Subject

Building decision tree with TANAGRA, ORANGE and WEKA. Error rate estimation using a cross-validation.

When we build a decision tree from a dataset, we much follow the following steps (not necessarily in the same order):

- Import the dataset in the software;
- Select the class attribute (TARGET) and the descriptors (INPUT);
- Choose the induction algorithm, according the implementation we can obtain slightly different results;
- Learning process and viewing the decision tree;
- Use cross-validation in order to obtain an honest error rate estimate.

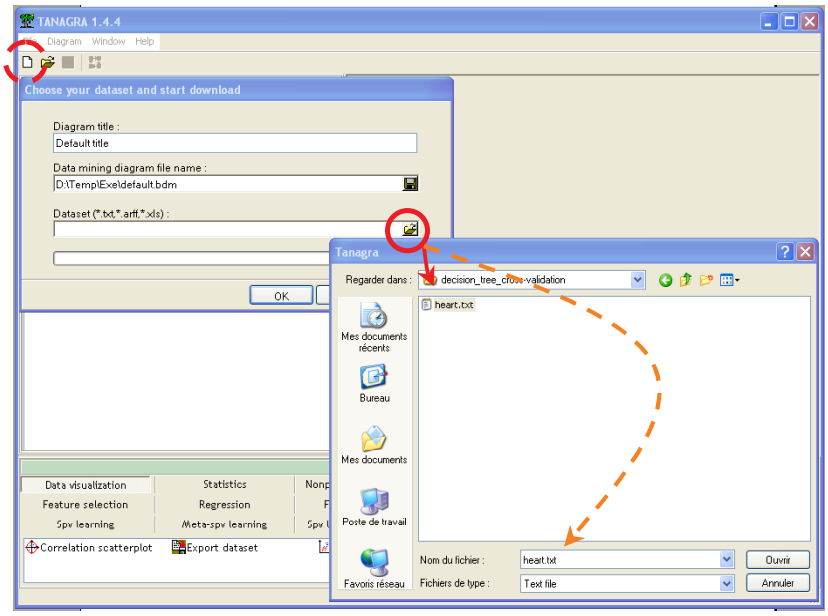
Dataset

We use the HEART.TXT (UCI IRVINE), some attributes are deleted; there are 270 examples in the dataset.

Building a decision tree with TANAGRA

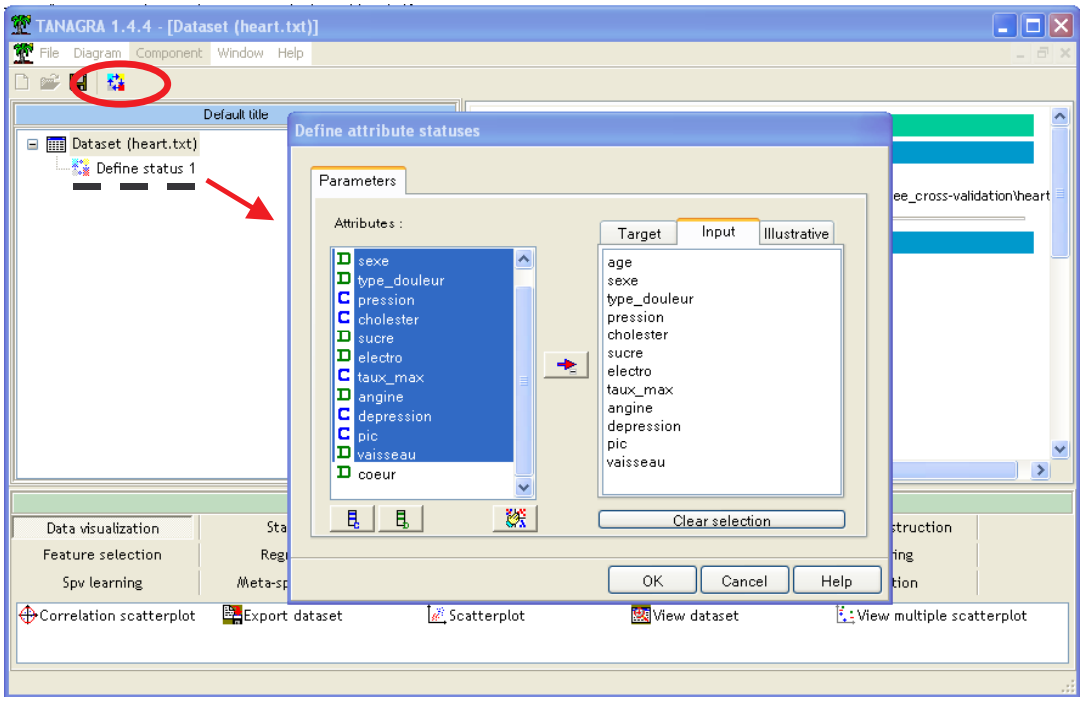
Importing the dataset

We create a new diagram and import the dataset with the FILE/NEW menu.



