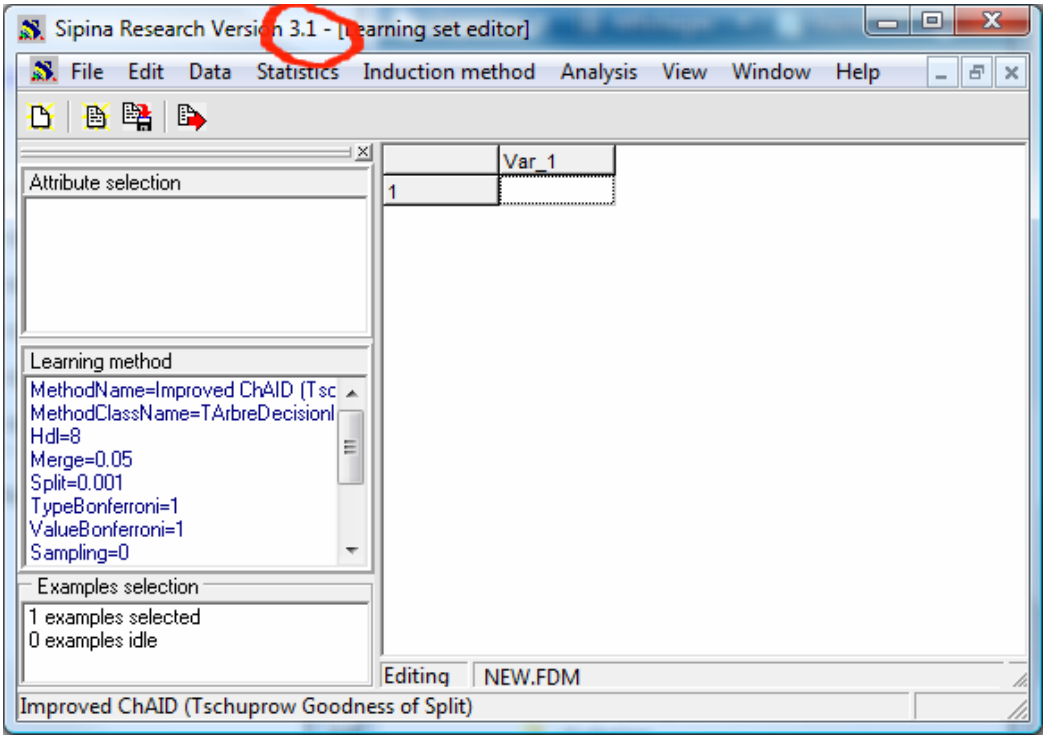
1 Introduction

Handling Missing values in SIPINA.

Caution: We must use the 3.1 version (or later) of Sipina Research for this tutorial. Please, check the version number in the title bar.



Dealing with missing values is a difficult problem. The programming in itself is not a problem; we just report the missing value by a specific code. In contrast, the treatment before or during data analysis is very complicated. We must take into consideration two facets[1]:

- The nature of the missing value. The simplest way to consider missing value is MCAR (missing completely at random) i.e. none of the variables have missing scores related to the value of the variable itself. It is a purely theoretical point of view. The MAR (missing at random) is a more realistic hypothesis. We cannot predict the location of missing scores after controlling the values of the other variables. These two kinds of mechanism are said to be ignorable if the parameters that govern the missing data process are unrelated to the parameters to be estimated. At this time, it is not required to model first the missing value mechanism before implementing the data mining process.

[1] See P.D. Allison, « Missing Data », in Quantitative Applications in the Social Sciences Series n°136, Sage University Paper, 2002.
See also http://faculty.chass.ncsu.edu/garson/PA765/missing.htm

- The second aspect that we must consider is the type of the machine learning algorithm or statistical method that we implement in the data mining process. Indeed, some approaches for handling missing value are less or more appropriate according to the learning algorithm. For instance, in the linear regression context, the "listwise deletion" approach (or "casewise deletion" i.e. instances with at least one missing value is removed form the dataset), which seems much unsophisticated, has good properties even if the subsequent dataset is dramatically reduced (Allison, 2001).

TANAGRA has essentially an educational purpose; I did not want to introduce automated tools for managing missing data. It does not seem desirable that the students can click a button and remove the issue casually. They must prepare its data in full awareness of what he does before launching a statistical treatment in good conditions.

I had not the same point of view when I programmed SIPINA (which is anterior to Tanagra). Various techniques are available in order to handle missing values. In this tutorial, we show how to implement them; and what are their consequences on the induced decision tree (C4.5 algorithm; Quinlan, 1993).

We note that the handling of the missing values is done in a data preparation phase with Sipina. Another approach, not implemented here, is embedding the missing value handling in the core of the learning process (such as the surrogate split mechanism of CART for instance; Breiman and al., 1984).

2 Dataset

Our data file comes form the Gilles Hunault's website (University of Angers[2]). We want to predict snoring of individuals based on their characteristics (age, weight, gender, etc.). We have extracted 30 observations; we removed some values completely at random. We treat two data files in this tutorial:

- RONFLEMENT_ALL.FDM is the data file with all the values. We will use it to create the reference decision tree.
- RONFLEMENT_WITH_MISSING.FDM is the data file with some missing values. We use this data file in order to show the different strategies to processing missing values. The aim is to subsequently create a decision tree which is as similar as possible with the reference decision tree.

These files are gathered in an archive[3]. We use the binary file format (FDM). We have prepared the files for easy manipulation in this tutorial. But SIPINA can import data files in text format (tab separator) with missing data. The missing value is symbolized by the character "?" or the blank character.

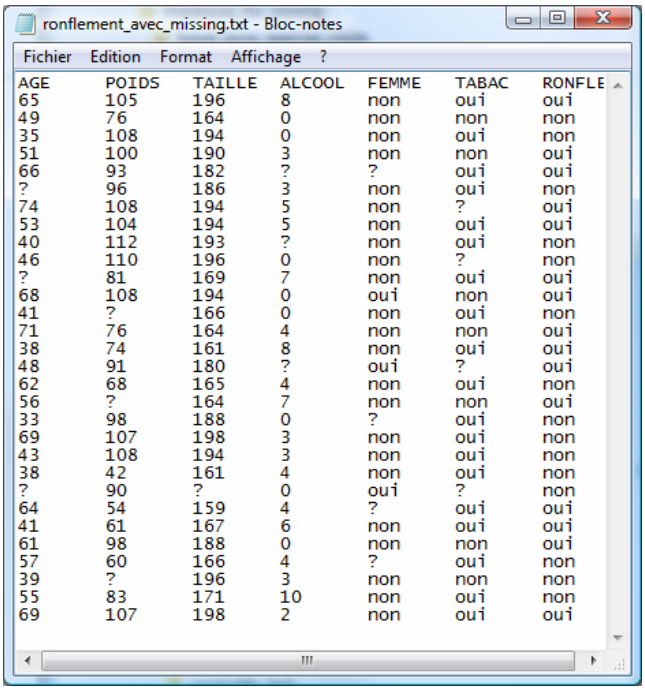
In addition, we can use also the add-in to transfer the data from Excel to Sipina[4]. In this case, the missing value is symbolized by an empty cell into the spreadsheet.

[2] http://www.info.univ-angers.fr/~gh/Datasets/datasets.htm

[3] http://eric.univ-lyon2.fr/~ricco/dataset/ronflement_missing_data.zip

[4] The installation of the add-in is described here: http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_installation.htm; its use is described here: http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_processing.htm.

