1 Topic

Interactive decision tree with SPAD.

About the data mining tools, commercial and academic tools are not (often) intended to the same users. The academic tools are especially intended to the researchers or the students. The goal is to raise to the users a large number of algorithms which enable them to analyze their behavior on various dataset. The users must be able to implement experiments easily. The R software (http://www.r-project.org/) is perhaps the most representative of this kind of tools. In contrast, the commercial tools are especially intended to final users, including the researchers in other domains. They need to implement the complete cycle of the data mining process, starting from the data access into databases, to the deployment and the production of reports. The tool must make easier their work by taking charge the repetitive and tedious tasks (e.g. data cleansing, graphical representation, creating reports...). Of course, the distinction is not so clear. Many academic tools incorporate features that would interest the industrial users (e.g. RapidMiner which is a significant extension of Yale, an academic tool). At the opposite, many commercial tools try to incorporate the algorithms libraries – which are very considerable – of academic tools.

In this tutorial, we will be interested in SPAD (http://www.spad.eu/). This is a French software specialized in exploratory data analysis which evolved much these last years. We would perform a sequence of analysis from a dataset collected into 3 worksheets of a Excel data file: (1) we create a classification tree from the learning sample into the first worksheet, we try to analyze deeply some nodes of the tree to highlight the characteristics of covered instances, we try also to modify interactively (manually) the properties of some splitting operation; (2) we apply the classifier on unseen cases of the second worksheet; (3) we compare the prediction of the model with the actual values of the target attribute contained into the third worksheet.

Of course, we can perform this process using free tools such as SIPINA (the interactive construction of the tree - http://eric.univ-lyon2.fr/~ricco/sipina.html) or R (the programming of the sequence of operations, in particular the applying of the model on unlabeled dataset). But with Spad or other commercial tools (e.g. SPSS Modeler, SAS Enterprise Miner, STATISTICA Data Miner…), we can very easily specify the whole sequence, even if we are not especially familiarized with data mining tools.

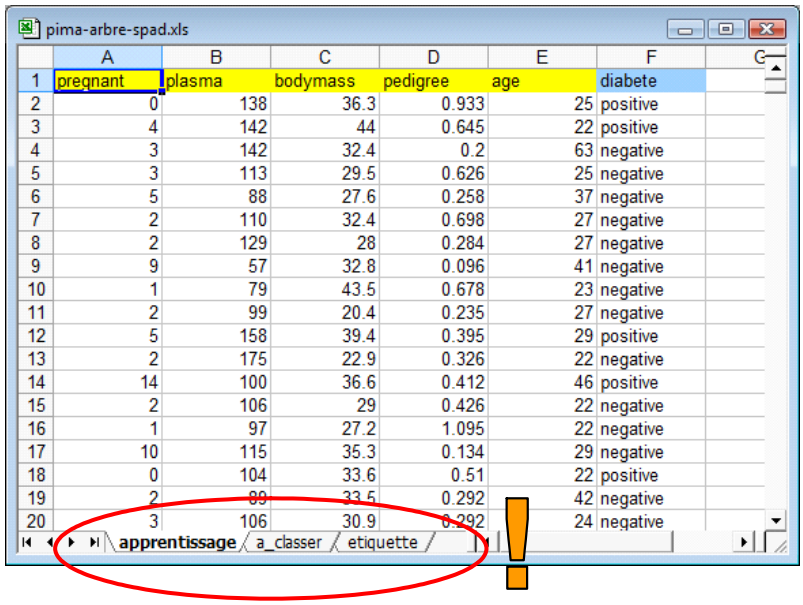
2 Dataset

Our data file "PIMA-ARBRE-SPAD.XLS[1]" is a version of the "Pima Indian Diabetes" dataset available from the UCI Server[2]. The aim is to detect the individuals (females) tested positive for diabetes from their characteristics (number of times pregnant; body mass index; age; etc.).

We show below the Excel workbook with 3 worksheets: (1) "**apprentissage**" is the labeled training sample, we use this dataset for the construction of the decision tree; (2) "**à classer**" is the unlabeled dataset on which we apply the model learned before; (3) "**étiquette**" contains the labels of the

[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/pima-arbre-spad.zip
[2] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

instances from the second worksheet, this column is not usually available in a data mining process, we use this information in a pedagogical goal only in this tutorial.



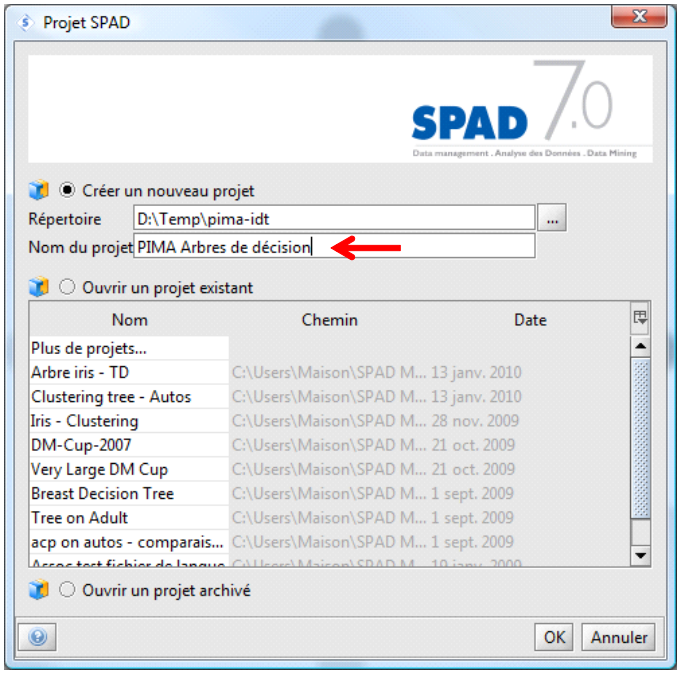
	A	B	C	D	E	F
1	pregnant	plasma	bodymass	pedigree	age	diabete
2	0	138	36.3	0.933	25	positive
3	4	142	44	0.645	22	positive
4	3	142	32.4	0.2	63	negative
5	3	113	29.5	0.626	25	negative
6	5	88	27.6	0.258	37	negative
7	2	110	32.4	0.698	27	negative
8	2	129	28	0.284	27	negative
9	9	57	32.8	0.096	41	negative
10	1	79	43.5	0.678	23	negative
11	2	99	20.4	0.235	27	negative
12	5	158	39.4	0.395	29	positive
13	2	175	22.9	0.326	22	negative
14	14	100	36.6	0.412	46	positive
15	2	106	29	0.426	22	negative
16	1	97	27.2	1.095	22	negative
17	10	115	35.3	0.134	29	negative
18	0	104	33.6	0.51	22	positive
19	2	89	33.5	0.292	42	negative
20	3	106	30.9	0.292	24	negative

3 Analysis under SPAD

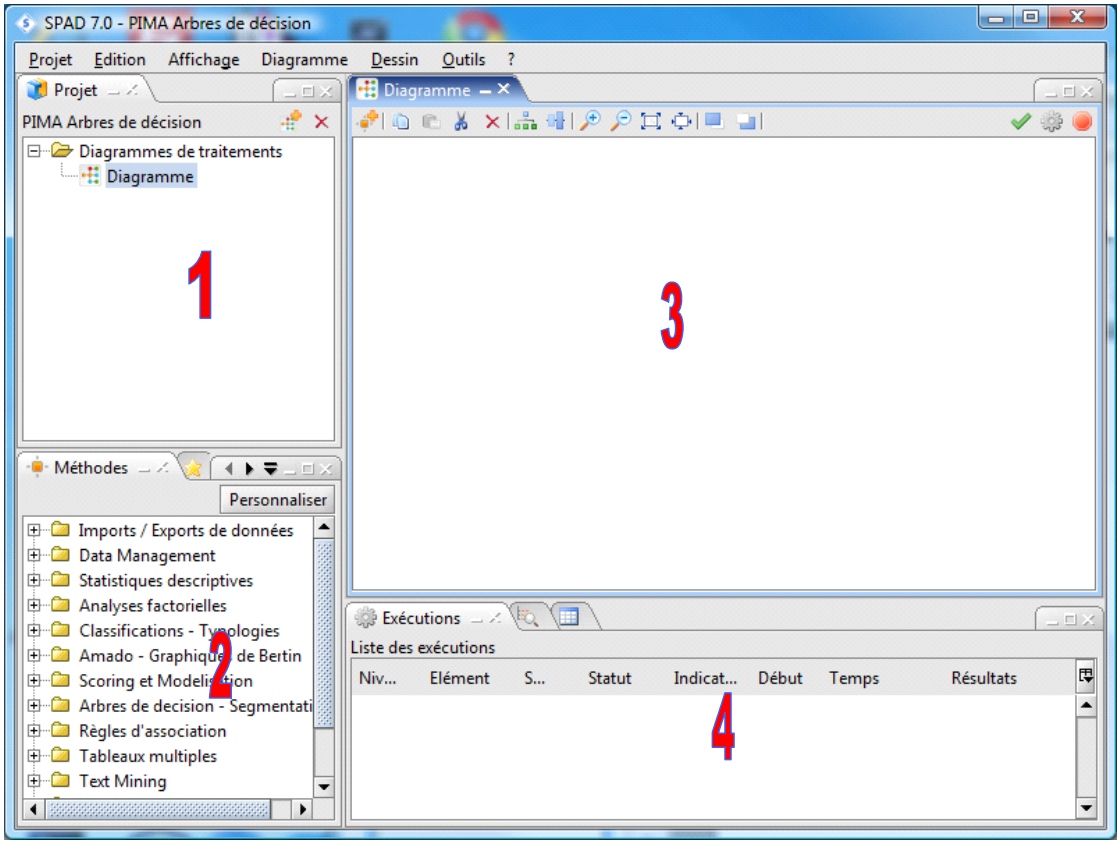
3.1 Learning and storing a classifier

3.1.1 Creating a project

We start SPAD and we create a new project “PIMA Arbres de decision”.

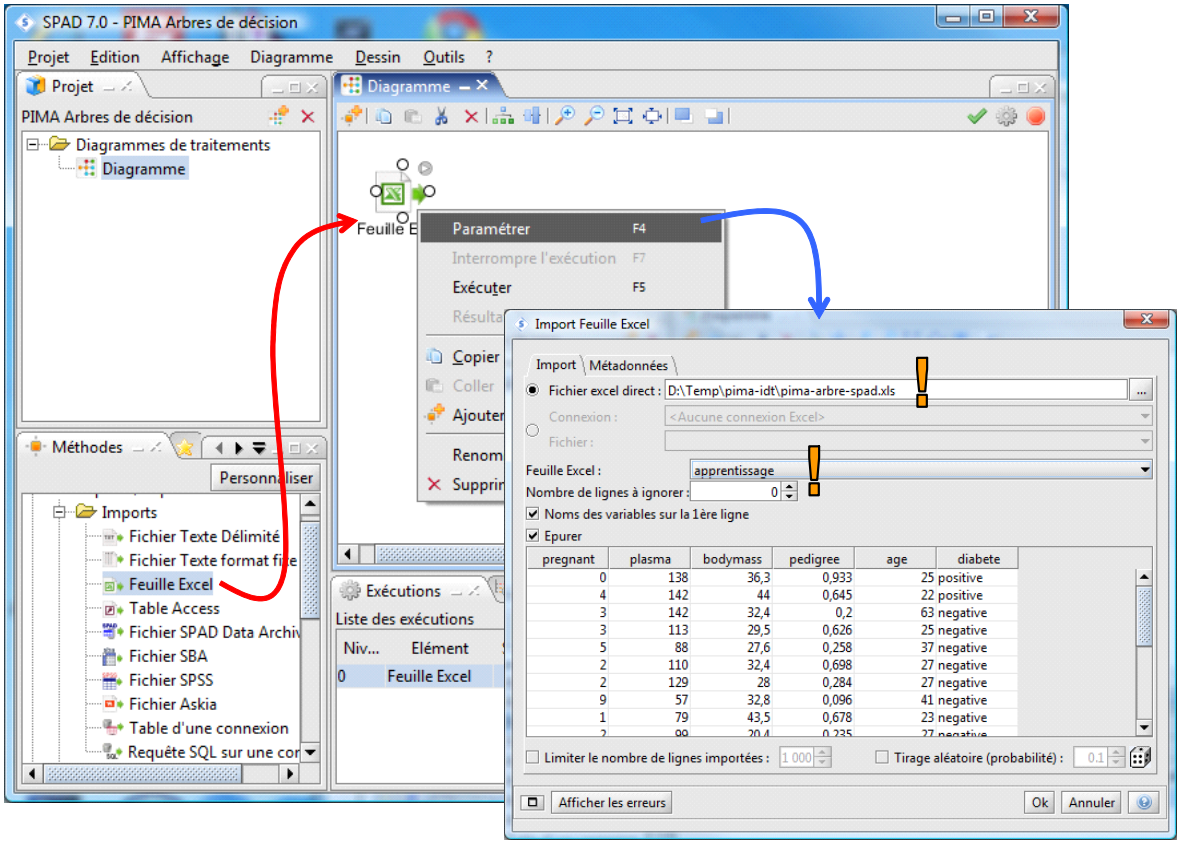


A visual programming interface enables us to draw diagrams of data operations. The interface is complying with the standard interface used in many data mining tools. We observe in the main window of SPAD: (1) Project manager; (2)The palettes contain the nodes for data mining operations; (3) The workspace enables to define the sequence of treatments; (4)“Exécution” enables to follow the sequences of operations.