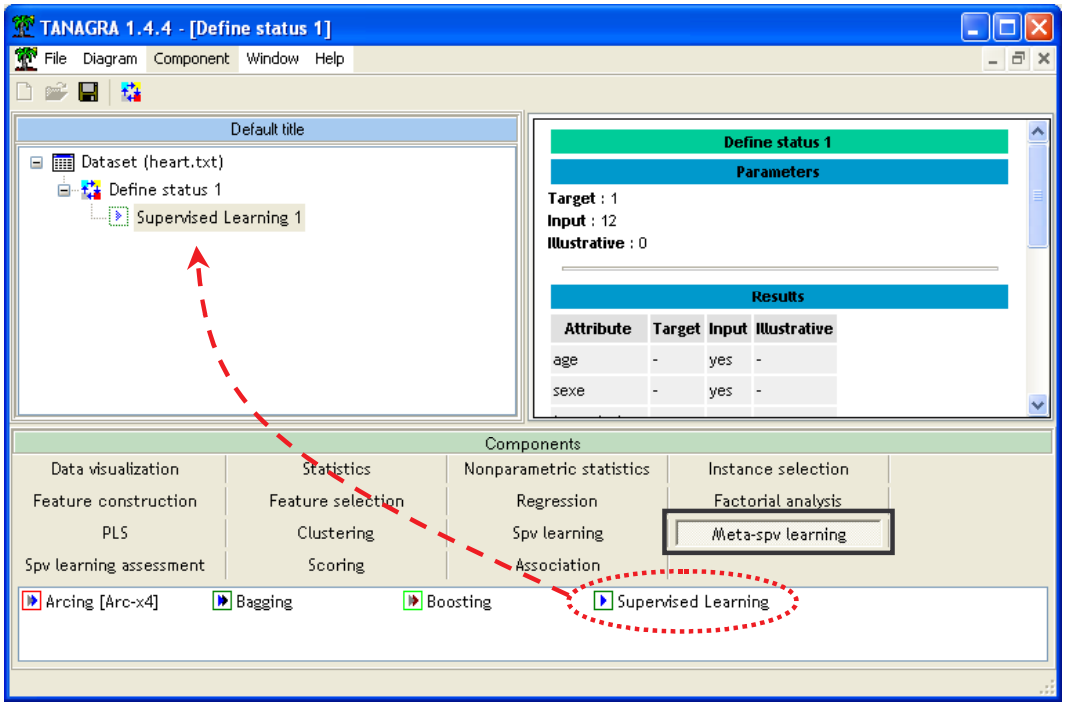
Defining the role of attributes

We add the DEFINE STATUS component (we use the button in the toolbar) in the diagram. We set COEUR as TARGET, the other attributes as INPUT.