We show bellow the dataset with the missing values. They are identified with the symbol "?".

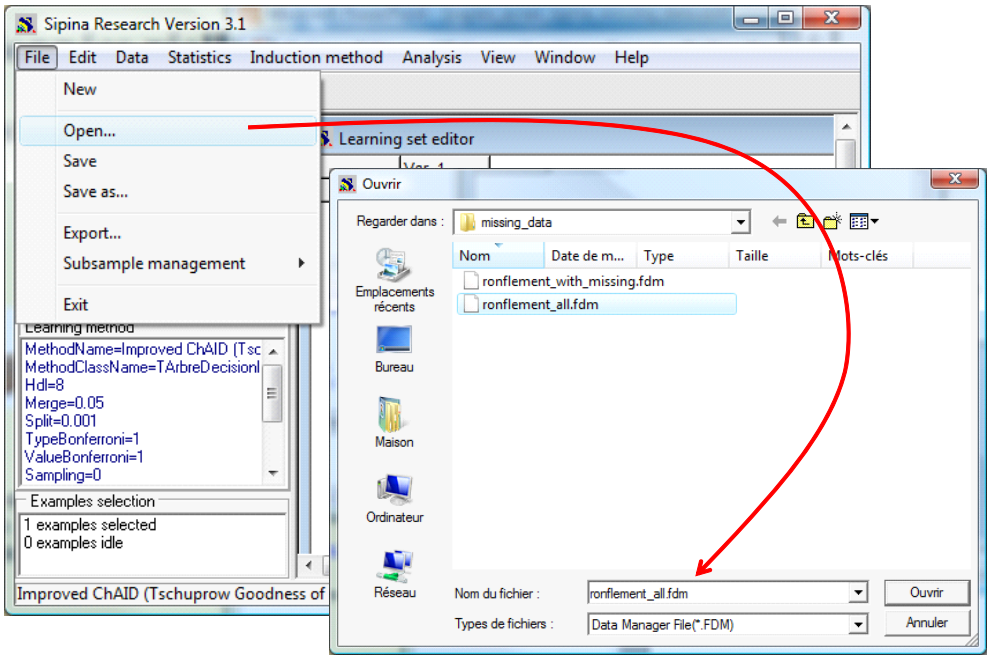


AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
65	105	196	8	non	oui	oui
49	76	164	0	non	non	non
35	108	194	0	non	oui	non
51	100	190	3	non	non	oui
66	93	182	?	?	oui	oui
?	96	186	3	non	oui	non
74	108	194	5	non	?	oui
53	104	194	5	non	oui	oui
40	112	193	?	non	oui	non
46	110	196	0	non	?	non
?	81	169	7	non	oui	oui
68	108	194	0	oui	non	oui
41	?	166	0	non	oui	non
71	76	164	4	non	non	oui
38	74	161	8	non	oui	oui
48	91	180	?	oui	?	oui
62	68	165	4	non	oui	non
56	?	164	7	non	non	oui
33	98	188	0	?	oui	non
69	107	198	3	non	oui	non
43	108	194	3	non	oui	non
38	42	161	4	non	oui	non
?	90	?	0	oui	?	non
64	54	159	4	?	oui	oui
41	61	167	6	non	oui	oui
61	98	188	0	non	non	oui
57	60	166	4	?	oui	non
39	?	196	3	non	non	non
55	83	171	10	non	oui	non
69	107	198	2	non	oui	oui

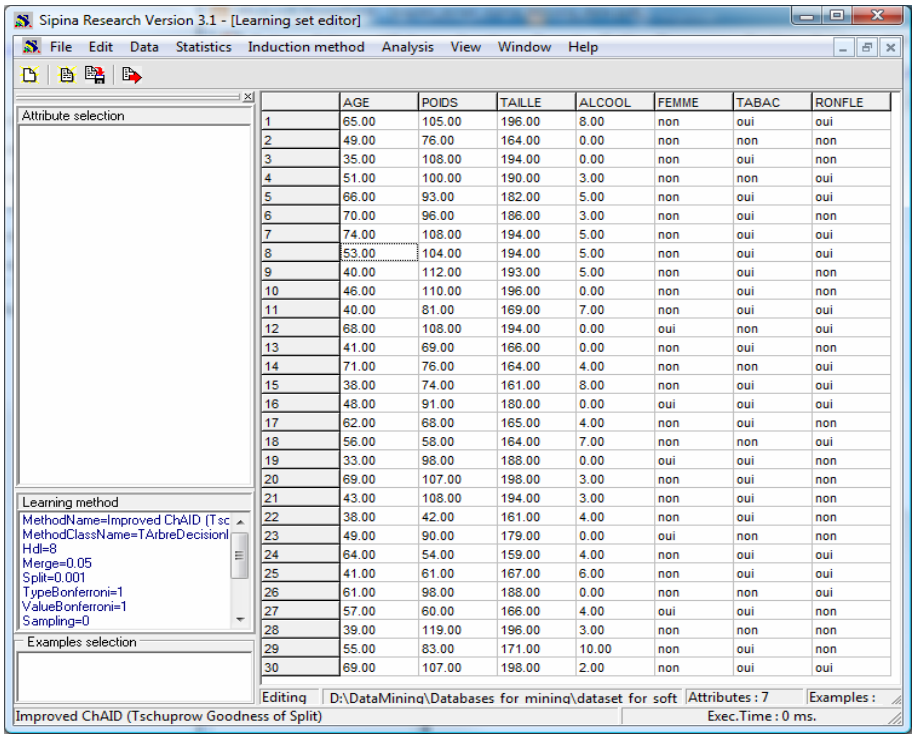
3 Treatment of the dataset without missing values

3.1 Loading the data file

After launching Sipina, we load the data file without missing values (RONFLEMENT_ALL.FDM) by clicking on the FILE / OPEN menu.