3.1.2 Importing the data file

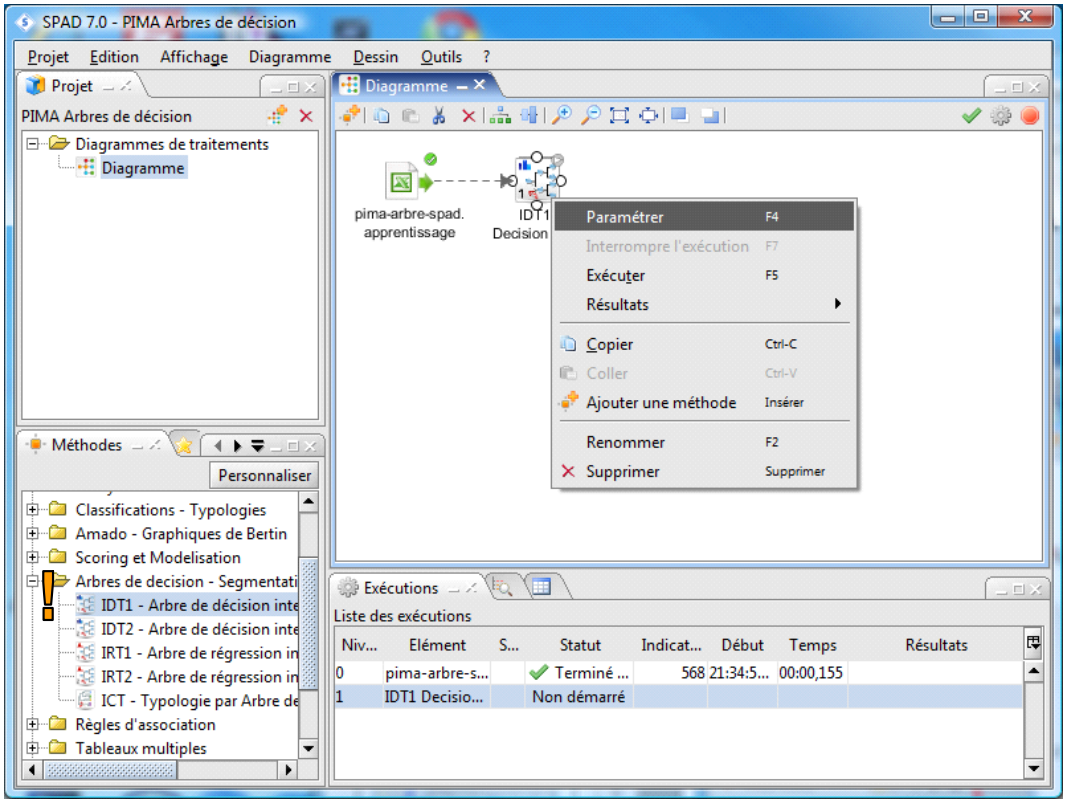
First, we want to import the learning sample using the **Feuille Excel** tool. We click on the "Paramétrer (F4)" contextual menu to define the settings. We set the names of the workbook and the worksheet.



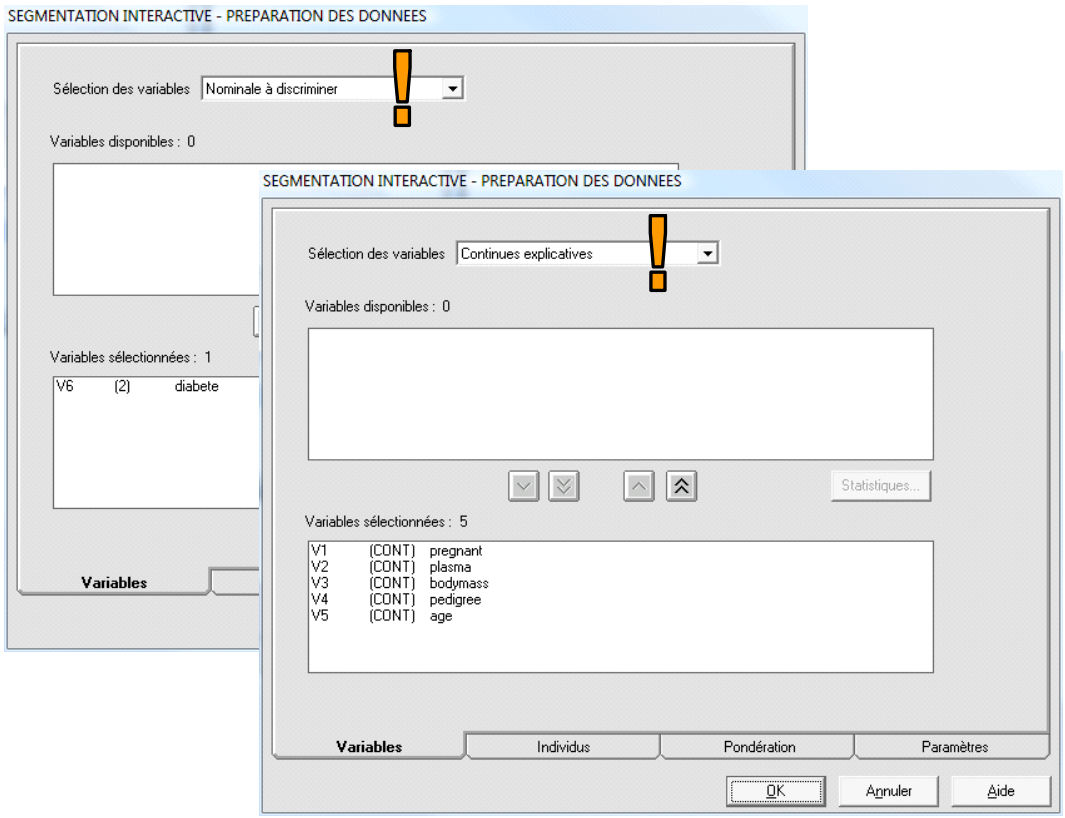
We click on the OK button. The dataset is loaded.

3.1.3 Specifying the variables for the analysis

IDT1 node enables to define the target variable (diabète) and the explanatory variables (the others).

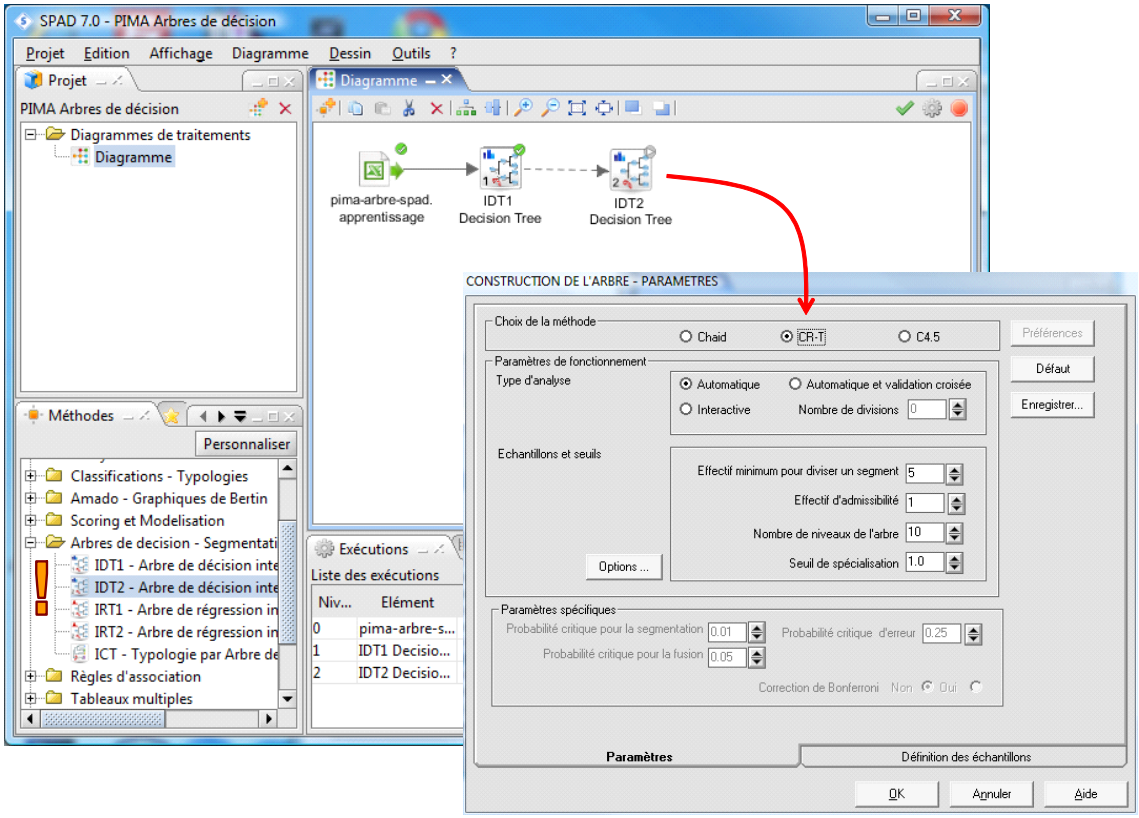


We set the settings as follows.

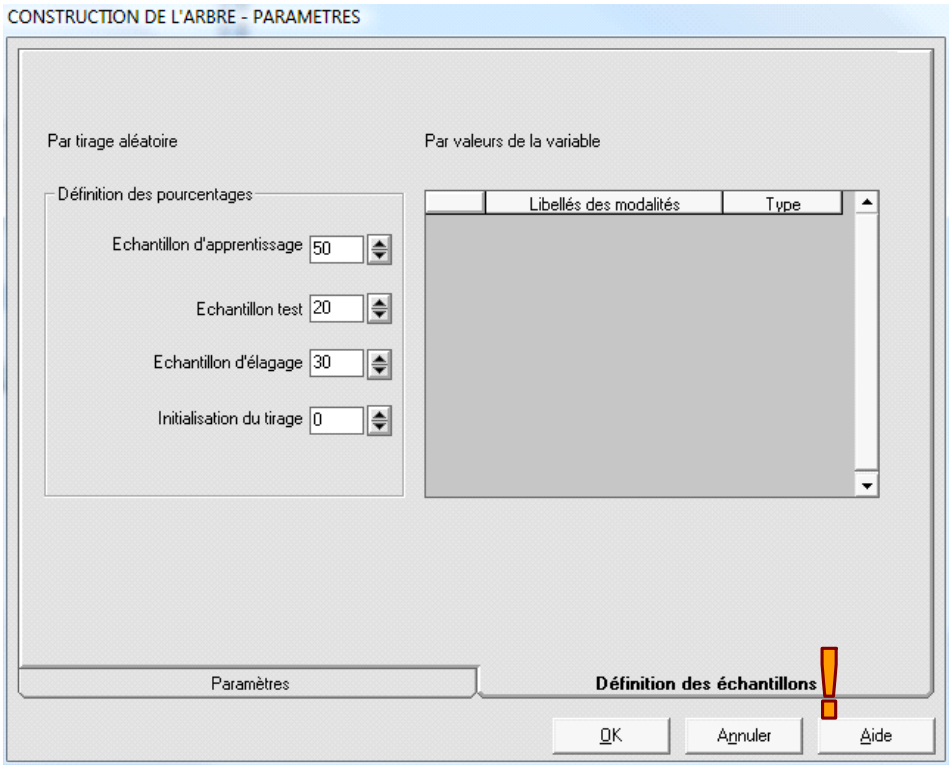


3.1.4 Choosing the learning algorithm

IDT2 enables to define the learning algorithm and its settings. We use for instance the C-RT approach.



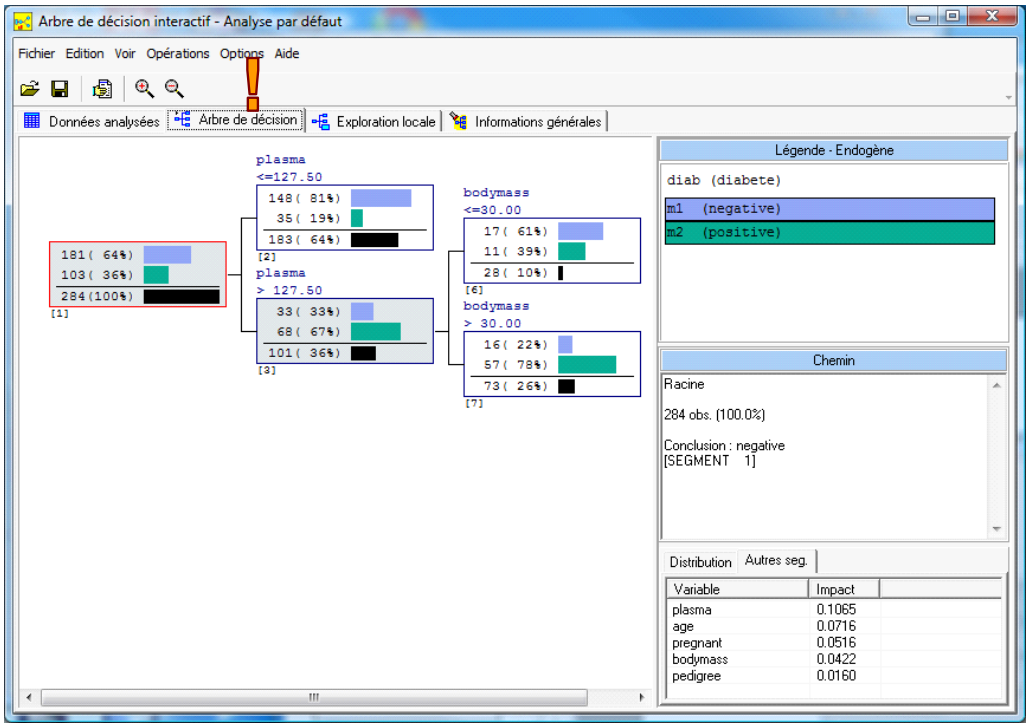
The “Définition des échantillons” enables to specify the used samples.



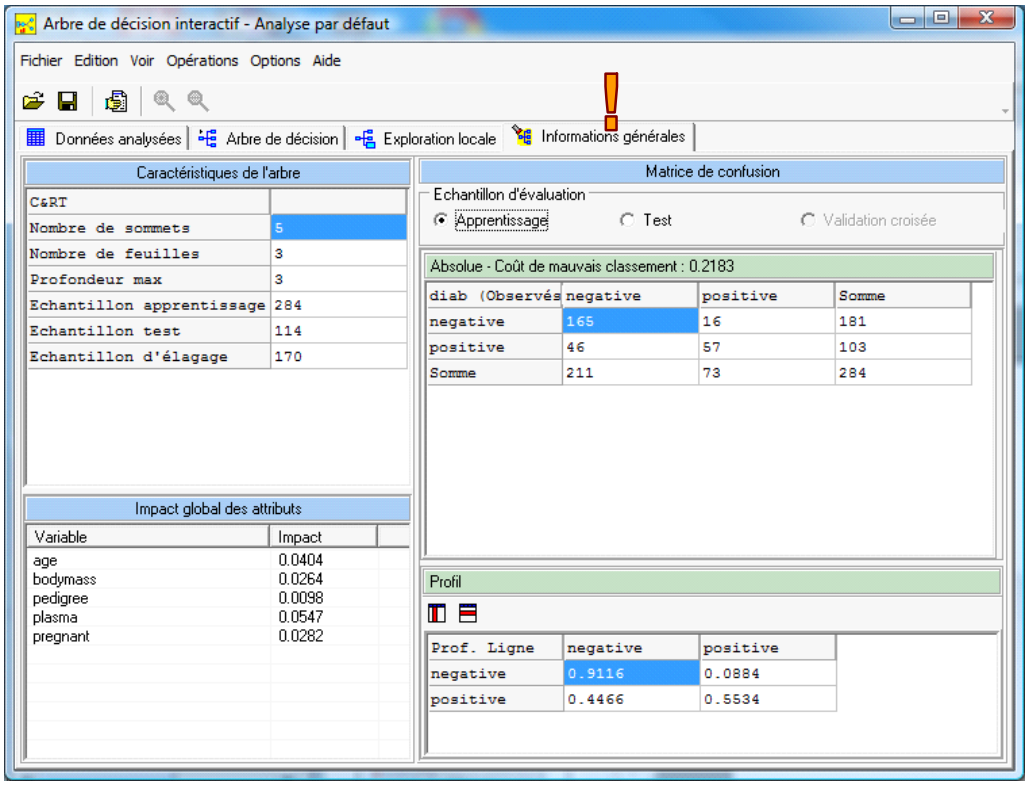
"Echantillon d'apprentissage" corresponds to the growing sample (50%); "Echantillon d'élagage" is the pruning sample (30%); "Echantillon test" is the testing sample (20%).