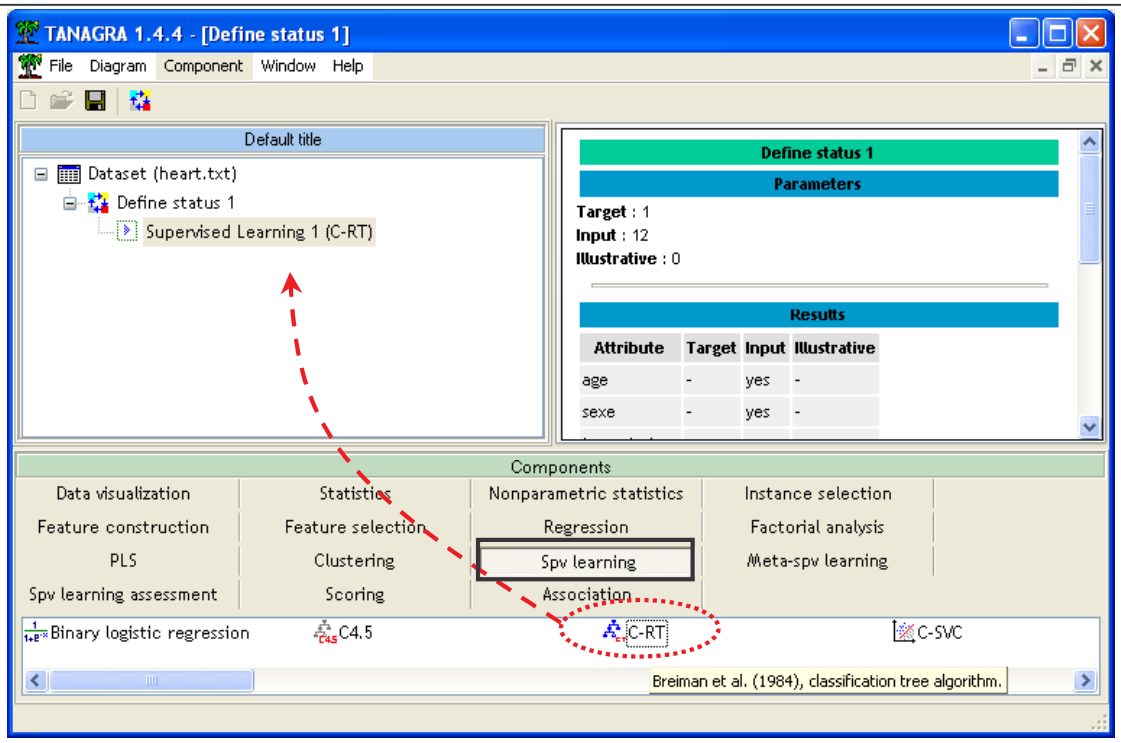


Selecting the learning algorithm

We want to use the *Classification and Regression Tree* (Breiman et al.) algorithm. There are two steps when we want to define a supervised learning process in TANAGRA: (a) we insert a meta-supervised learning component from the META SPV LEARNING tab…

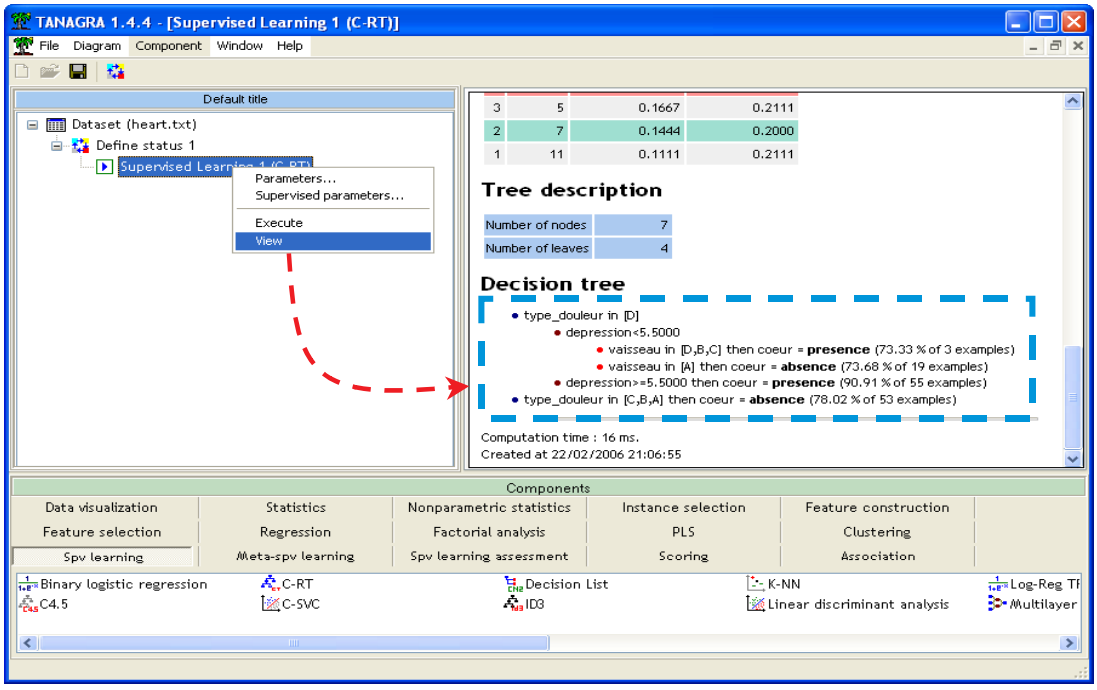


(b) … and embed the learning algorithm, C-RT, from the SPV learning tab.



Displaying the results

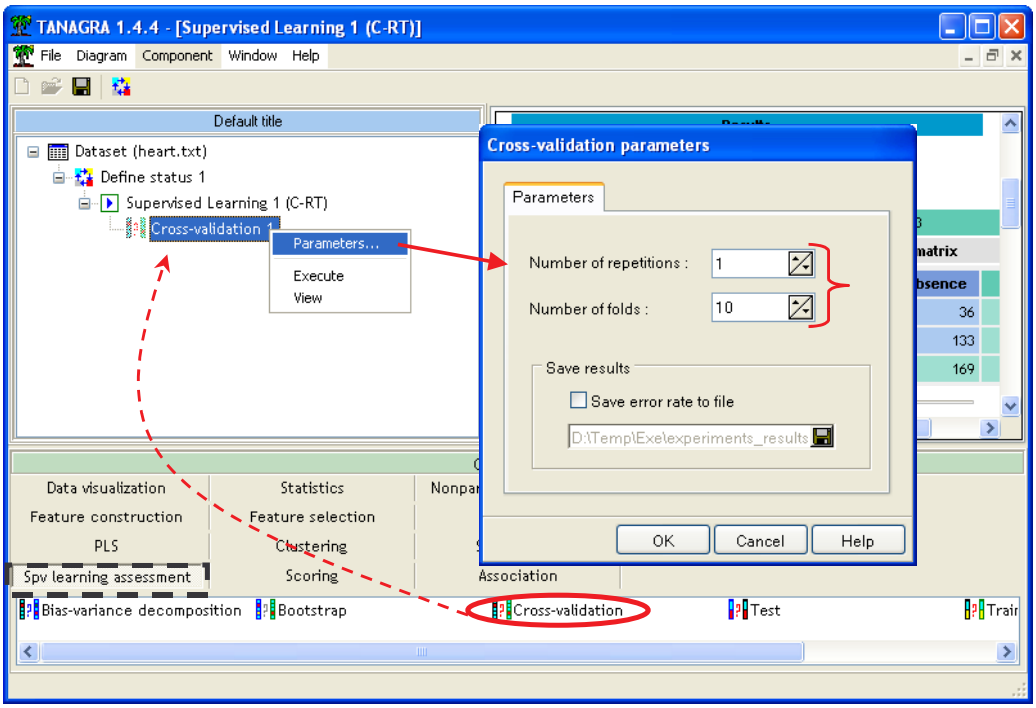
In order to view the decision tree, we click on the VIEW menu of the last component, we see the tree[1]: the resubstitution error rate is 19.63%; the tree has 4 leaves (4 rules).