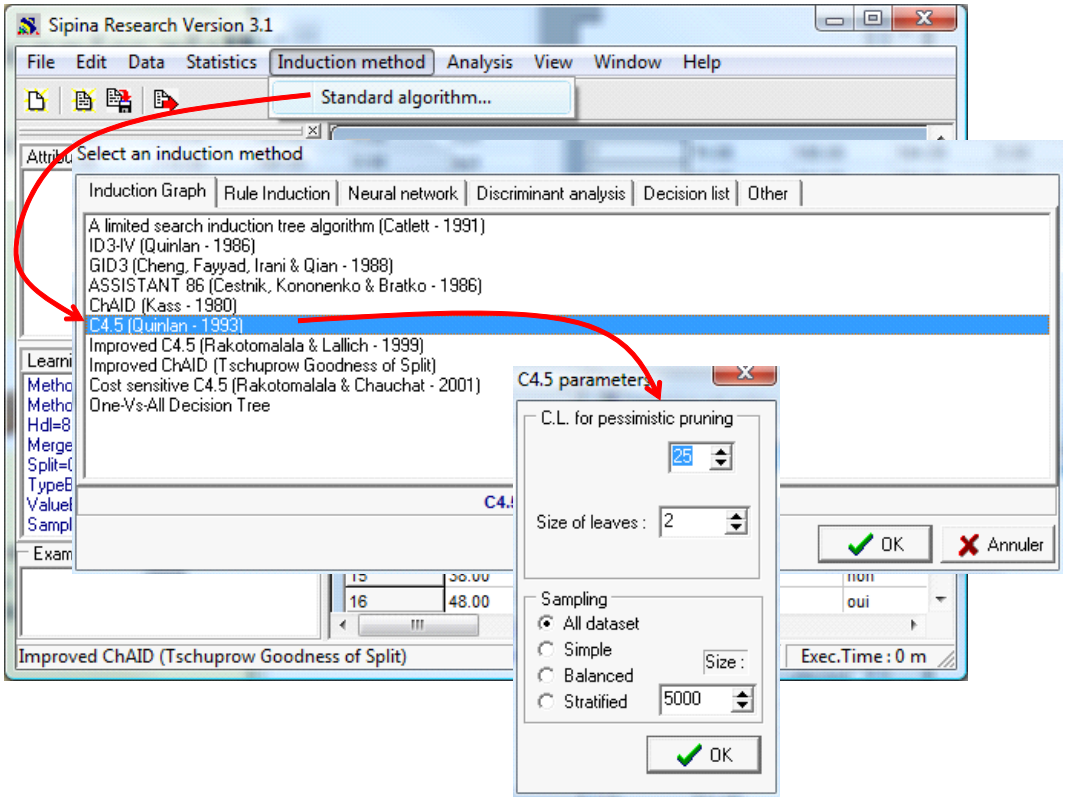
The 30 instances are displayed into the visualization grid.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	5.00	non	oui	oui
6	70.00	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	oui	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	5.00	non	oui	non
10	46.00	110.00	196.00	0.00	non	oui	non
11	40.00	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	69.00	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	0.00	oui	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	58.00	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	oui	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	49.00	90.00	179.00	0.00	oui	non	non
24	64.00	54.00	159.00	4.00	non	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	oui	oui	non
28	39.00	119.00	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

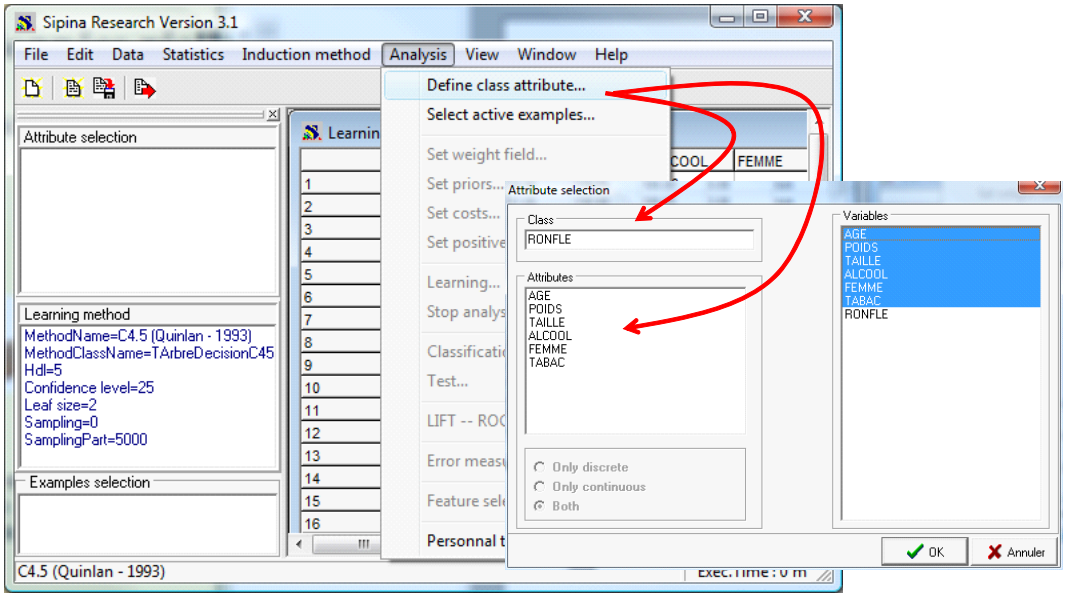
3.2 Choosing the learning algorithm

In order to select the learning approach, we click on the INDUCTION METHOD / STANDARD ALGORITHM menu. We choose the C4.5 algorithm. We left the default settings. We note however that a node is not split if one or both the subsequent leaves contain less than two instances.



3.3 Specifying the target and input variables

In order to specify the target and the input attributes, we click on the ANALYSIS / DEFINE CLASS ATTRIBUTES menu. By drag and drop, we set RONFLE as class attribute (target), the other ones as descriptors (input).



3.4 Decision tree learning

We launch the analysis by clicking on the ANALYSIS / LEARNING menu.