3.1.5 Obtaining the decision tree

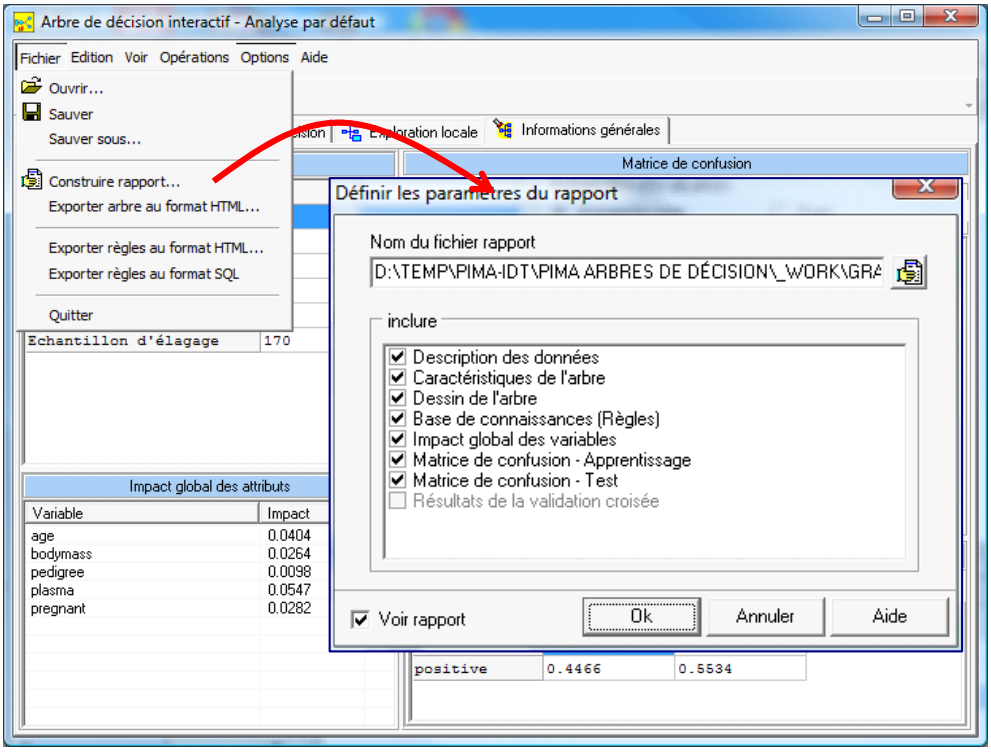
When we validate the settings, the calculations are automatically launched. We click on the RESULTATS / INTERACTIVE DECISION TREE menu to visualize the tree.



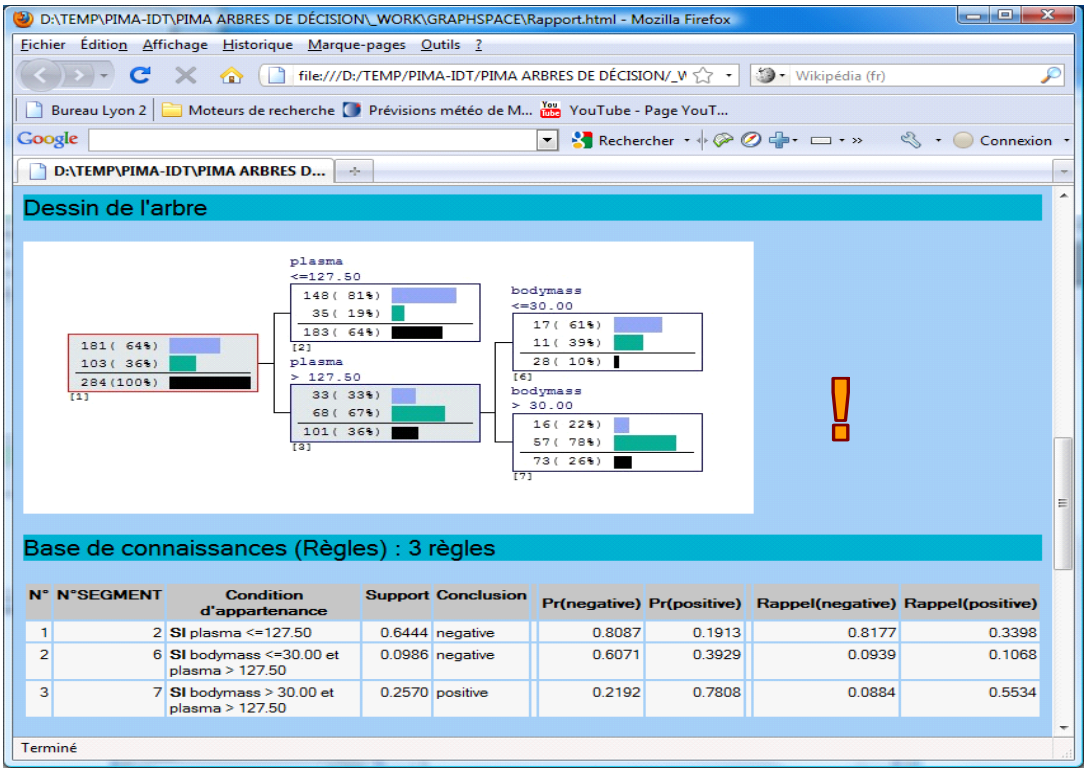
Into the « Informations générales » tab, we observe the confusion matrix computed on the learning "Apprentissage" and testing "Test" samples.



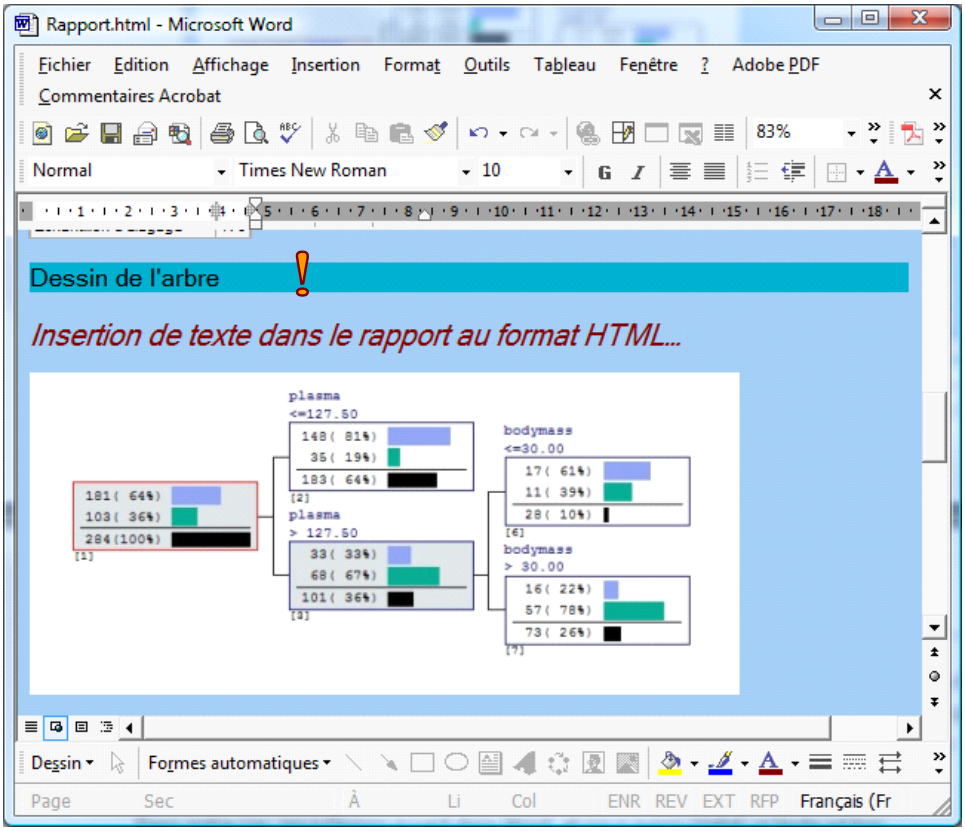
Creating a report about the analysis. We can create a report which summarizes the main results. We click on the FICHIER / CONSTRUIRE RAPPORT. We select the sections incorporated into the report.



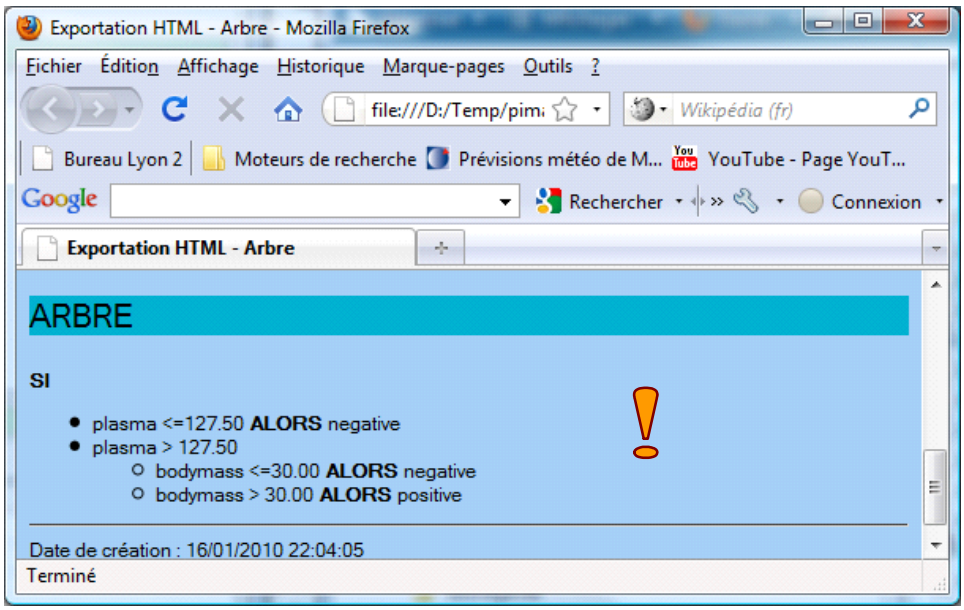
The report is in the HTML format. It is automatically loaded into the default browser of the system.



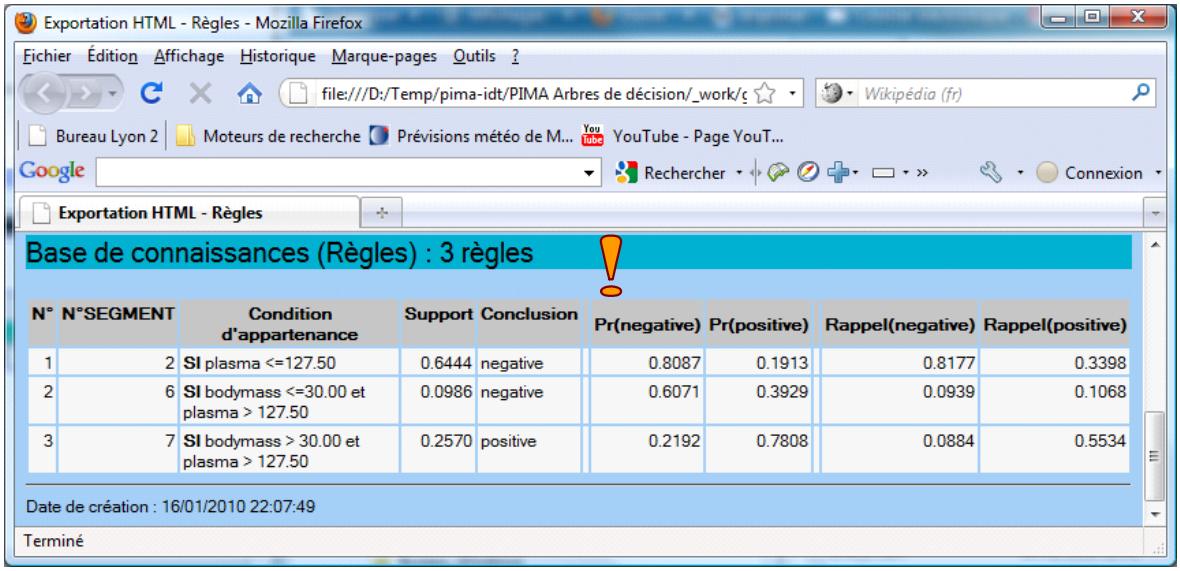
We see: the description of the dataset, a graphical representation of the tree, the rules extracted from the tree. Because the report is in the HTML format, we can edit it with any word processor.



Exporting the tree in the HTML format. We can also obtain a description of the tree in the HTML format. It is often more compact than the graphical description. We click on the FICHIER / EXPORTER ARBRE AU FORMAT HTML menu for that.



Exporting the rules in the HTML format. The tree can be transformed into a ruleset. We click on the FICHIER / EXPORTER REGLES AU FORMAT HTML menu. We obtain the following report.