[1] TANAGRA uses a textual representation, if you want a graphical representation, you can try the SIPINA software from the same author (http://eric.univ-lyon2.fr/~ricco/sipina.html).

Cross-validation

We want to compute the error rate with a cross-validation resampling method. We add the CROSS-VALIDATION component from the SPV LEARNING ASSESMENT tab. We set the number of folds to 10, and the number of repetition to 1.



The estimated error rate is 24.81%

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	10
Trials	1

Results

CV error rate

Range	
MIN	0.2481
MAX	0.2481
Trial	**Err rate**
1	0.2481

Overall cross-validation error rate

Error rate	0.2481

Values prediction

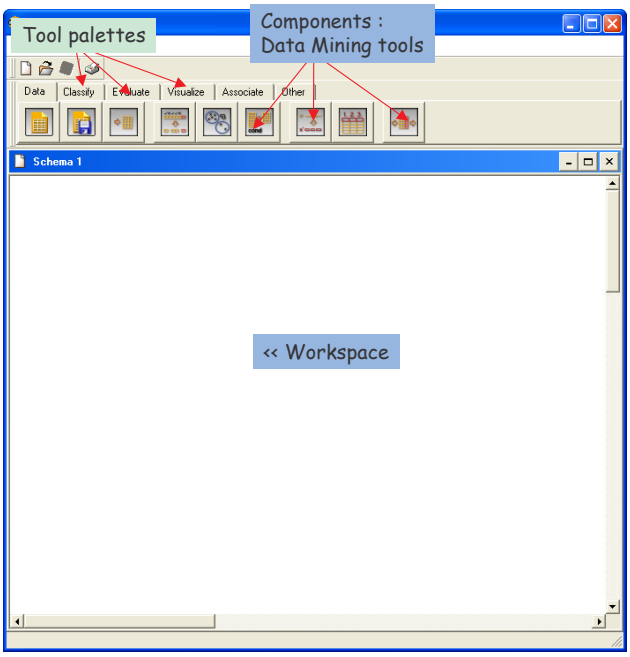
Value	Recall	1-Precision
presence	0.6250	0.2268
absence	0.8533	0.2601

Confusion matrix

	presence	absence	Sum
presence	75	45	120
absence	22	128	150
Sum	97	173	270

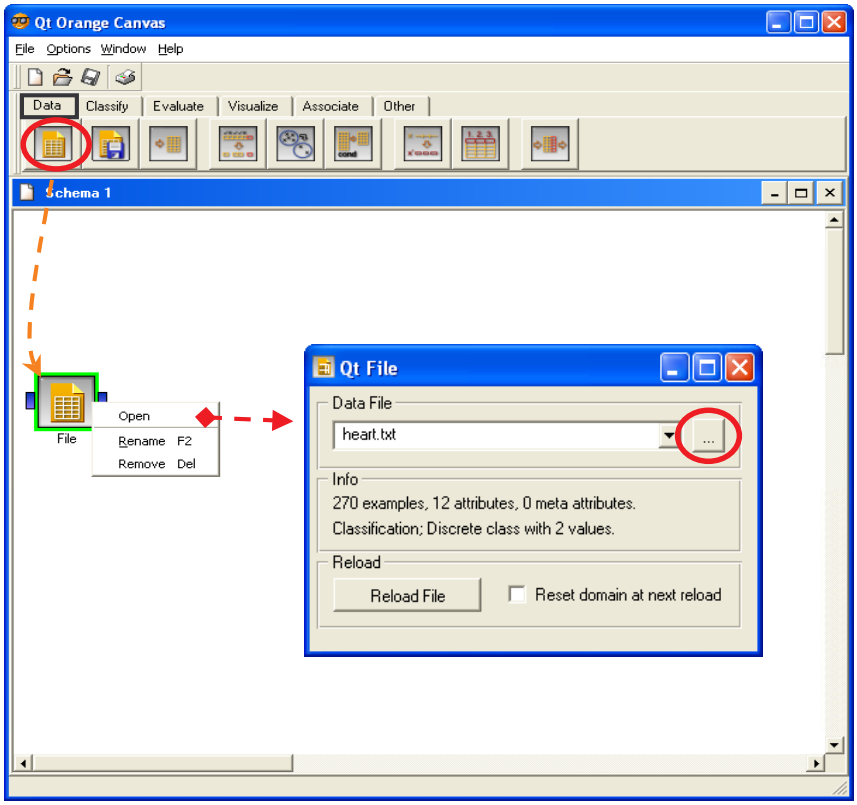
Building a decision tree with ORANGE

When we execute ORANGE, we have the following interface.