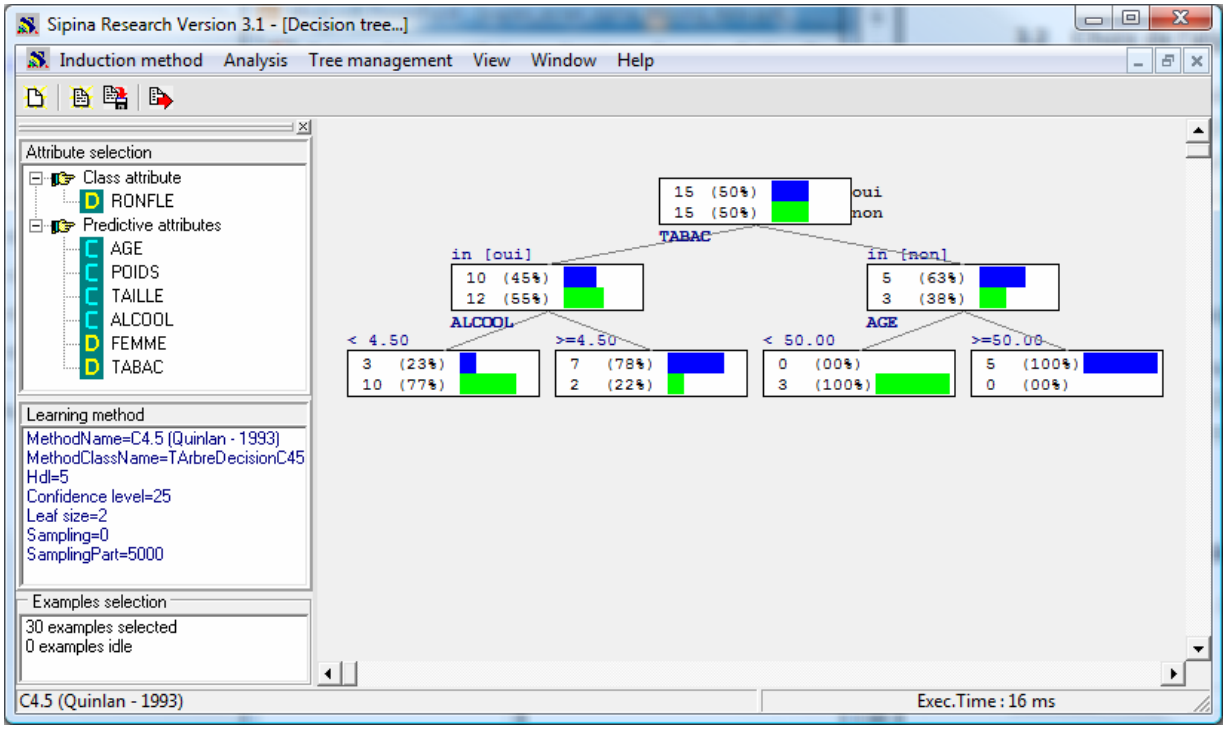


Figure 1 - The decision tree obtained from the full dataset

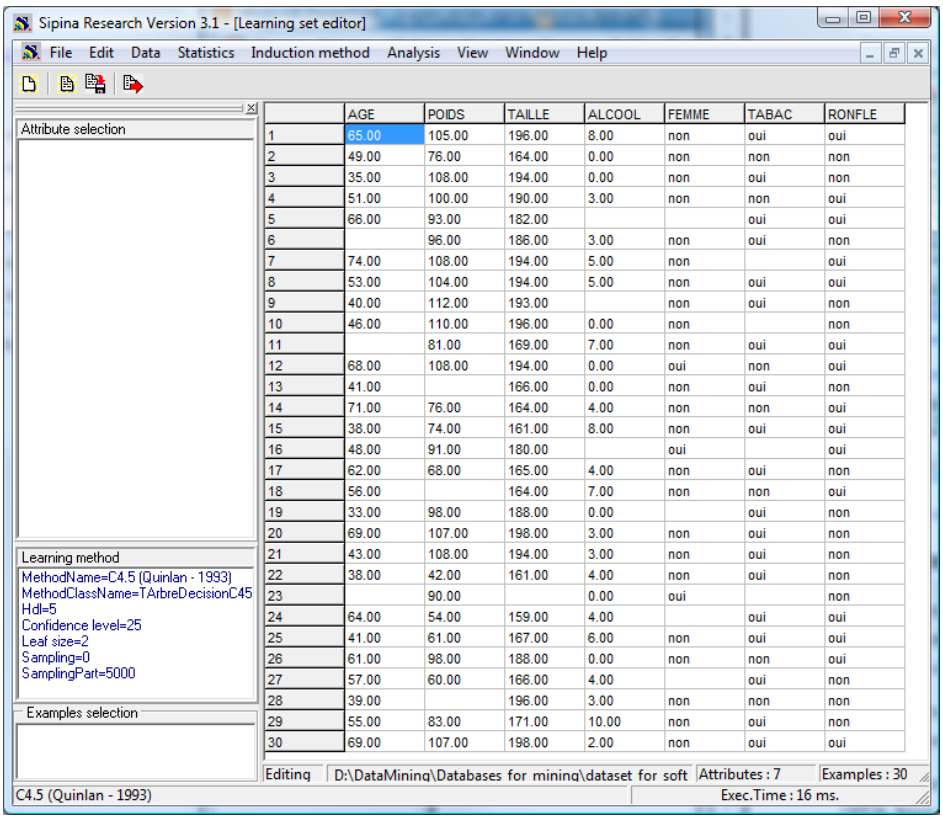
We obtain a classification tree with 3 levels (Figure 1). It is the reference tree. The relevant variables are TABAC (tobacco), ALCOOL (alcohol) and AGE (age).

4 Dealing with missing values

The idea now is to work on the data file including some missing values. The goal is to evaluate at which conditions the features proposed by Sipina to treat missing values allow to supply a dataset which, when we launch the C4.5 algorithm, to find again the reference tree (Figure 1).

We stop the current analysis by clicking on the ANALYSIS / STOP ANALYSIS menu. We click on the FILE / NEW menu in order to clear the visualization grid.

Then, we load the data file with some missing values by clicking on the FILE / OPEN menu. We select RONFLEMENT_WITH_MISSING.FDM.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00			oui	oui
6		96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non		oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00		non	oui	non
10	46.00	110.00	196.00	0.00	non		non
11		81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00		166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00		oui		oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00		164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00		oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23		90.00		0.00	oui		non
24	64.00	54.00	159.00	4.00		oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00		oui	non
28	39.00		196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

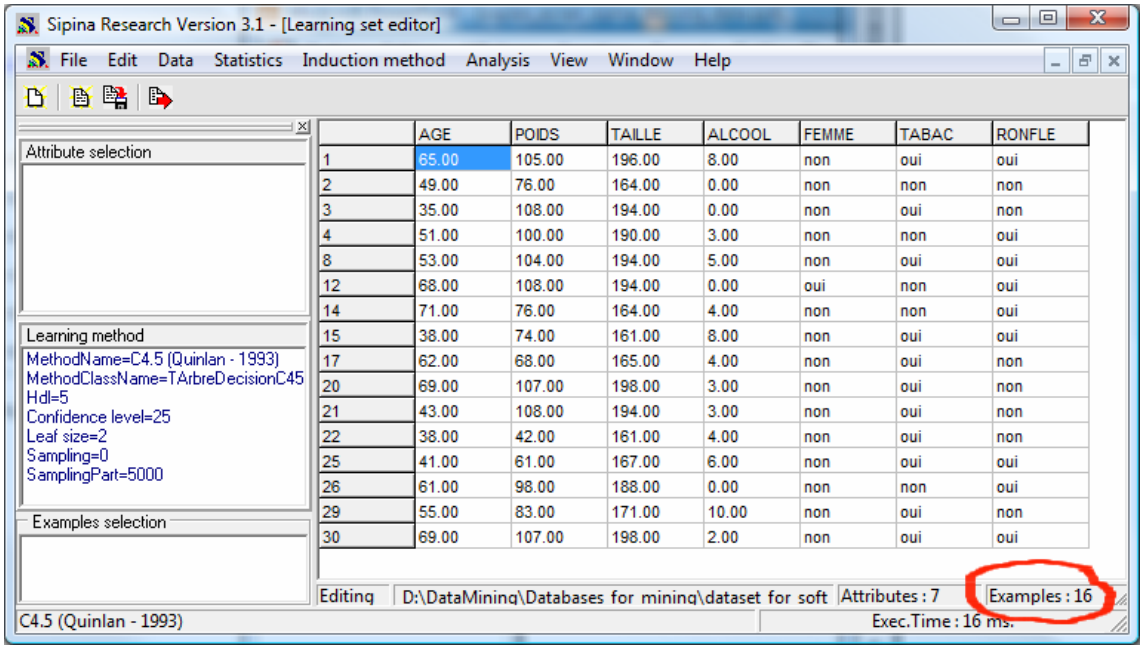
Figure 2 - Data visualization grid with some missing values

The empty cells symbolize the missing values.