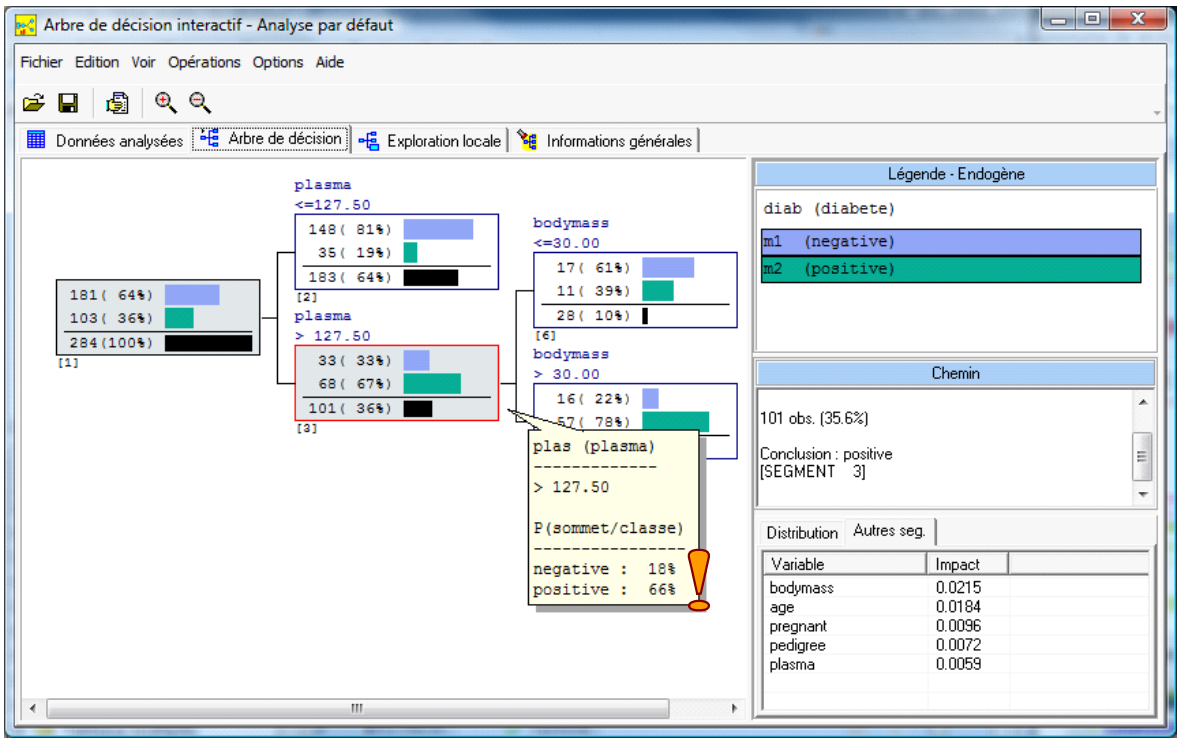


N°	N°SEGMENT	Condition d'appartenance	Support	Conclusion	Pr(negative)	Pr(positive)	Rappel(negative)	Rappel(positive)
1	2	SI plasma <=127.50	0.6444	negative	0.8087	0.1913	0.8177	0.3398
2	6	SI bodymass <=30.00 et plasma > 127.50	0.0986	negative	0.6071	0.3929	0.0939	0.1068
3	7	SI bodymass > 30.00 et plasma > 127.50	0.2570	positive	0.2192	0.7808	0.0884	0.5534

3.1.6 Exploring a node

SPAD incorporates the usual features of this kind of tools. Into the "Arbre de décision" tab, we can explore deeply each node of the tree.

For the node n°3 for instance, we observe 101 cases (36% of 284), 33 among them are negatives (diabetes = negative, 33% = 33 / 101) and 68 are positives. By clicking on the node, we observe the description of the rule associated to the node and some conditional probabilities [e.g. P(Node / Negative) = 18% = 33 / 181].

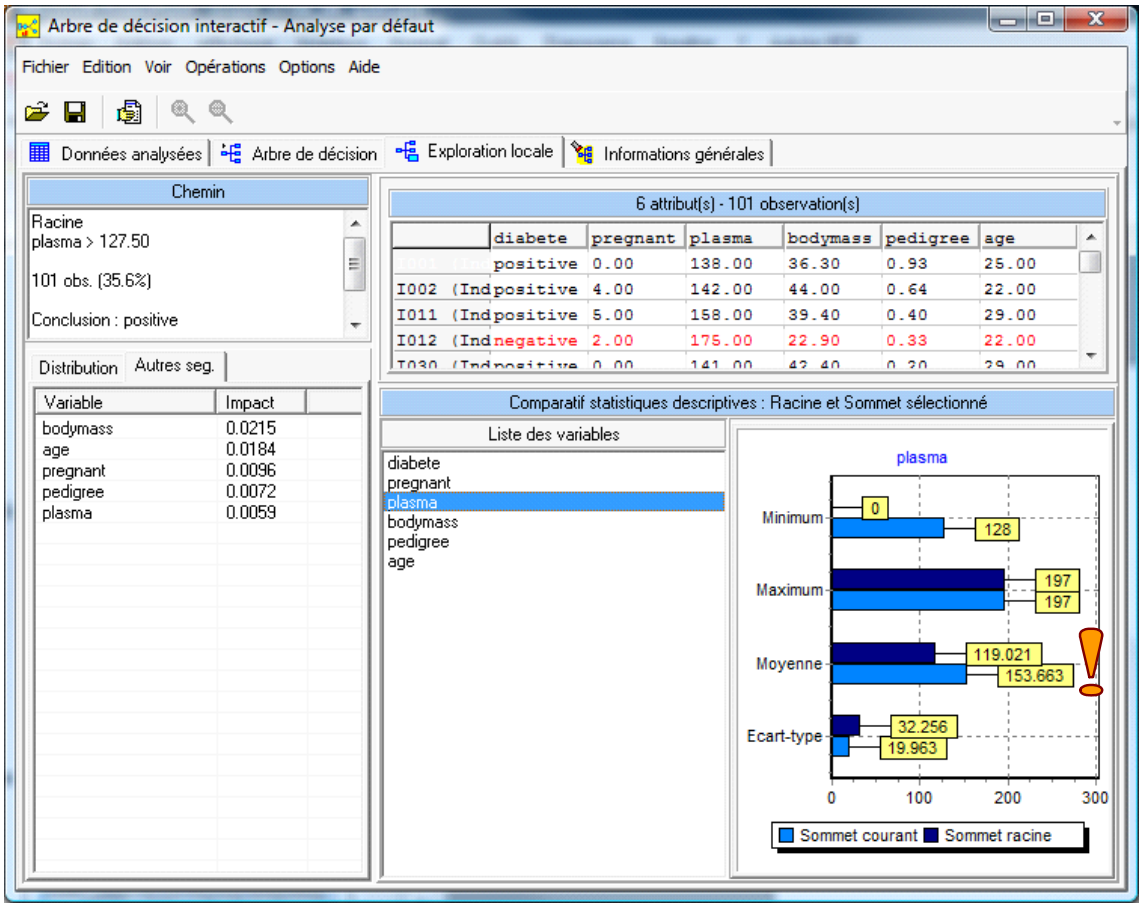


In the lower right par of the window, we see the goodness of split of each explanatory variable (Gini index for the C-RT approach). The most relevant variable if we want to split the node is BODYMASS with a gain = 0.0215. This is precisely the splitting operation used to obtain the leaves n°6 and n°7.

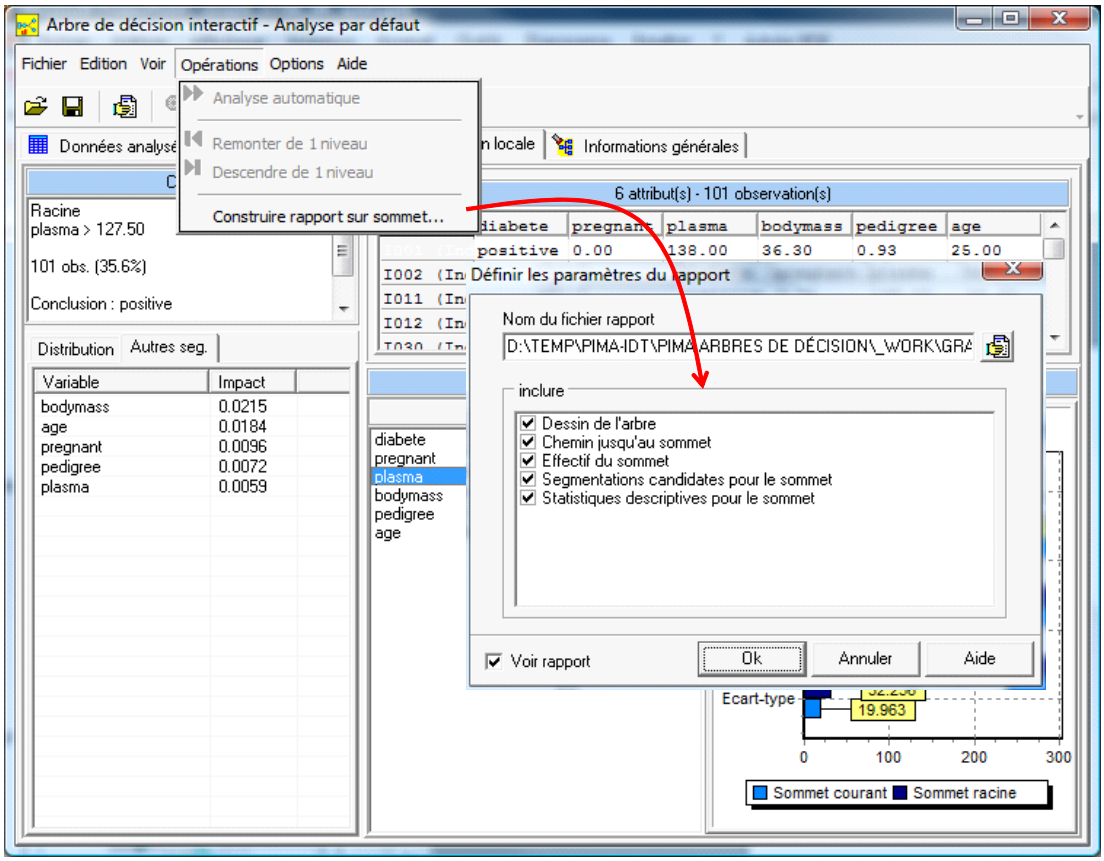
To deeply explore the node n°3, we activate the “Exploration locale” tab:

- We can visualize the instances covered by the node. Because the majority of the instances belong to the positive class value, the negative instances are highlighted.
- In the low part of the window, we observe the descriptive statistics comparing the root of the tree (the whole population) and the selected node (the sub population corresponding to the node). If we select PLASMA, we observe that its mean is 119.021 for the whole instances. For the individuals covered by the selected node, the mean becomes 153.663. The values of PLASMA seem higher for this subpopulation.

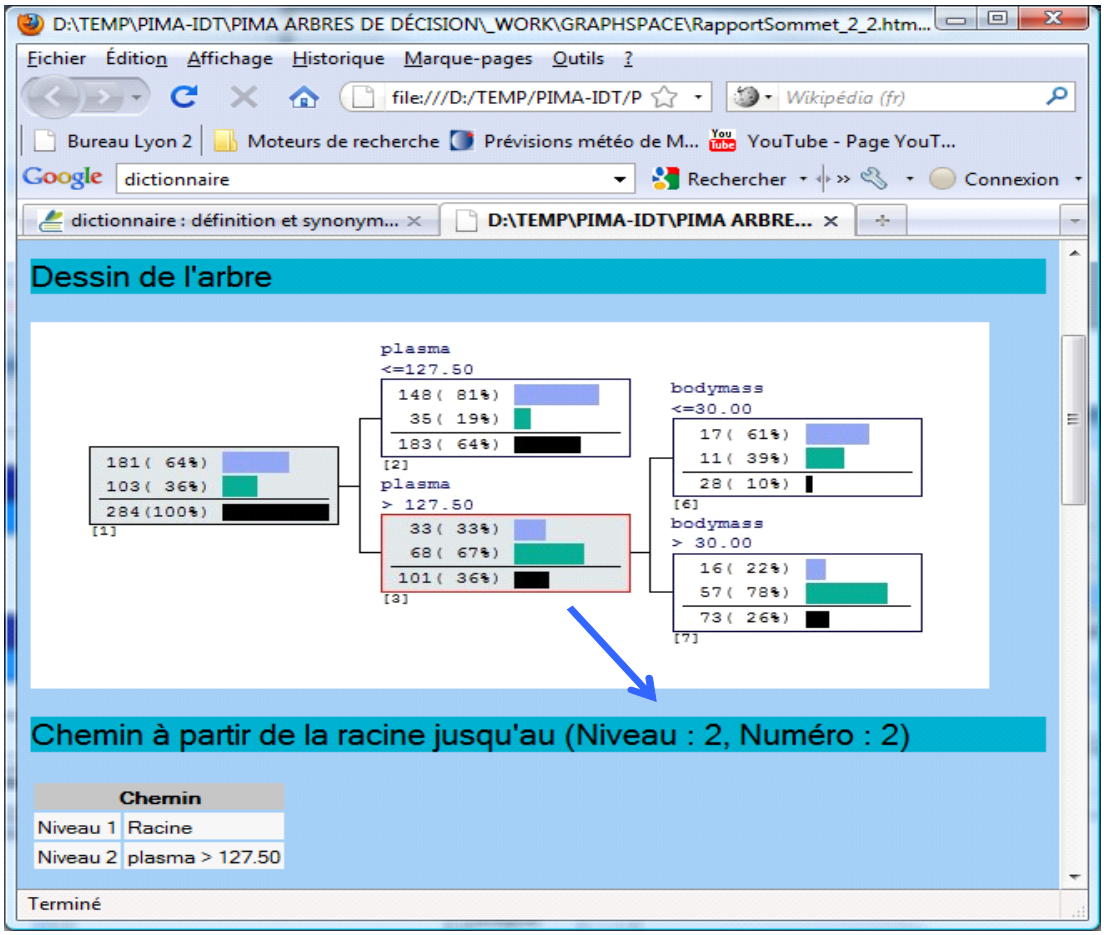
We can perform this kind of analysis for each variables of the dataset.



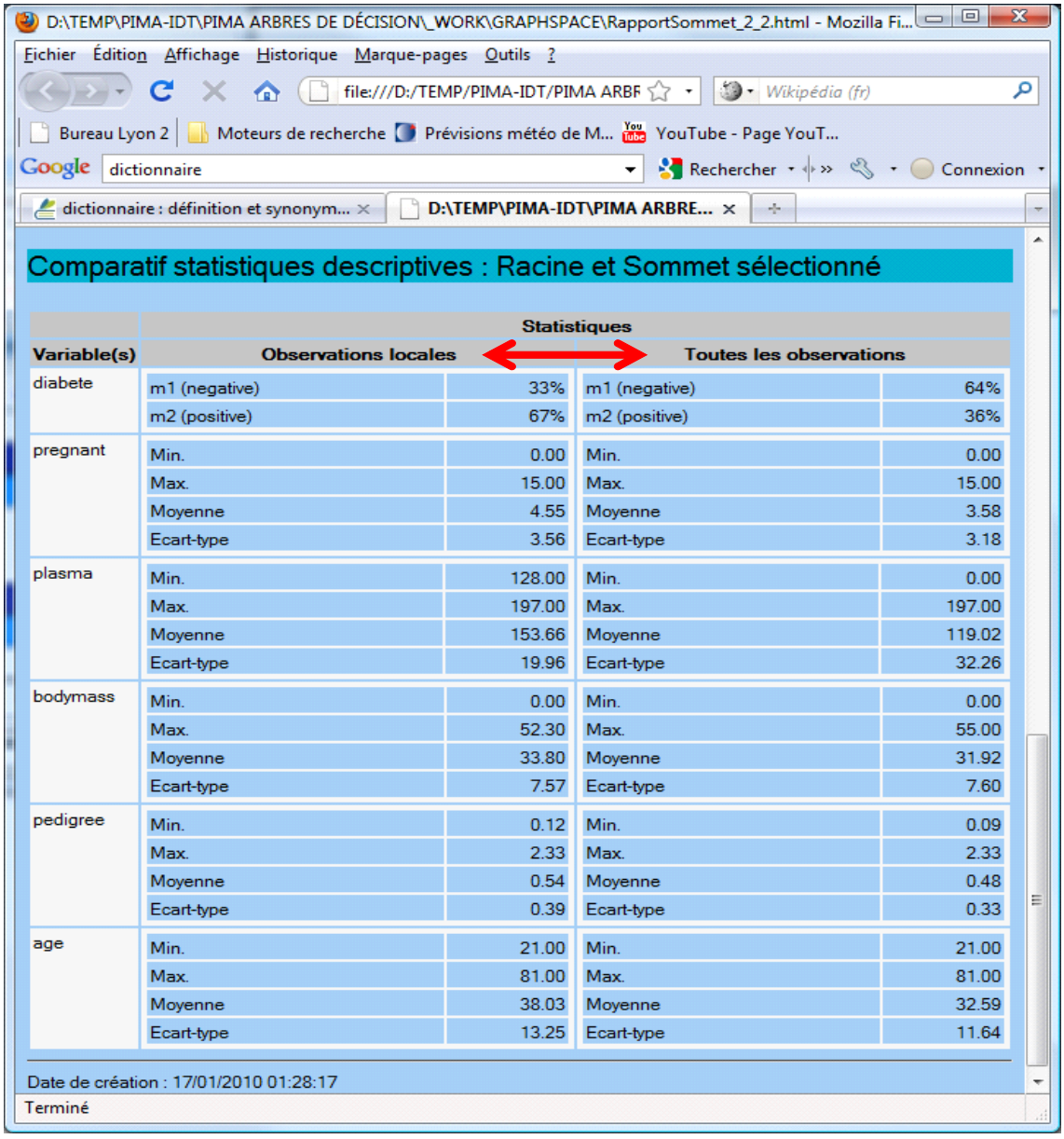
Creating a report about a node. We can also create a report for the selected node. We click on the OPERATIONS / CONSTRUIRE RAPPORT SUR SOMMET menu for that.