Importing the dataset

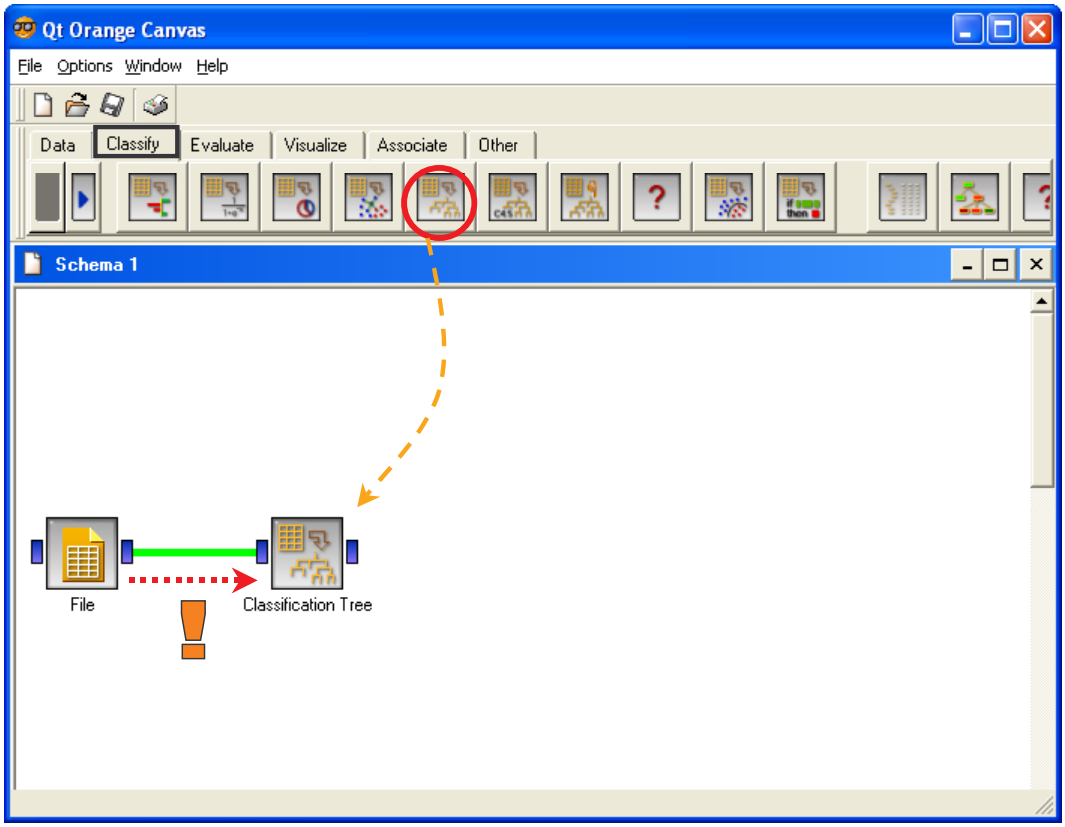
ORANGE can handle text file format (tabulation separator). When we select the tool, a new component is inserted in the diagram. We can select the file with the OPEN contextual menu.