4.1 Listwise deletion

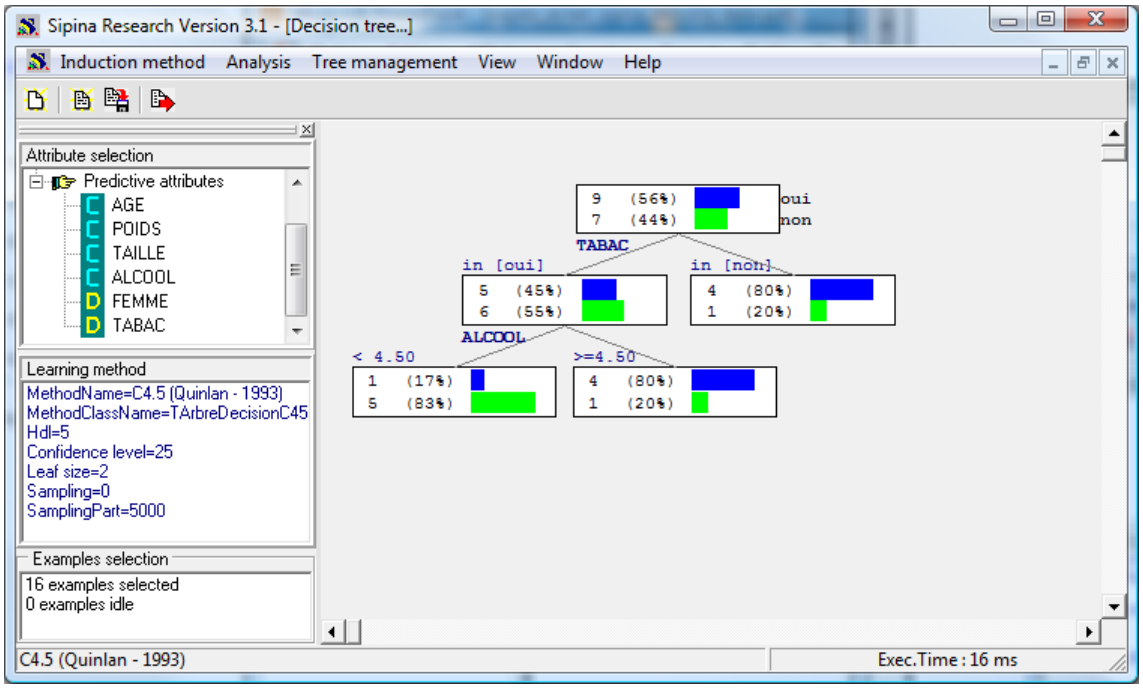
The first strategy seems rather unsophisticated. We remove all the instances with at least one missing value. If the missing values are randomly disseminated in the dataset, only a few instances are removed. At the opposite, if they are concentrated on one variable, we can empty out the data file!

We click on the STATISTICS / MISSING DATA / DELETE EXAMPLES menu. The rows with at least one missing value are removed. The resulting dataset contains 16 examples i.e. 50% of the original instances. The reduction is significant.



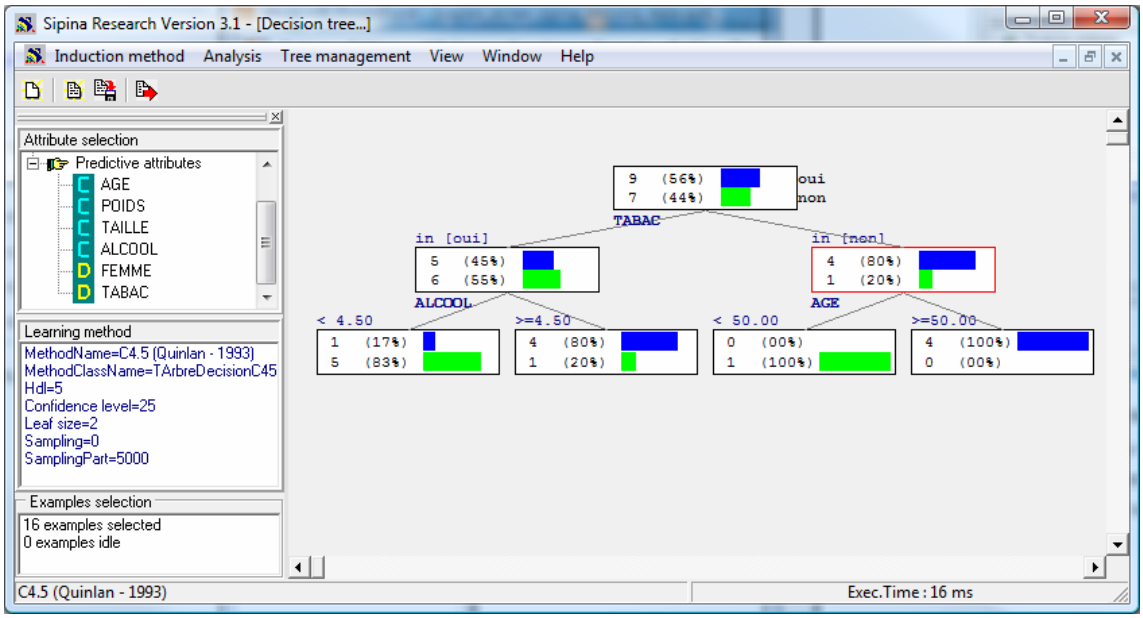
	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

We define the target and the input attributes (ANALYSIS / DEFINE CLASS ATTRIBUTES). Then we launch the learning process (ANALYSIS / LEARNING). We obtain the following decision tree.



Finally, despite the drastic reduction of the number of instances, we get a tree nearly identical to the reference tree (Figure 1). C4.5 could not split the node to the right because it has too few examples. It can not produce leaves with less than 2 instances according to the default settings of the method. If we

force even when the splitting, we would get the next tree.



We find the tree computed on the full dataset (without missing values) (Figure 1). The class values frequencies of each node is different because the number of instances is not the same in the used datasets.

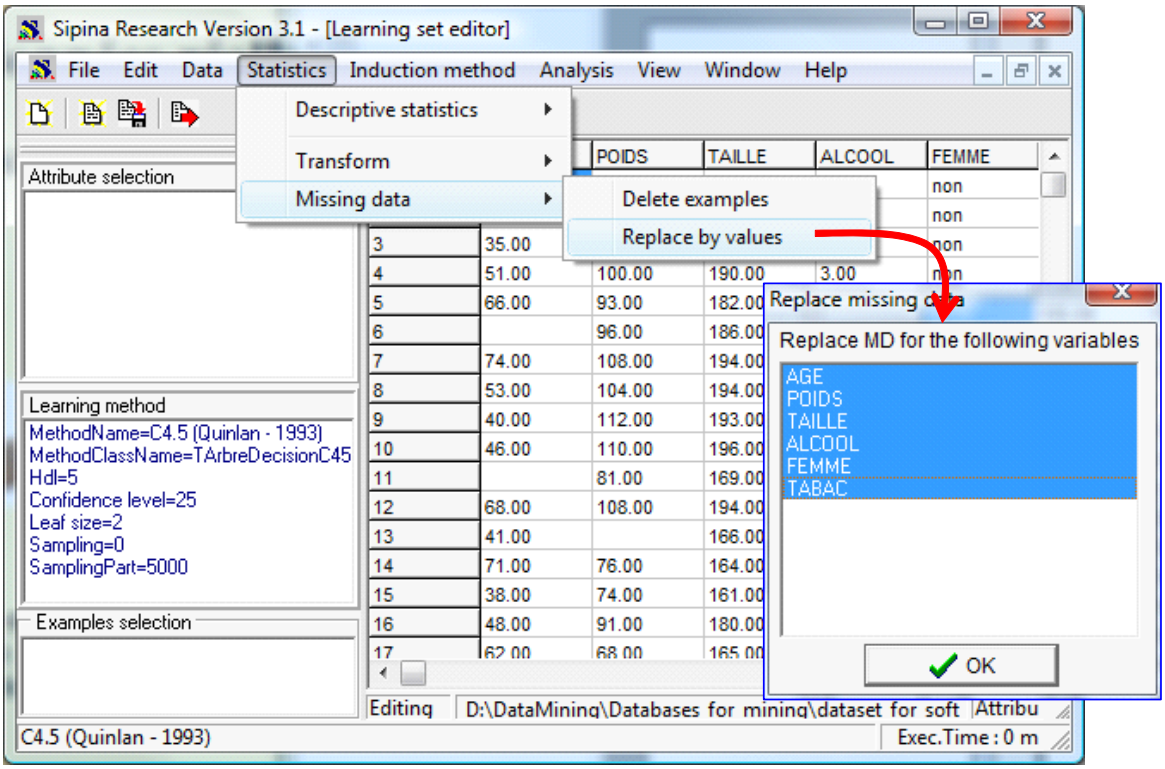
When the missing values mechanism is MCAR, we observe that the listwise deletion strategy allows obtaining a tree which is similar to the reference tree. But we must adjust the settings of the learning algorithm to the diminution of the number of instances.