The report is automatically visualized into the default browser of your system.



The selected node is highlighted in red in the graphical representation. Some comparative descriptive statistics, computed on each variable, enable to characterize the subpopulation covered by the node.



D:\TEMP\PIMA-IDT\PIMA ARBRES DE DÉCISION_WORK\GRAPHSPACE\RapportSommet_2_2.html - Mozilla Fi...

Fichier Édition Affichage Historique Marque-pages Outils ?

file:///D:/TEMP/PIMA-IDT/PIMA ARBF

Wikipédia (fr)

Bureau Lyon 2 | Moteurs de recherche | Prévisions météo de M... | YouTube - Page YouT...

Google dictionnaire

Rechercher | Connexion

dictionnaire : définition et synonym... | D:\TEMP\PIMA-IDT\PIMA ARBRE...

Comparatif statistiques descriptives : Racine et Sommet sélectionné

Variable(s)	Statistiques – Observations locales		Statistiques – Toutes les observations	
diabete	m1 (negative)	33%	m1 (negative)	64%
	m2 (positive)	67%	m2 (positive)	36%
pregnant	Min.	0.00	Min.	0.00
	Max.	15.00	Max.	15.00
	Moyenne	4.55	Moyenne	3.58
	Ecart-type	3.56	Ecart-type	3.18
plasma	Min.	128.00	Min.	0.00
	Max.	197.00	Max.	197.00
	Moyenne	153.66	Moyenne	119.02
	Ecart-type	19.96	Ecart-type	32.26
bodymass	Min.	0.00	Min.	0.00
	Max.	52.30	Max.	55.00
	Moyenne	33.80	Moyenne	31.92
	Ecart-type	7.57	Ecart-type	7.60
pedigree	Min.	0.12	Min.	0.09
	Max.	2.33	Max.	2.33
	Moyenne	0.54	Moyenne	0.48
	Ecart-type	0.39	Ecart-type	0.33
age	Min.	21.00	Min.	21.00
	Max.	81.00	Max.	81.00
	Moyenne	38.03	Moyenne	32.59
	Ecart-type	13.25	Ecart-type	11.64

Date de création : 17/01/2010 01:28:17

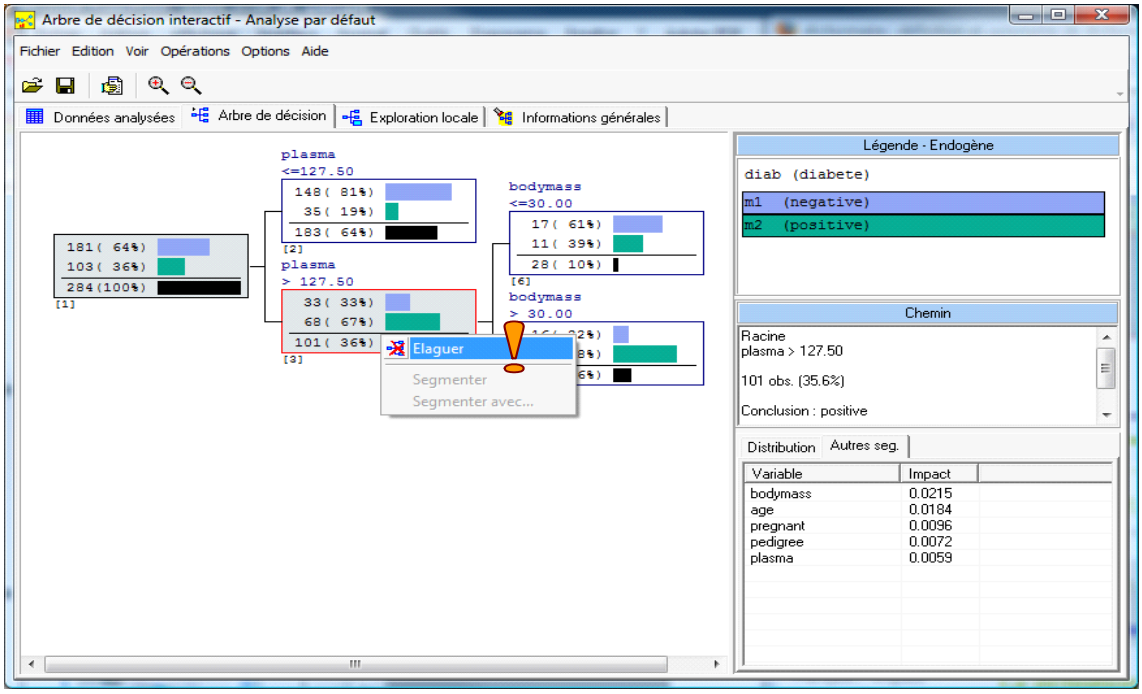
Terminé

We obtain the values given in the "Local exploration" tab. For instance, the mean for PLASMA is 119.02 into the whole population (the root node of the tree), it becomes 153.66 into the subpopulation associated to the current node.

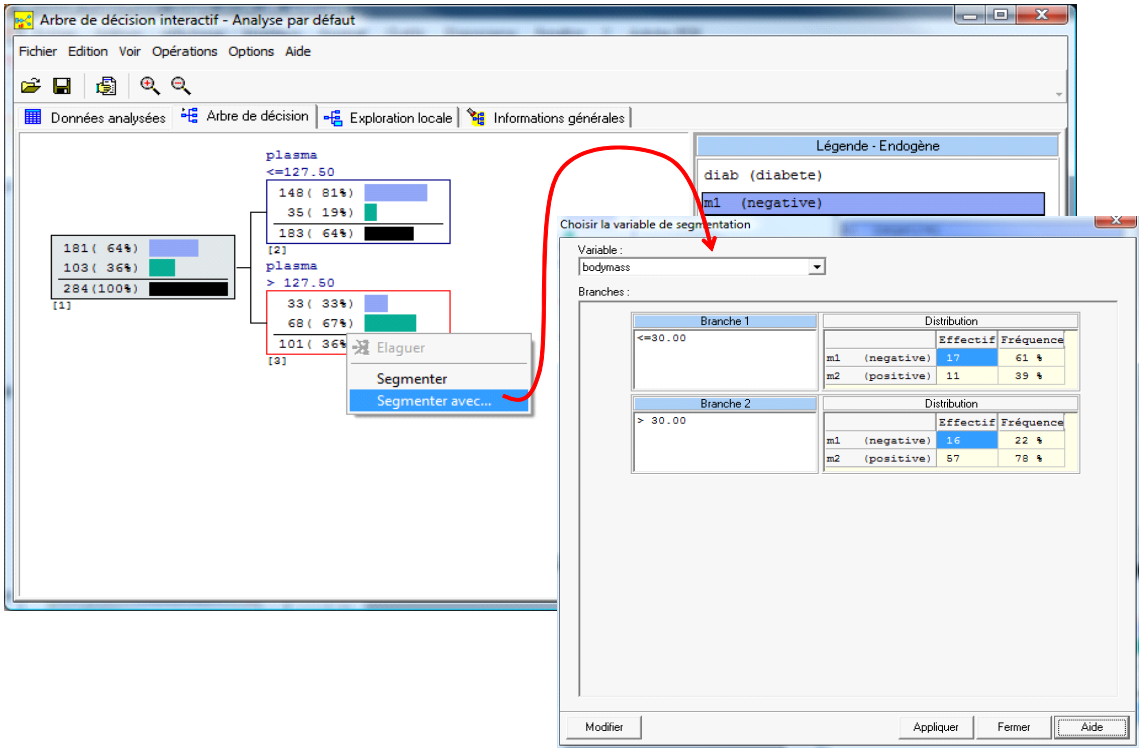
3.1.7 Interactive construction of the tree

Pruning manually a subtree – Selecting the variable for splitting. We come back to the "Arbre de décision" tab. We select the node n°3. We observe that BODYMASS is used for the partitioning. The Gini gain (impact) is 0.0215. But we note that AGE can be also a good splitting attribute. We want to replace BODYMASS by AGE. The operation is performed in few steps.

(1) We remove the subtree started from the current node by clicking on the "Elaguer" menu.

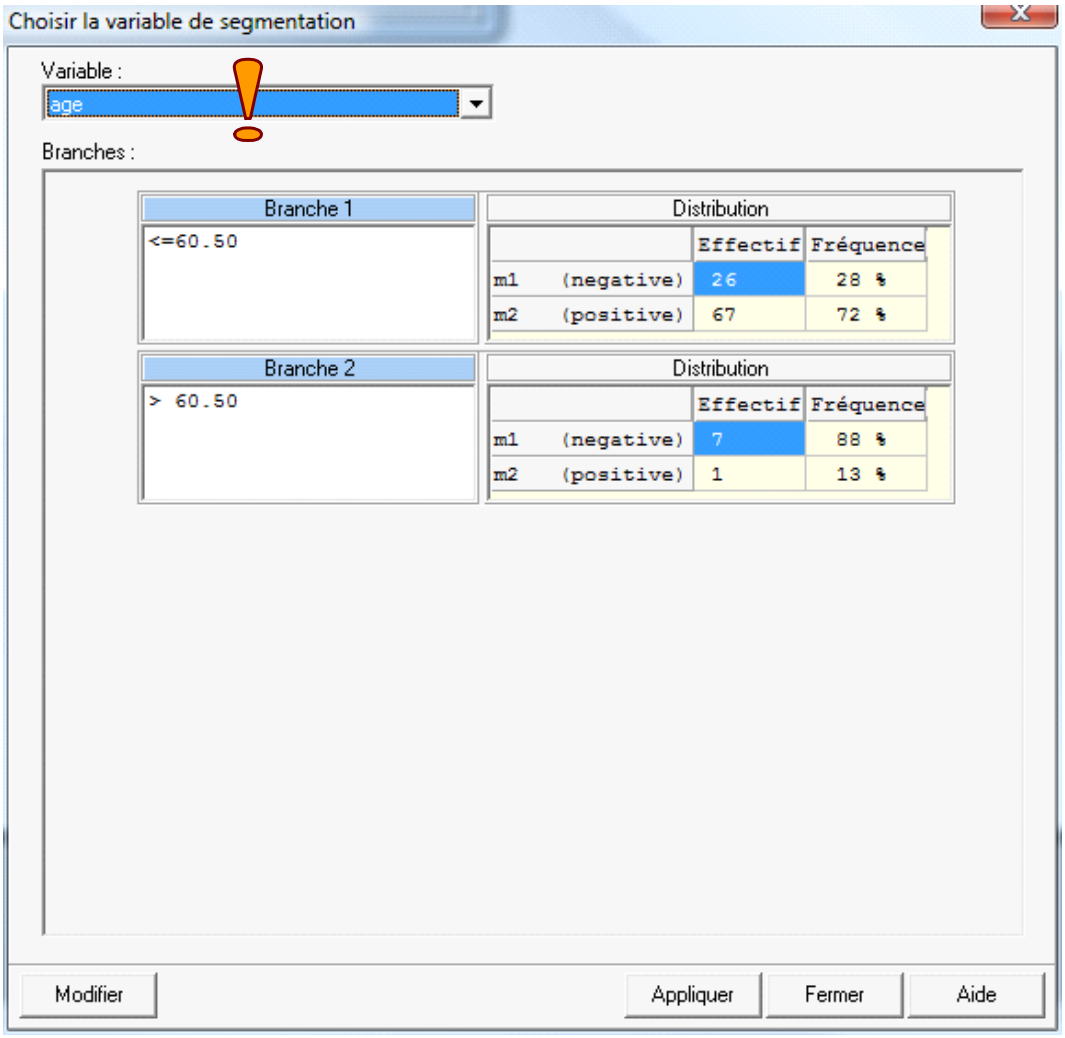


(2) Then we click on the « Segmenter avec… » menu. Into the dialog settings, we see the list of explanatory variables, ordered according to the Gini gain.

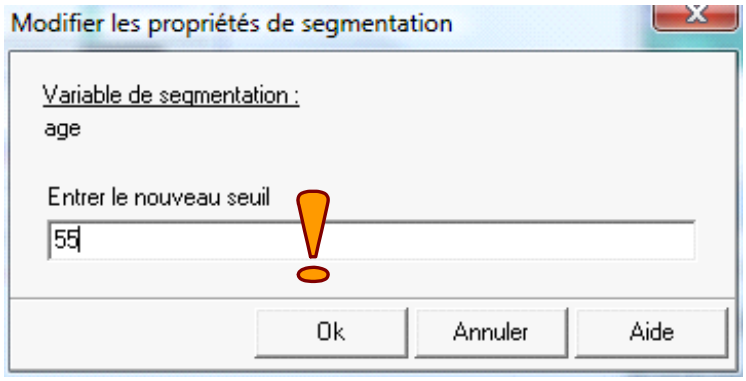


BODYMASS is the first because it has the best Gini gain. But we can select any explanatory variable.