Learning process

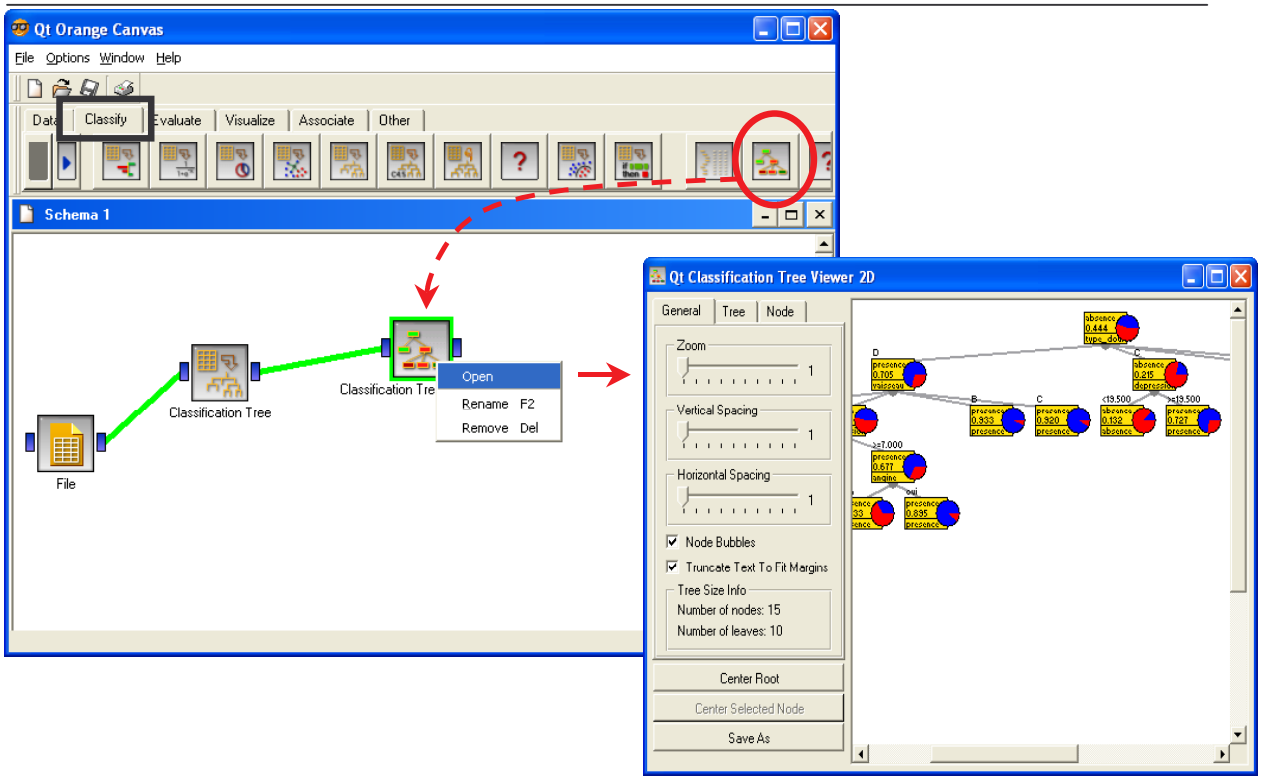
By default, the target attribute is the last column; the others are the input attributes. We have the right configuration in our dataset.

We can add the classification tree component (CLASSIFY tab) in our diagram. We connect this component with the dataset component.



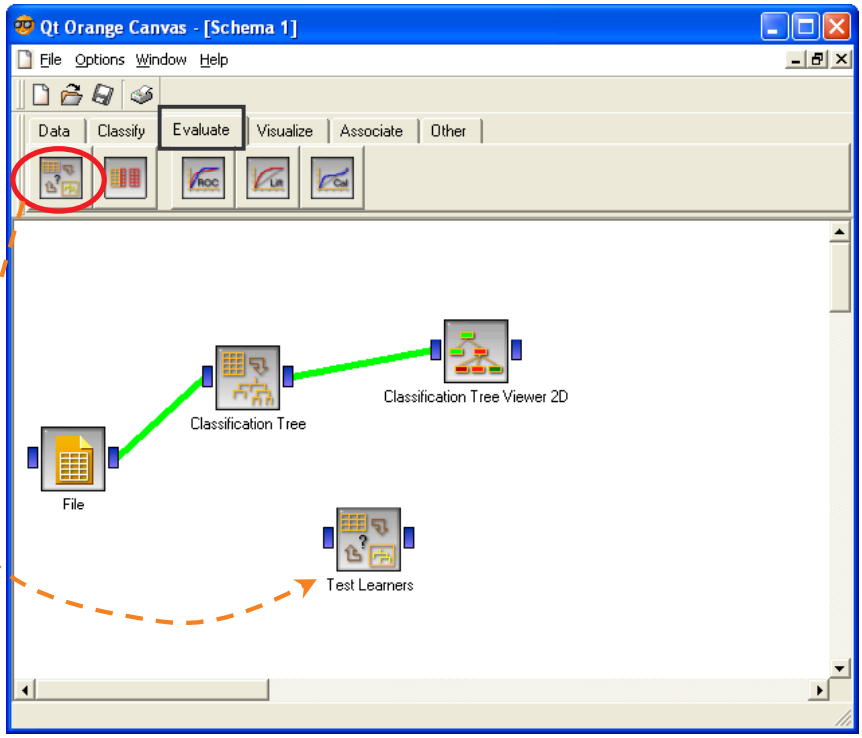
Decision tree visualization

We can display the tree in a text viewer, it is recommended if we have numerous nodes in the tree; there is also a graphical viewer that is more pleasant (CLASSIFICATION TREE VIEWER 2D – CLASSIFY tab). We connect the CLASSIFICATION TREE component to this last one. We click on the OPEN menu in order to display the tree. There are 10 rules (leaves) in our tree.