4.2 Data imputation – A

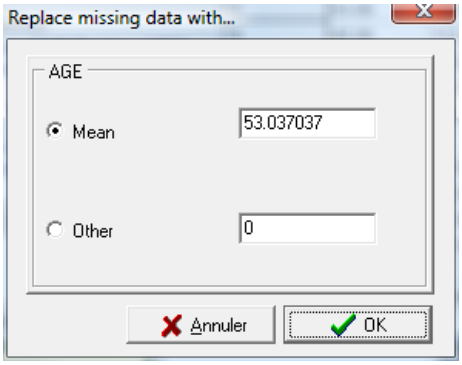
Now, we study the various data imputation strategies proposed by Sipina. The first strategy is also very simplistic: we replace the missing value by the mean for the continuous attributes; we use the most frequent value (the mode) for the discrete attributes.

We stop the current analysis (ANALYSIS / STOP ANALYSIS) and we clear the data grid (FILE / NEW). We load again the data file with missing values (FILE / OPEN). We find again our dataset with missing values (Figure 2).

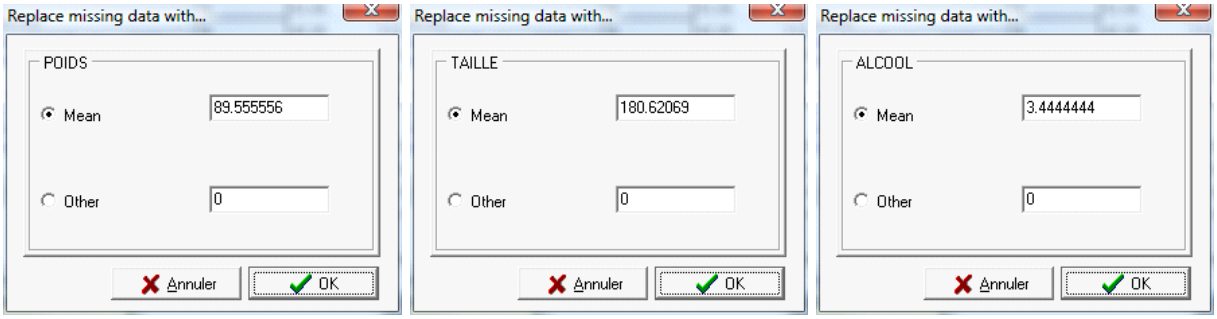
We click now on the STATISTICS / MISSING DATA / REPLACE BY VALUES menu. A dialog box with the attributes with at least one missing value appears. We notice that the column RONFLE, the target attribute, has no missing values. It does not appear in the list.



We select the variables and we click on the OK button. For each variable, Sipina suggests an imputation value. For a continuous variable, this is the mean. We click on OK for AGE.

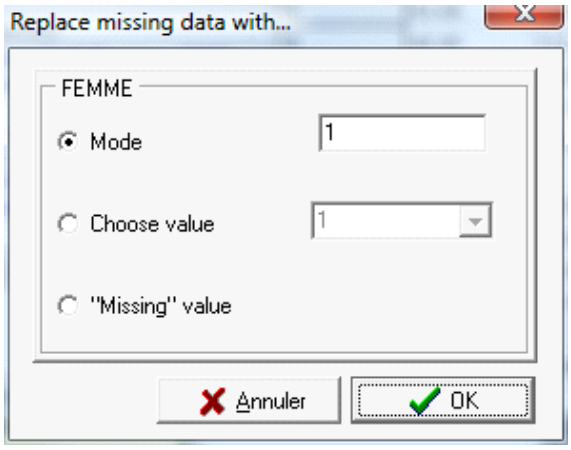


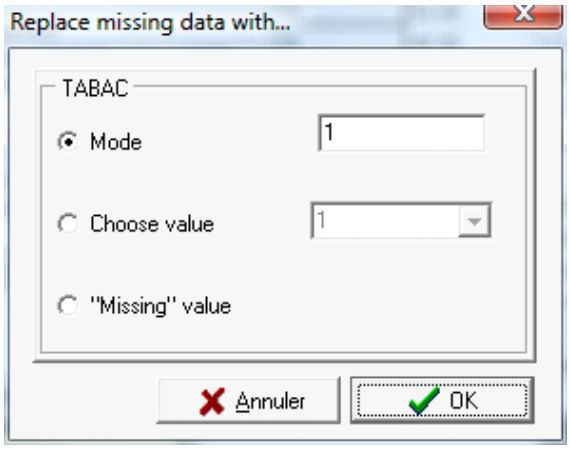
Ibid for POIDS, TAILLE and ALCOOL.



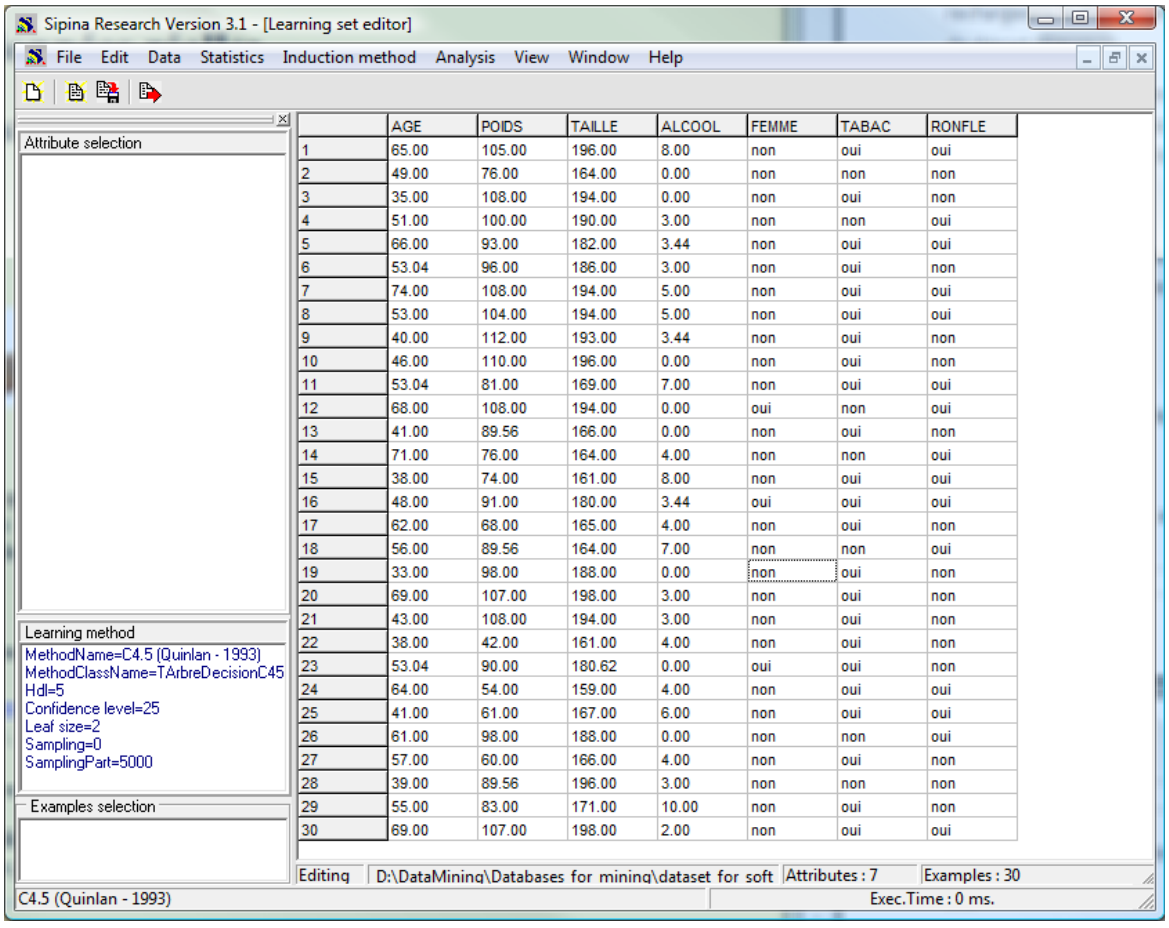
When we deal with a discrete attribute (FEMME), Sipina suggests using the most occurred value into the column (without accounting the missing values). Here, it is FEMME = 1 (gender = female). We use