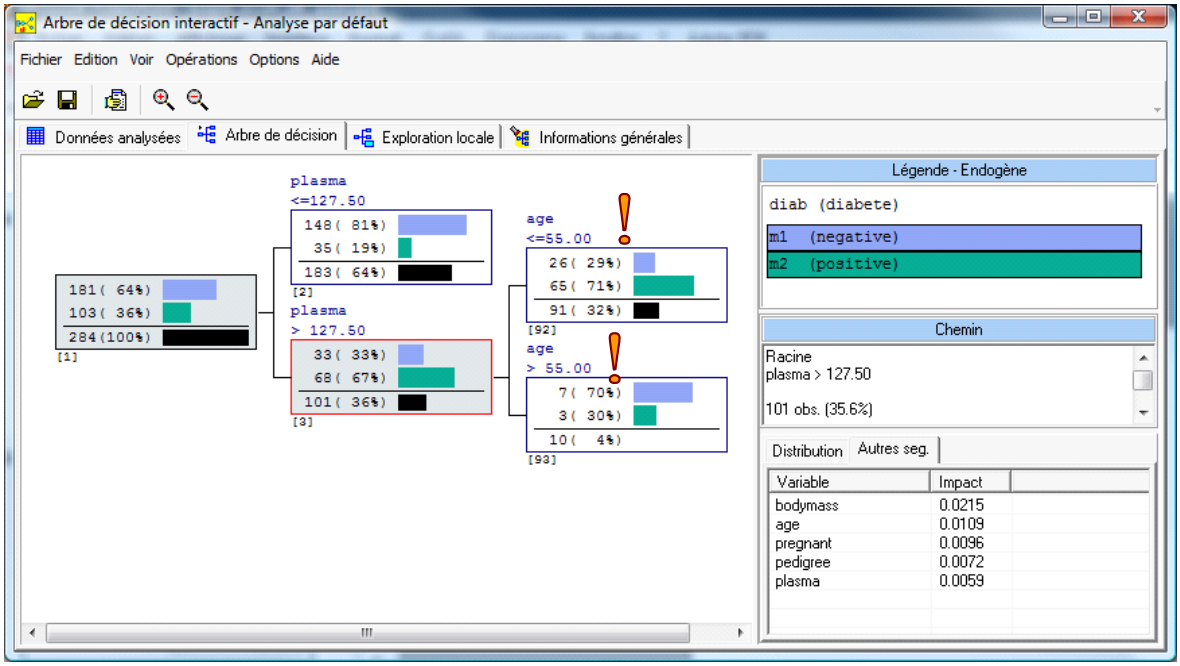
In our case, we select AGE. The leaves that we could obtain by the splitting operation are displayed into the dialog box.



Modifying the splitting characteristics. We can modify the characteristics of the split. For instance, if we want to transform the cut value (e.g. from 60.5 to 55), we click on the **MODIFIER** button in the lower part of the dialog box. We can set the new cut value.

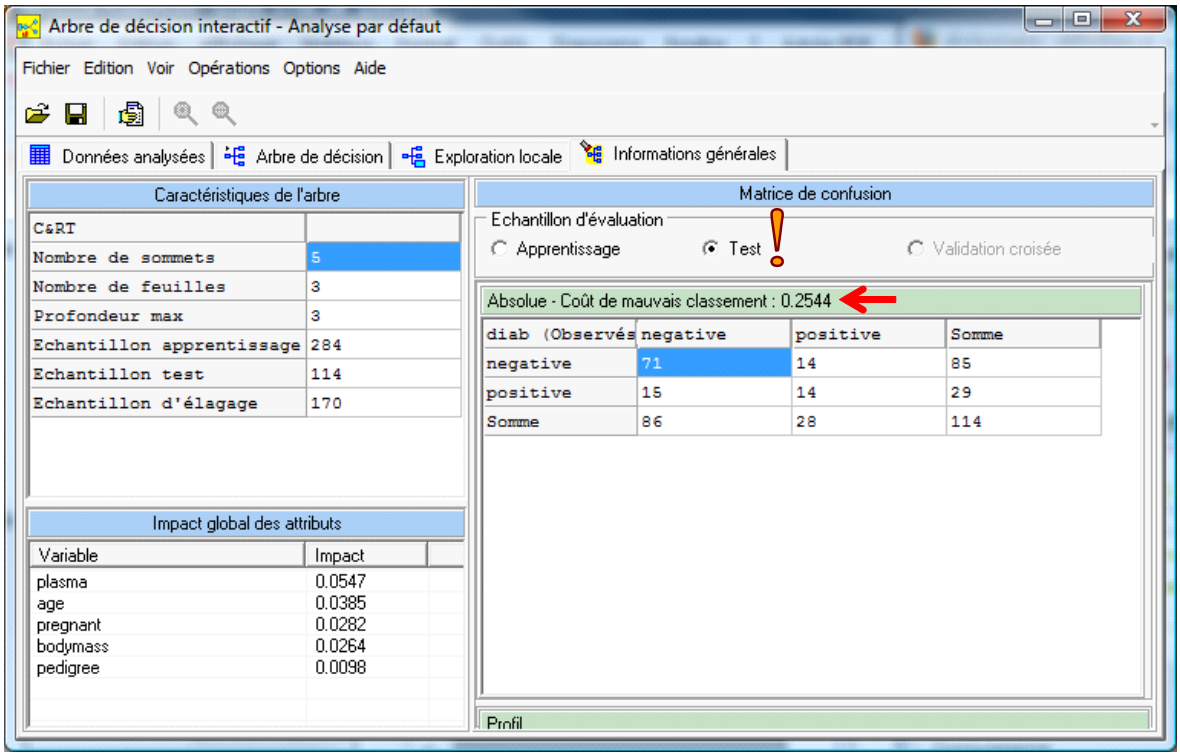


We validate the various options defined during this modification. We obtain a new version of the decision tree.



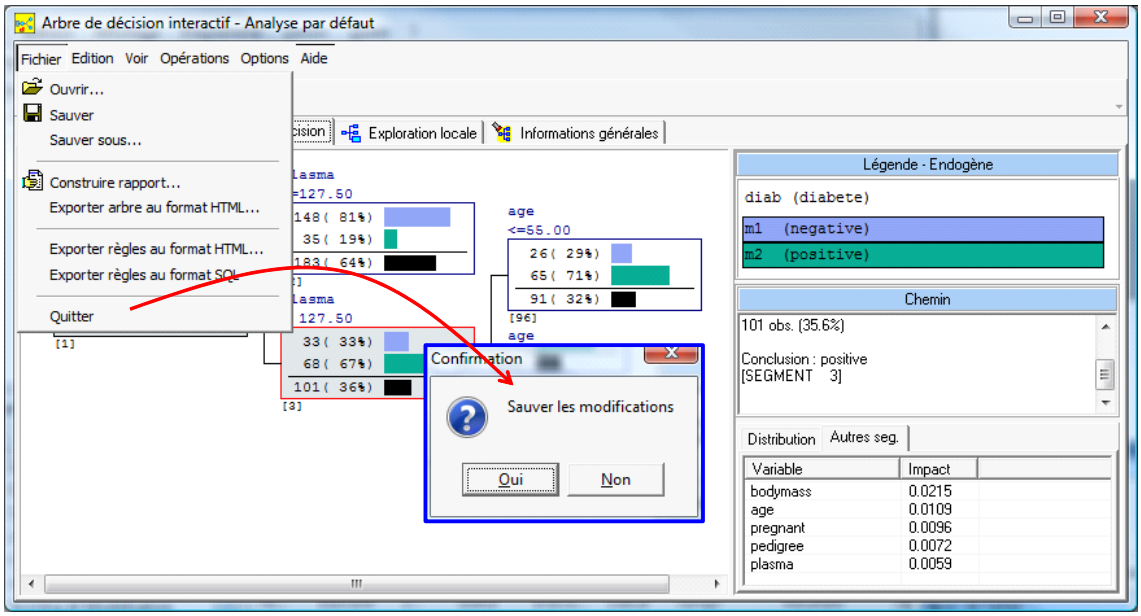
Of course, because we have modified the cut value, the Gini gain is also modified. It is 0.0109 now (vs. 0.0184 previously).

Into the “Informations Générales” tab, we observe the performance of the tree on the training and the testing samples. The test error rate is 25.44% here. That means if we apply the model on unseen cases, the probability of error is about 25.44%. We remember this value afterwards.

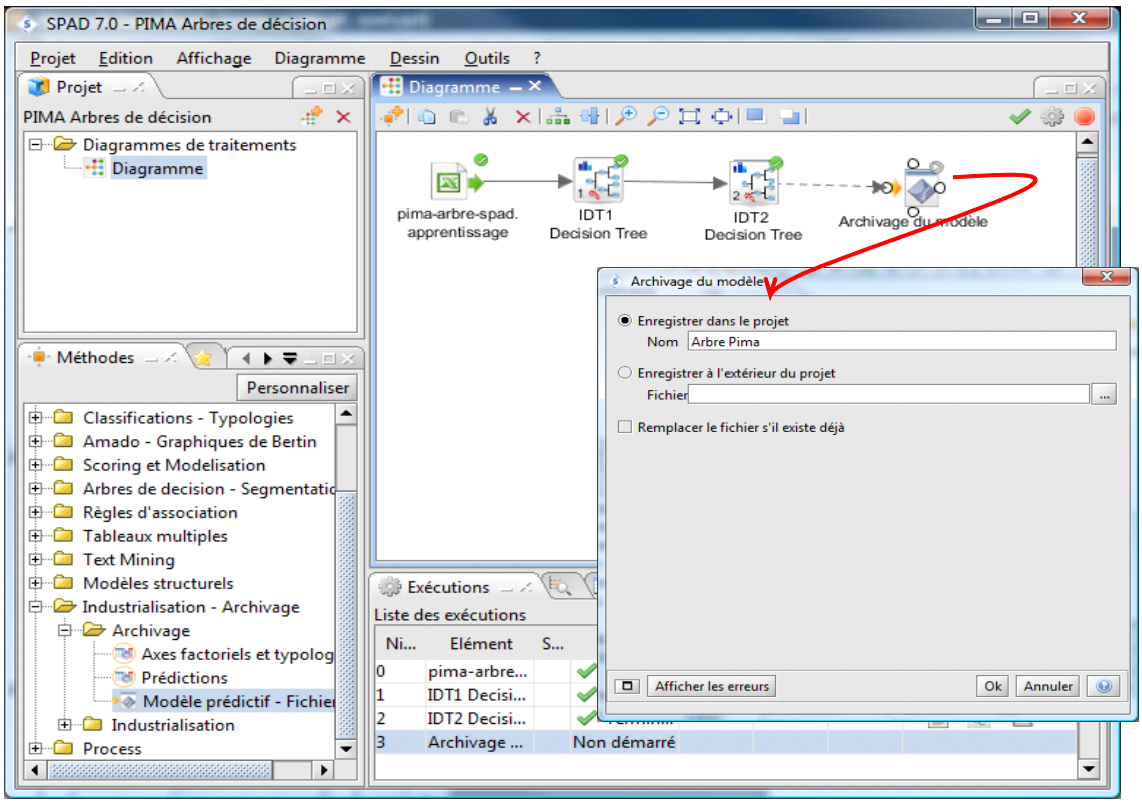


3.1.8 Storing the classifier

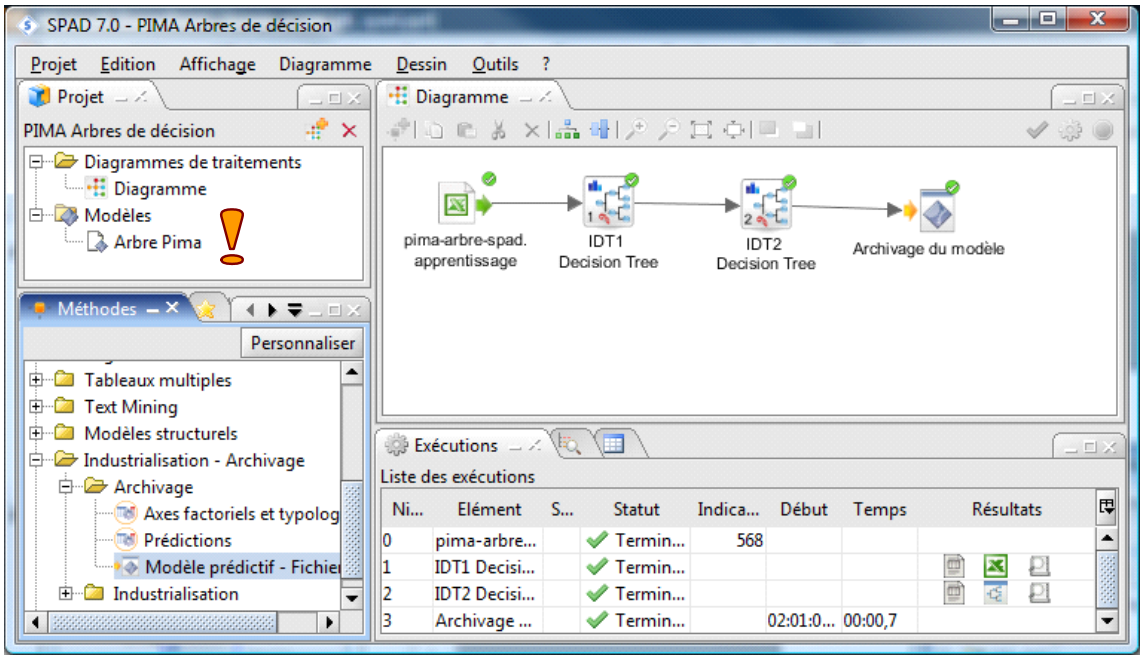
We close the "Arbre de décision interactif" application (FICHIER / QUITTER). We save the modified version of the tree.



We must archive the classifier before to apply it on other data source. We insert the "**Modèle prédictif – Fichier règles"** component into the workflow. We set the name of the model (PARAMETERS menu).