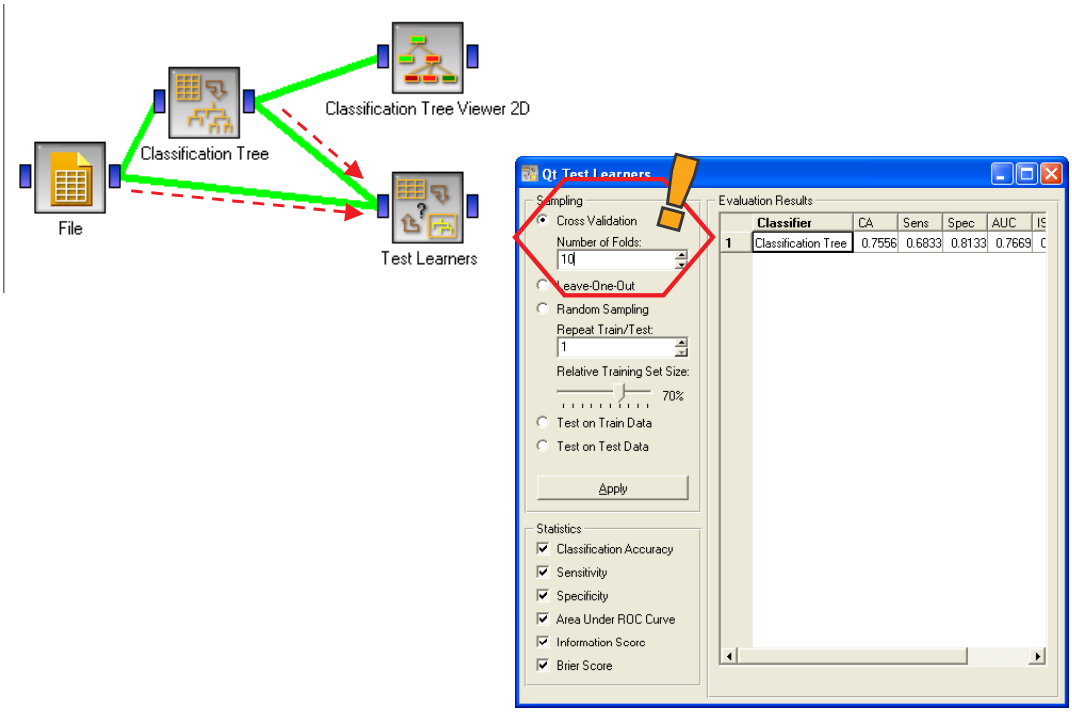
Cross-validation

The TEST LEARNERS component (EVALUATE tab) enables to compute the cross-validation error rate estimate. We connect to this new component the classification tree.



This component becomes operational when we will have specified the data source and the training method -- it is possible to connect simultaneously several learning methods, which

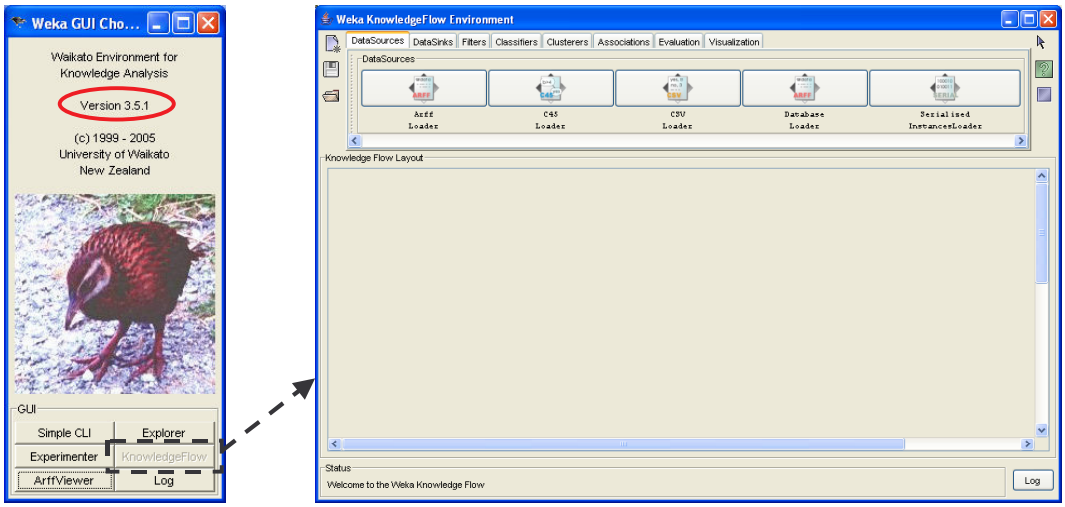
makes it possible to realize, very easily, the comparison of performances. We thus carry out the right connections, and then we display the results using the OPEN menu.



The classification accuracy is 75.56%; the error rate is 24.44%. Other statistics are available. We can also interactively choose another resampling method.