the same process for the TABAC attribute.



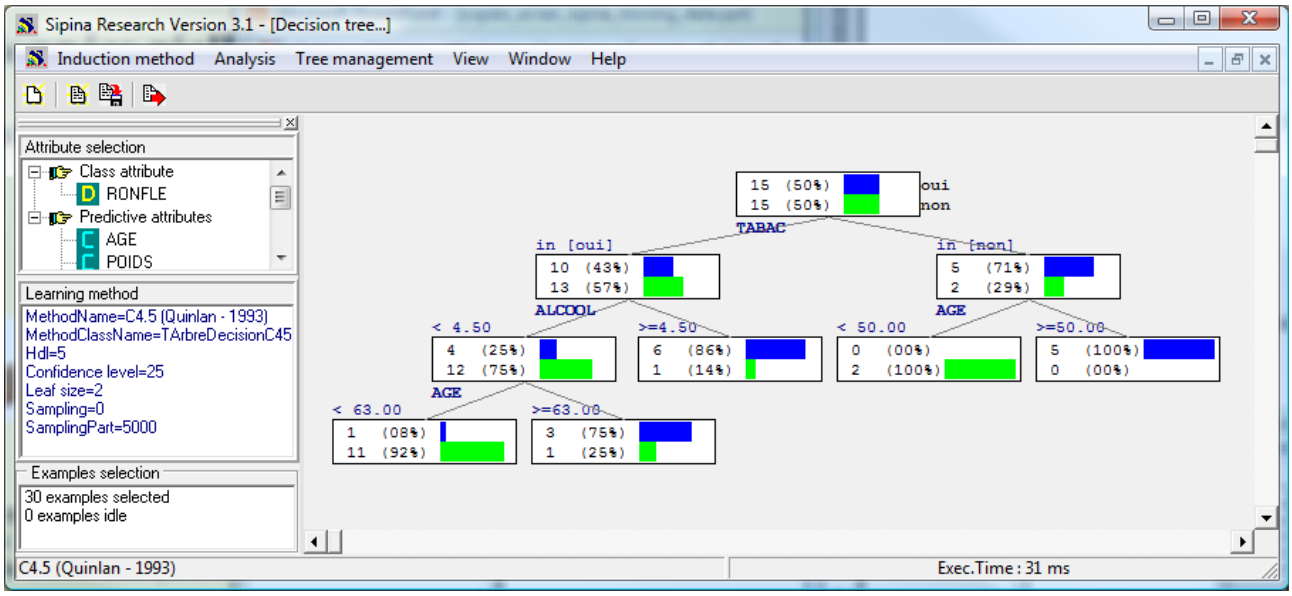


The empty cells of the data grid are then filled out.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	3.44	non	oui	oui
6	53.04	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	oui	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	3.44	non	oui	non
10	46.00	110.00	196.00	0.00	non	oui	non
11	53.04	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	89.56	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	3.44	oui	oui	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	89.56	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	non	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	53.04	90.00	180.62	0.00	oui	oui	non
24	64.00	54.00	159.00	4.00	non	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	non	oui	non
28	39.00	89.56	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

Now, we can launch the learning process (selecting the target and the input attributes, launching the analysis). We obtain the following decision tree.

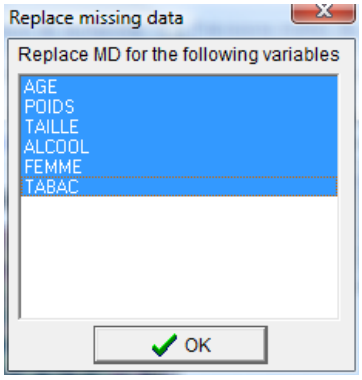


The three first levels are identical to the reference tree (Figure 1). But the frequencies are not the same. The additional examples for TABAC = 1, which is the most occurred value before the data imputation, affects the splitting process in the left part of the tree. An additional split is performed.

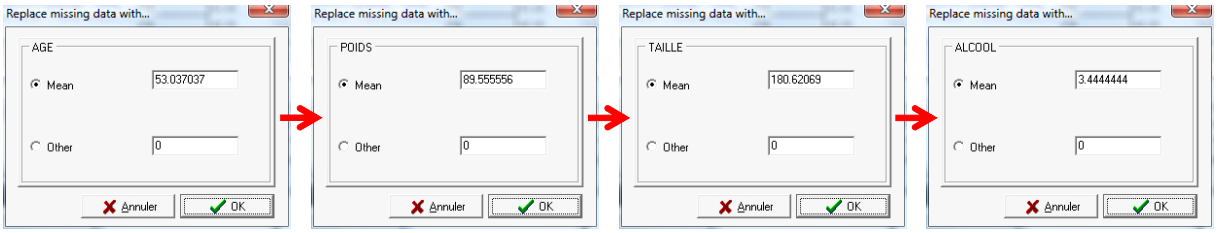
4.3 Data imputation – B

In this section, the missing values for discrete attributes are replaced by a new value "missing". For the continuous attribute, we use the mean.

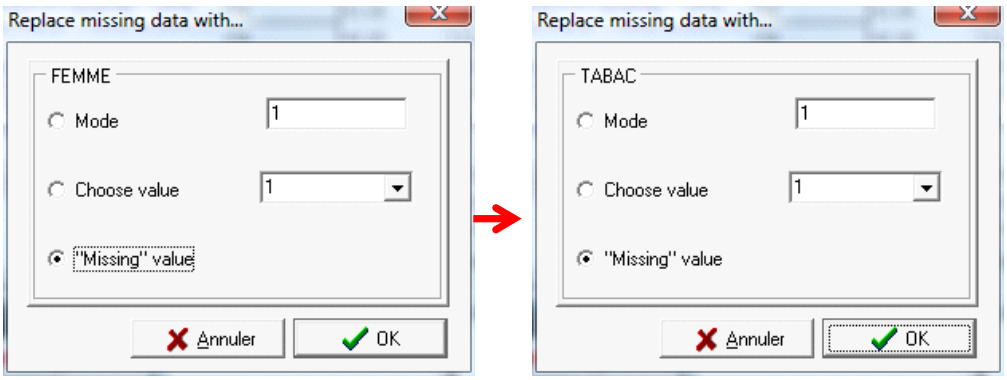
We stop the current analysis (ANALYSIS / STOP ANALYSIS) and we clear the grid (FILE / NEW). Then, we load again the RONFLEMENT_WITH_MISSING.FDM data file (FILE / OPEN) (Figure 2). We click on the STATISTICS / MISSING DATA / REPLACE BY VALUES menu. In the dialog box, we select all the variables.