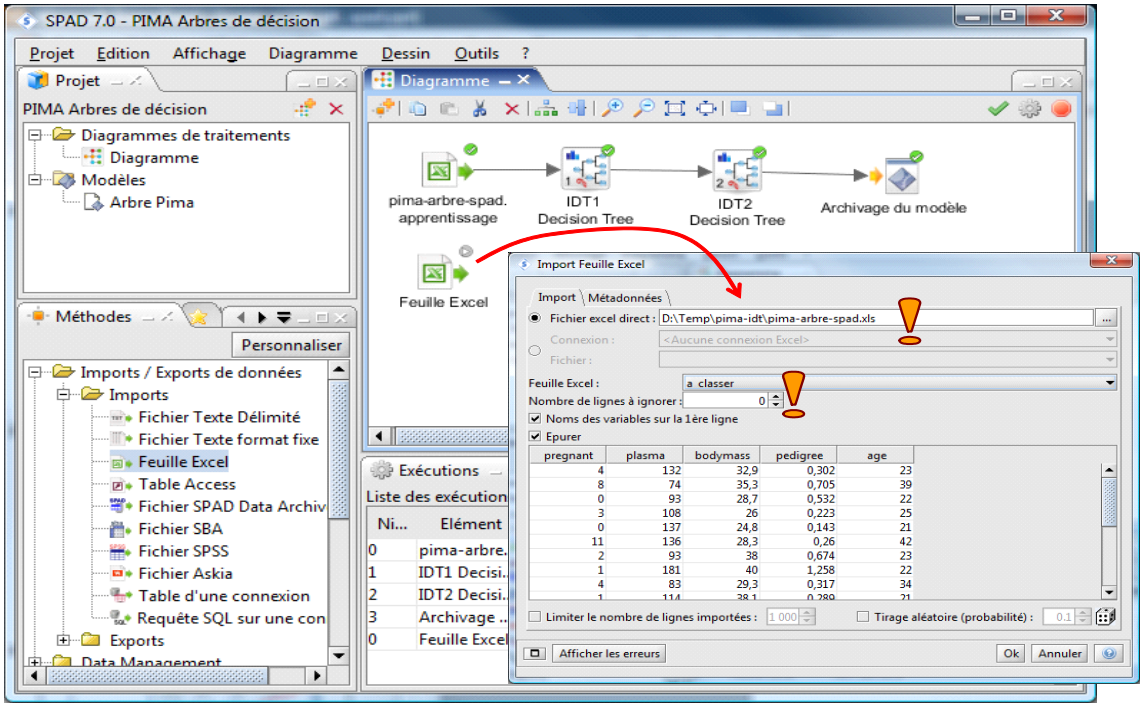
When we validate, we observe that a new icon is available within the branch "Modèle" of the project manager (left part of the window).



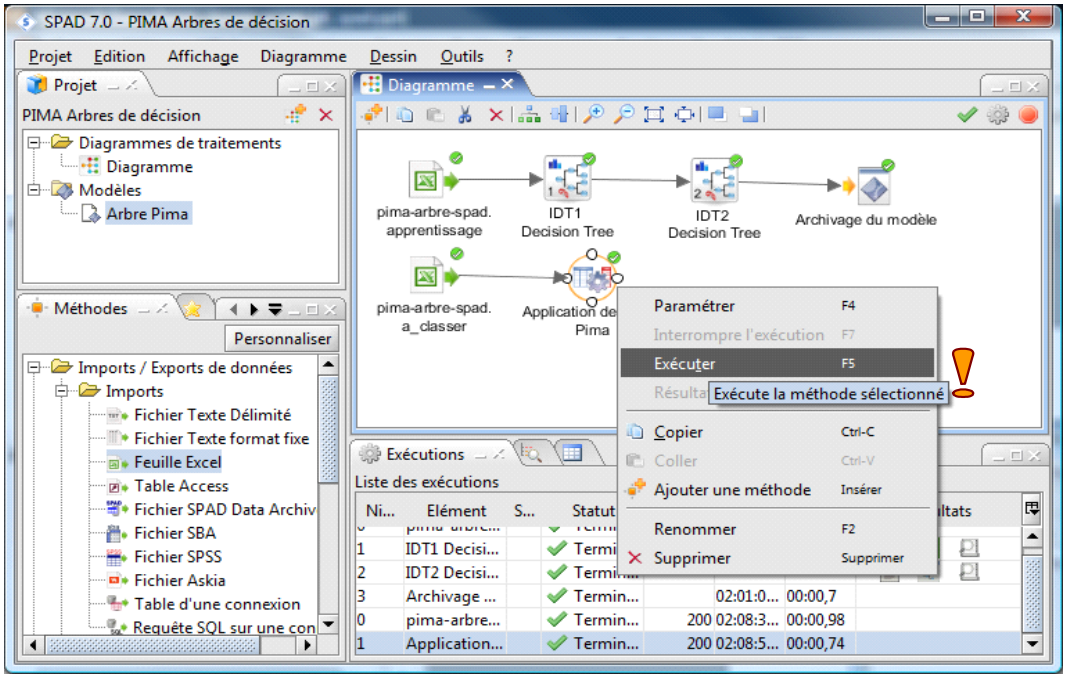
The first step of our analysis is achieved. We have learned a classifier. It is archived now. We can apply it on a new data source.

3.2 Applying the classifier on the « à classer » worksheet

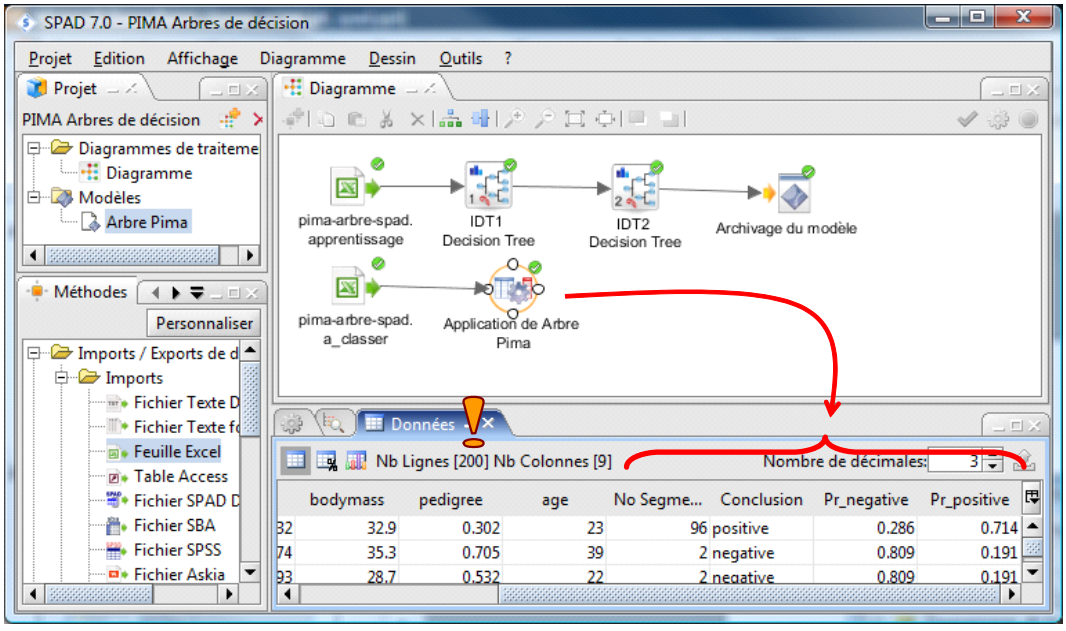
To apply the model on the unlabeled dataset, we must load this last one. We add the **Feuille Excel** component. We select the "à classer" worksheet.



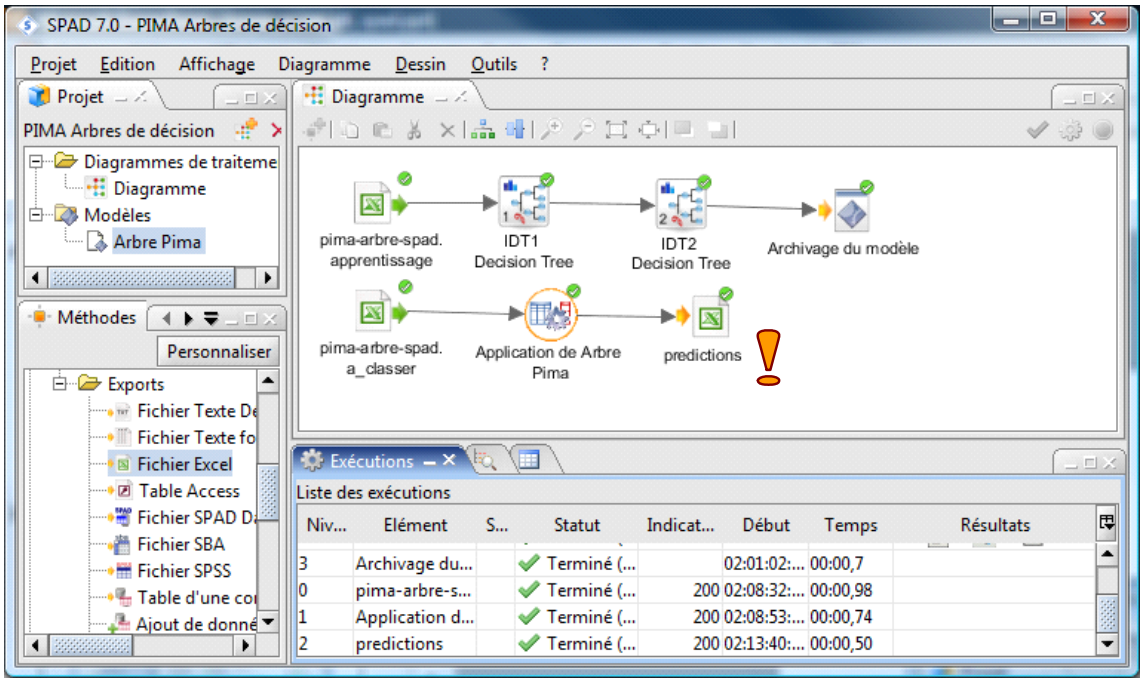
We insert the **Arbre Pima** model into the workspace. We set the connexion with the unlabeled instances. We click on the EXECUTER menu (F5).



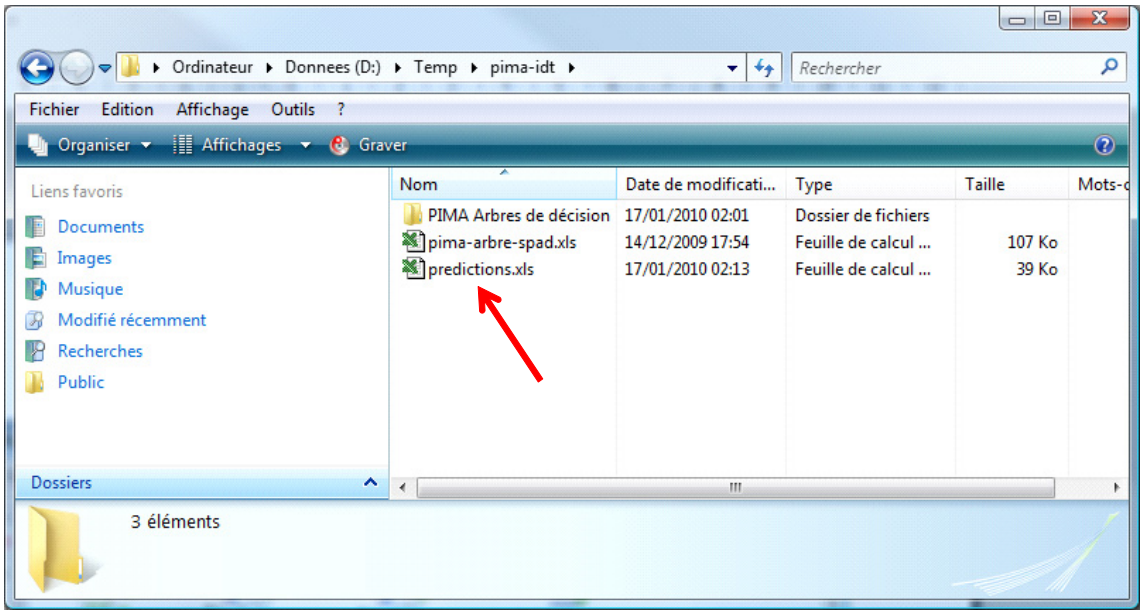
The icon is automatically named “Application de Arbre Pima”. Into the lower part of the main window, we can visualize: the dataset from the worksheet; the internal number used by IDT to identify the leaves of the tree; the model prediction; the estimation of the posterior class probabilities (positive or negative).



We save these information to a new file (Excel format) using the **Fichier Excel** component from the **Exports** branch of the palettes. The data file name is “predictions.xls”.



We confirm the operation. The data file is created on the disk as we see below.



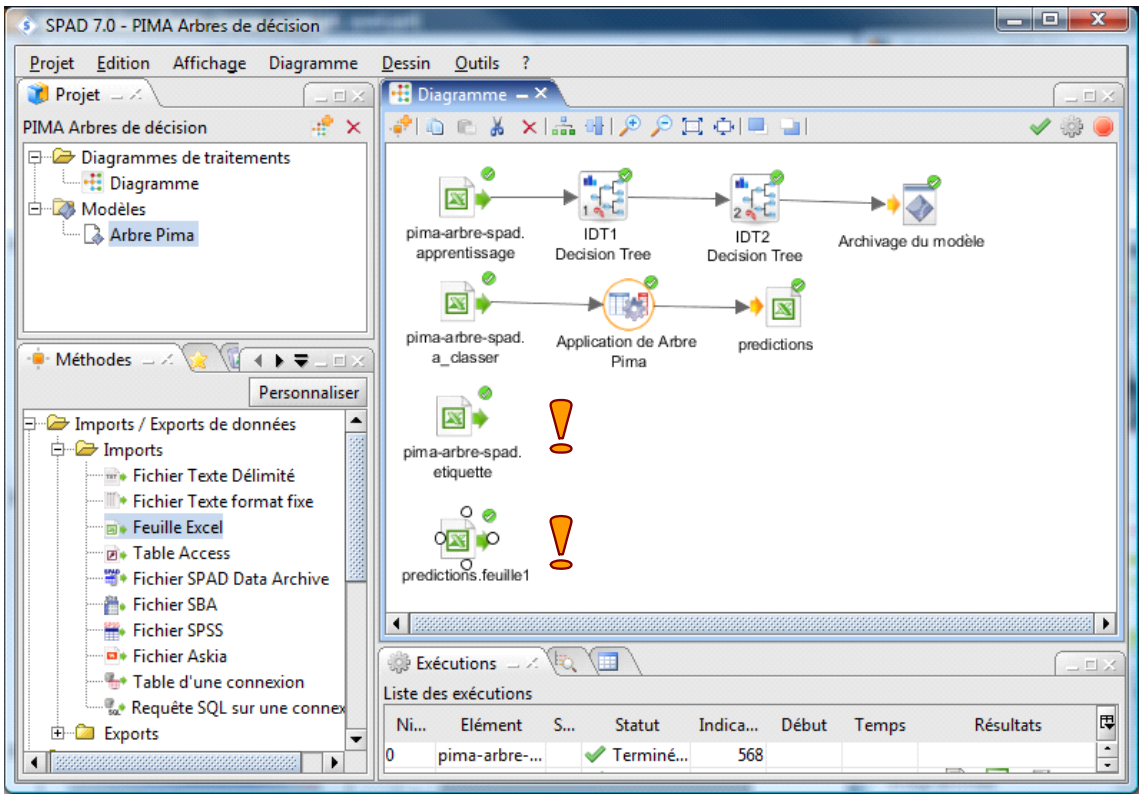
3.3 Comparing the model predictions and the actual values

In practical situation, our analysis is finalized. But, in this tutorial, the actual values of the target attribute are in fact available in the third worksheet "étiquette". We can thus compare the prediction of the model and the observed values to obtain the generalization error rate (this is another estimation of the true error rate of the model).

