In certain situations, we want to launch a new analysis on a subpopulation related to a node of the tree. For instance, we want to explain/predict the REST_ELECTRO variable for the subsample described by the rule “CHEST_PAIN = (ASYMPT OR TYP_ANGINA) **AND** EXERCICE_ANGINA = YES”.

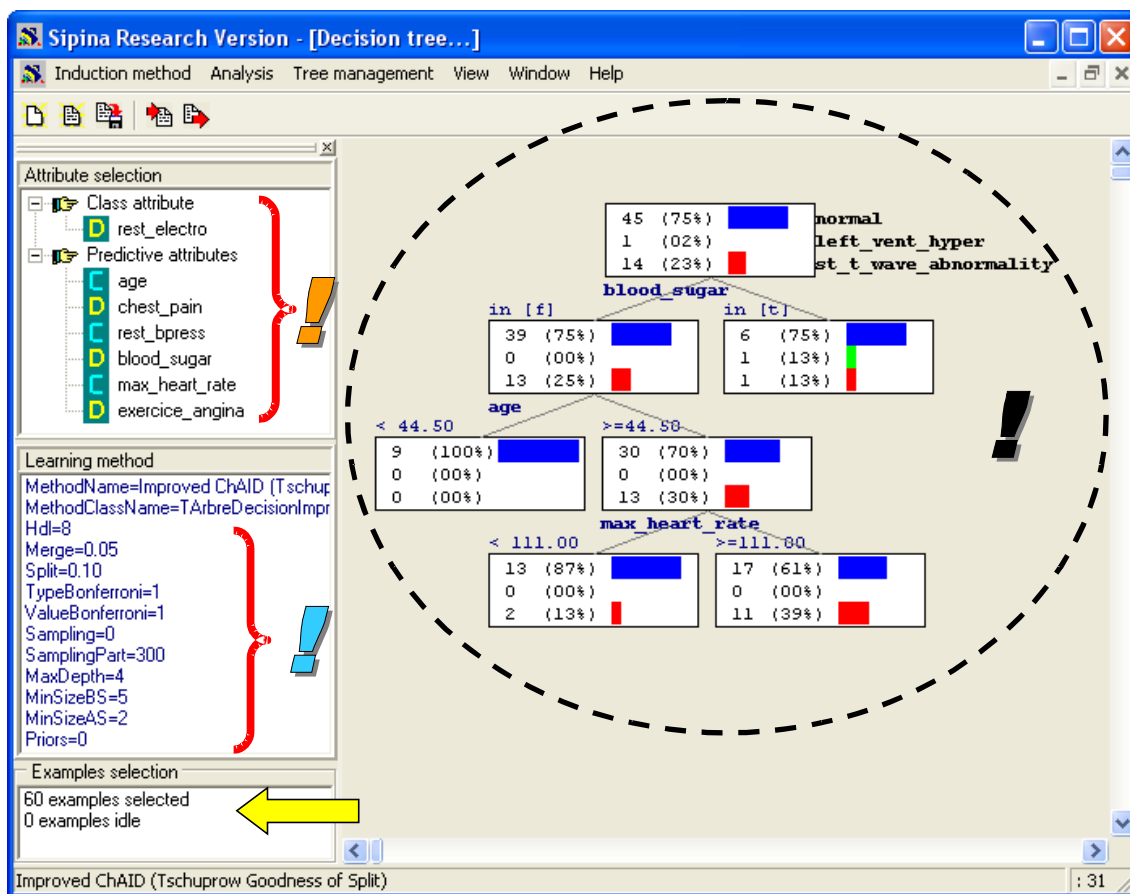
With the contextual menu, we select the OTHER SIPINA SESSION option.



A new SIPINA session is launched. The subsample (8 variables and 60 examples) is automatically downloaded.



We define again the TARGET and the INPUT variables. Then we select the adequate parameters of the learning algorithm. We obtain, for instance, the following classification tree.



7 Conclusion

In this tutorial, we wanted to describe the descriptive statistics tools of SIPINA. These features are not really extraordinary. But combined with the interactive exploration of a tree of decision, they turn out very productive.