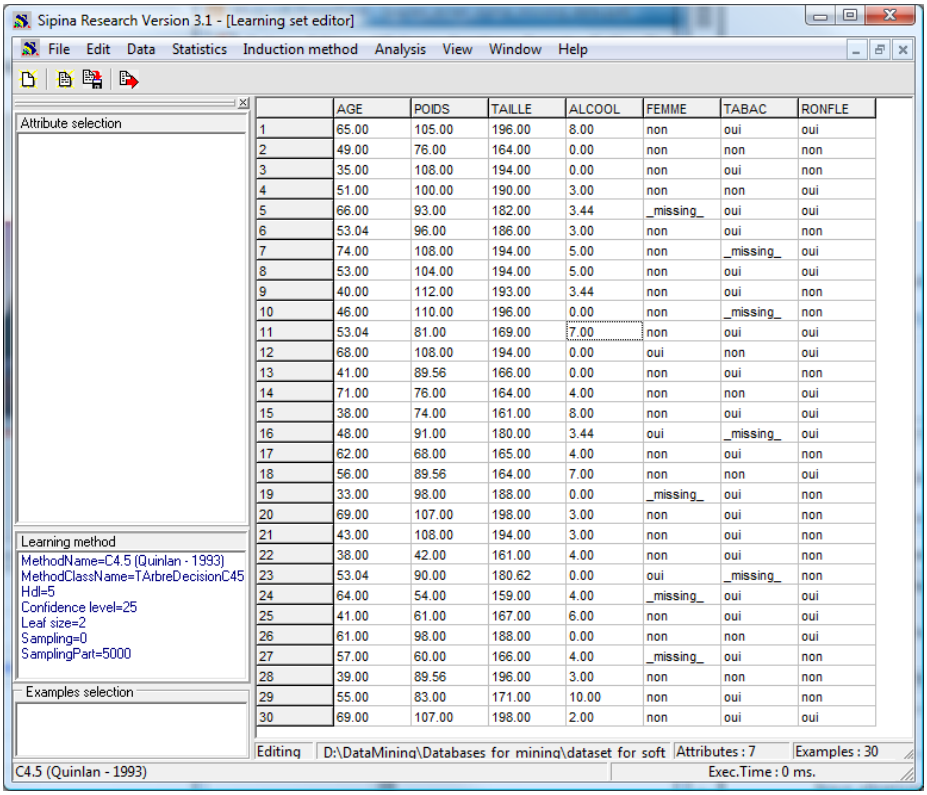
Again, we use the mean for AGE, POIDS, TAILLE and ALCOOL.



For the discrete attributes, we select now the « MISSING » VALUE option.

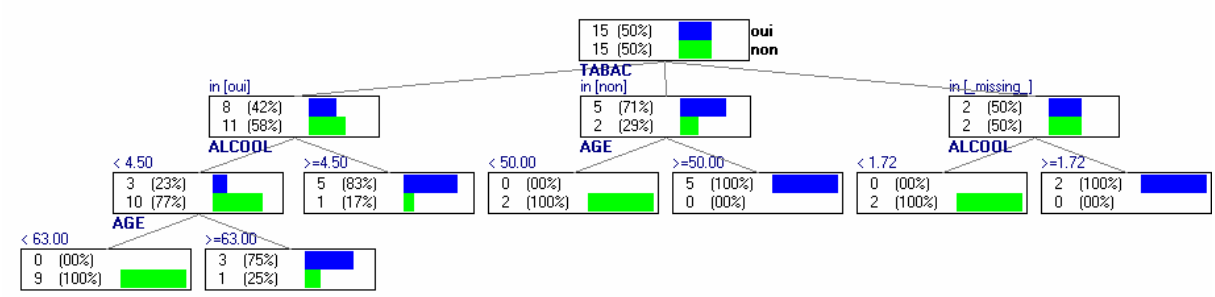


In the data visualization grid, the « _MISSING_ » value code is visible. Of course, the number of rows is not modified.



	AGE	POIDS	TAILLE	ALCOOL	FEMME	TABAC	RONFLE
1	65.00	105.00	196.00	8.00	non	oui	oui
2	49.00	76.00	164.00	0.00	non	non	non
3	35.00	108.00	194.00	0.00	non	oui	non
4	51.00	100.00	190.00	3.00	non	non	oui
5	66.00	93.00	182.00	3.44	_missing_	oui	oui
6	53.04	96.00	186.00	3.00	non	oui	non
7	74.00	108.00	194.00	5.00	non	_missing_	oui
8	53.00	104.00	194.00	5.00	non	oui	oui
9	40.00	112.00	193.00	3.44	non	oui	non
10	46.00	110.00	196.00	0.00	non	_missing_	non
11	53.04	81.00	169.00	7.00	non	oui	oui
12	68.00	108.00	194.00	0.00	oui	non	oui
13	41.00	89.56	166.00	0.00	non	oui	non
14	71.00	76.00	164.00	4.00	non	non	oui
15	38.00	74.00	161.00	8.00	non	oui	oui
16	48.00	91.00	180.00	3.44	oui	_missing_	oui
17	62.00	68.00	165.00	4.00	non	oui	non
18	56.00	89.56	164.00	7.00	non	non	oui
19	33.00	98.00	188.00	0.00	_missing_	oui	non
20	69.00	107.00	198.00	3.00	non	oui	non
21	43.00	108.00	194.00	3.00	non	oui	non
22	38.00	42.00	161.00	4.00	non	oui	non
23	53.04	90.00	180.62	0.00	oui	_missing_	non
24	64.00	54.00	159.00	4.00	_missing_	oui	oui
25	41.00	61.00	167.00	6.00	non	oui	oui
26	61.00	98.00	188.00	0.00	non	non	oui
27	57.00	60.00	166.00	4.00	_missing_	oui	non
28	39.00	89.56	196.00	3.00	non	non	non
29	55.00	83.00	171.00	10.00	non	oui	non
30	69.00	107.00	198.00	2.00	non	oui	oui

We launch again the analysis (Specifying the target and the input attributes; launching the learning process). We obtain the following decision tree.



The tree is roughly similar to one built on the complete data (Figure 1). We note that a new branch was formed on the right, composed from the value _MISSING_ of TABAC. C4.5 extracts a perfect segmentation, with pure leaves. This is an artifact. Missing values have been introduced completely at random in our dataset. This new rule is certainly not reproducible on another sample.

5 Conclusion

In this tutorial, we show various approaches to handling missing values which are supplied by Sipina. Let us not draw any particular conclusions from this small illustrative example. It intended solely to show that there are not perfect approaches in the missing value problem.

We preferred the qualitative analysis of the results here; we compare the trees produced after the treatment of the missing values. A more systematic approach is possible. It is often found in publications. The idea is to analyze the impact of the missing value treatment on the generalization accuracy of the predictive model.