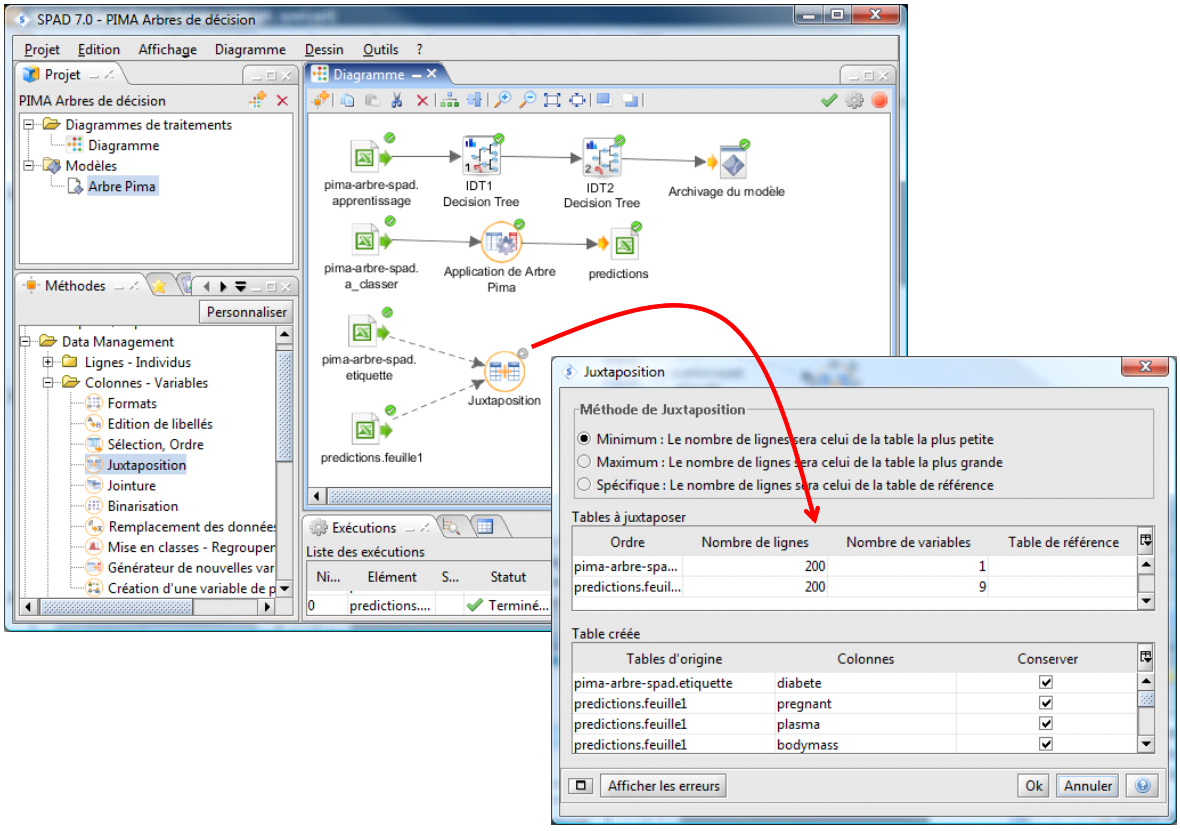
For this, we must: read the two data sources ("prediction.xls" and the third worksheet of the "pima-arbre-spad.xls"); merge them by making the concordance between the individuals; then, compute the error rate by comparing the two columns in a confusion matrix.

3.3.1 Importing and merging the two data sources

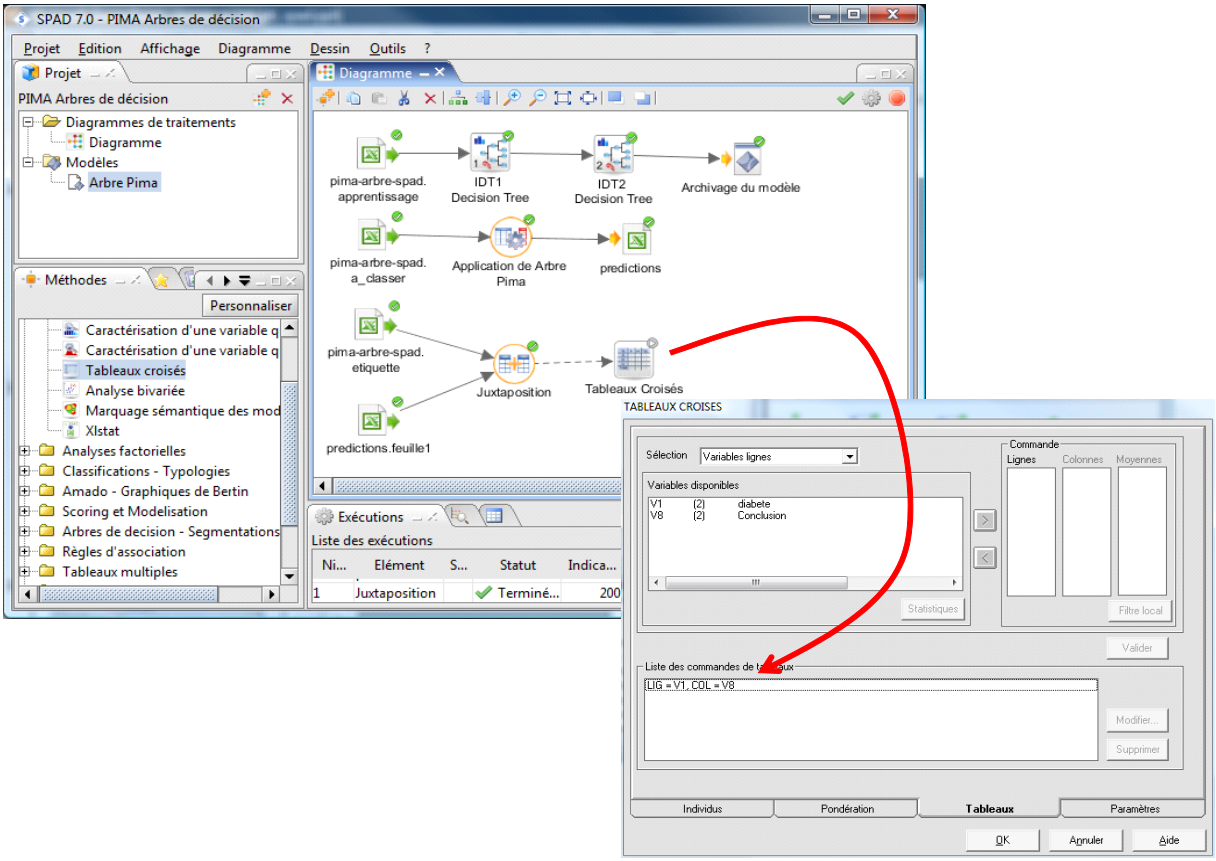
Feuille Excel enables to load the two datasets.



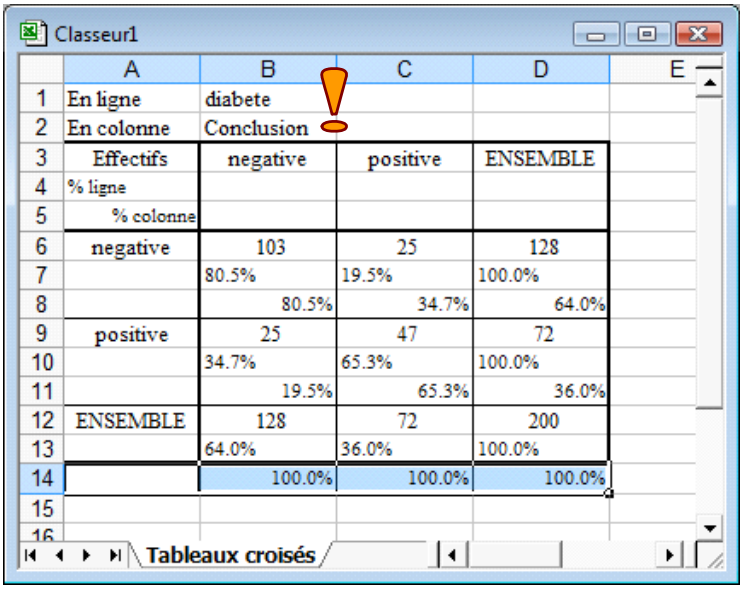
We make the merging operation by using the **Juxtaposition** component. Into the dialog box, we specify the creation settings of the new dataset.



Then we create the cross tab using the **Tableau Croisé** component.



We obtain the following results by activating the RESULTATS / SORTIES EXCEL (F9) menu.



	A	B	C	D	E
1	En ligne	diabete			
2	En colonne	Conclusion			
3	Effectifs	negative	positive	ENSEMBLE	
4	% ligne				
5	% colonne				
6	negative	103	25	128	
7		80.5%	19.5%	100.0%	
8		80.5%	34.7%	64.0%	
9	positive	25	47	72	
10		34.7%	65.3%	100.0%	
11		19.5%	65.3%	36.0%	
12	ENSEMBLE	128	72	200	
13		64.0%	36.0%	100.0%	
14		100.0%	100.0%	100.0%	
15					
16					

The generalization error rate is

Taux d'erreur = (25 + 25) / 200 = 1.0 – (103 + 47) / 200 = 25%

We note that it is very close to the test error rate computed previously (25.4%, see section 3.1.7). The cross tab supplies also various indicators (e.g. the sensibility is 47 / 72 = 65.3%; the specificity is 103 / 128 = 80.5%; etc.).

4 Analysis under R software

The various operations above can be achieved using the vast majority of data mining tools, but perhaps not with the same ease. For instance, a few lines of code are sufficient under the R software (http://www.r-project.org/). Of course, we must know the right instructions and know how to organize them. The drawback is that the trees are not interactive under R[3].

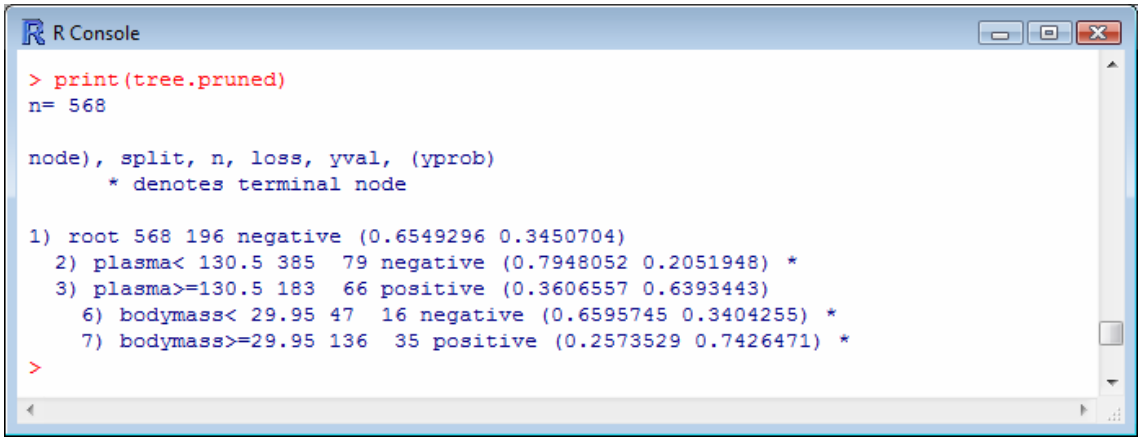
Here is the source code used for the analysis.

```
#clear the memory
rm (list=ls())

#load the dataset
library(xlsReadWrite)
#load the various worksheets of the workbook
pima.train <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="apprentissage")
pima.unlabeled <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="a_classer")
pima.label <- read.xls(file="pima-arbre-spad.xls",colNames=T,sheet="etiquette")

#learning the decision tree using the rpart package
library(rpart)
#creating the unpruned tree
tree.unpruned <- rpart(diabete ~ ., data = pima.train)
#post pruning using the cp parameter
plotcp(tree.unpruned)
#lwe use cp = 0.04
tree.pruned <- prune(tree.unpruned,cp=0.04)
#visualizing the decision tree
print(tree.pruned)

#applying the model and computing of the confusion matrix
#applying the tree on the unseen cases
prediction <- predict(tree.pruned,newdata = pima.unlabeled,type="class")
#comparing the predictions and the actual values of the target attribute
conf.matrix <- table(pima.label$diabete,prediction)
print(conf.matrix)
#computing the generalization error rate
error.rate <- 1.0 - (conf.matrix[1,1]+conf.matrix[2,2])/sum(conf.matrix)
print(error.rate)
```

[3] With Sipina (http://eric.univ-lyon2.fr/~ricco/sipina.html) on the other hand, we can create interactively the tree, but we cannot define a sequence of operations as lines of codes.

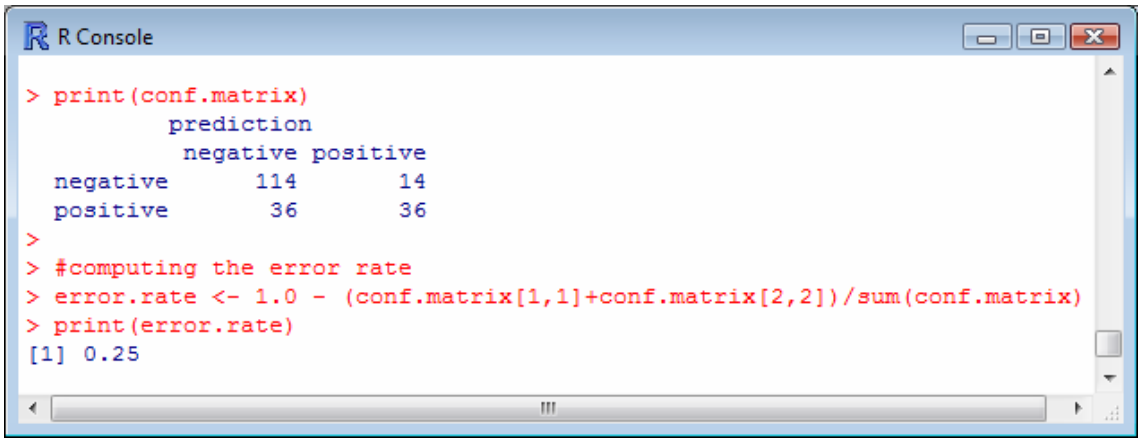
Here are the most important results.

(1) The decision tree obtained after the post pruning process.



« Plasma » and « Bodymass » are the used explanatory variables.

(2) The confusion matrix and the generalization error rate



5 Conclusion

Software is not intended to replace data miners. However, it must give us the tools to free ourselves from a lot of tedious, repetitive tasks: access to files which can be dispersed, data preparation (e.g. cleaning, creation of intermediate variables), formatting reports, etc.

In this tutorial we showed a sequence of treatments using the SPAD software. We had loaded the data file, created a decision tree, applied the model to new instances, assess its validity by comparing the predicted and actual values of the target attribute.

As we tell before, most of the data mining tools enable to perform this kind of process. The difference lies in the easiness to achieve each operation and the ability to link them (easily).