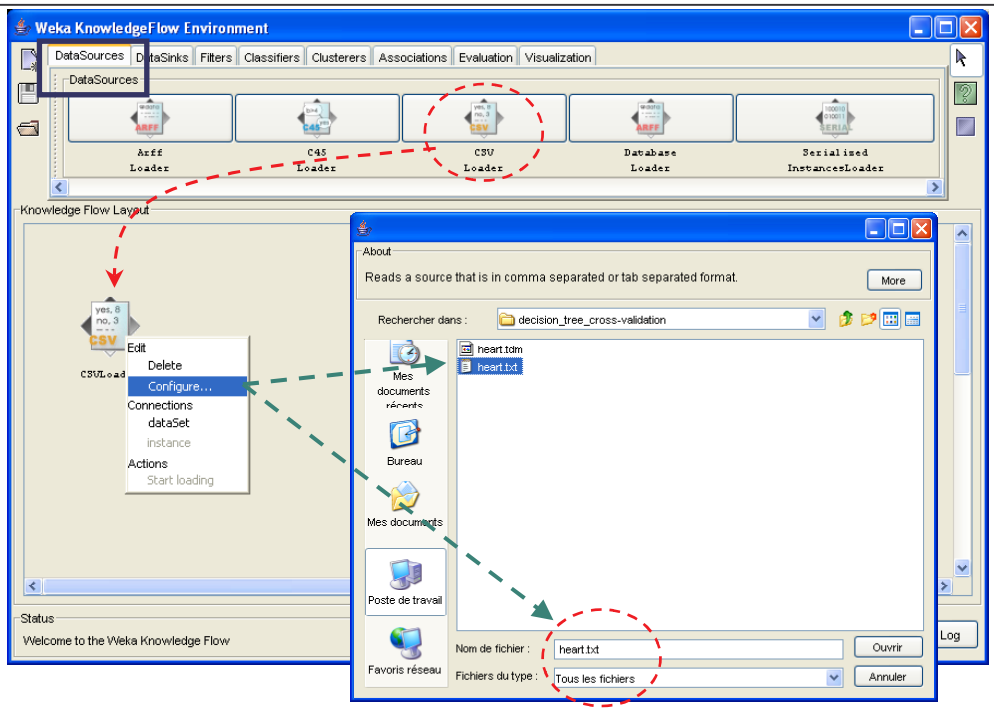
Building a decision tree with WEKA

A dialog box appears when we execute WEKA; we choose the **KNOWLEDGE FLOW** paradigm. We have used the **3.5.1** version.



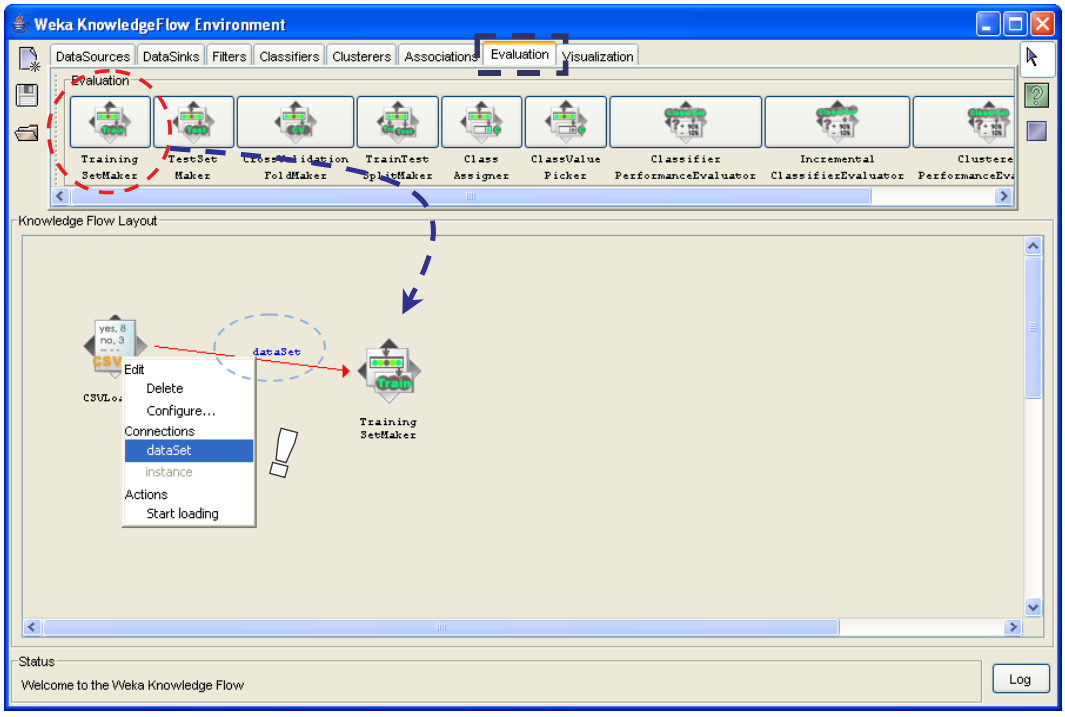
Importing the dataset

The CSV LOADER enables to handle text file format. We select the HEART.TXT dataset with the CONFIGURE contextual menu.



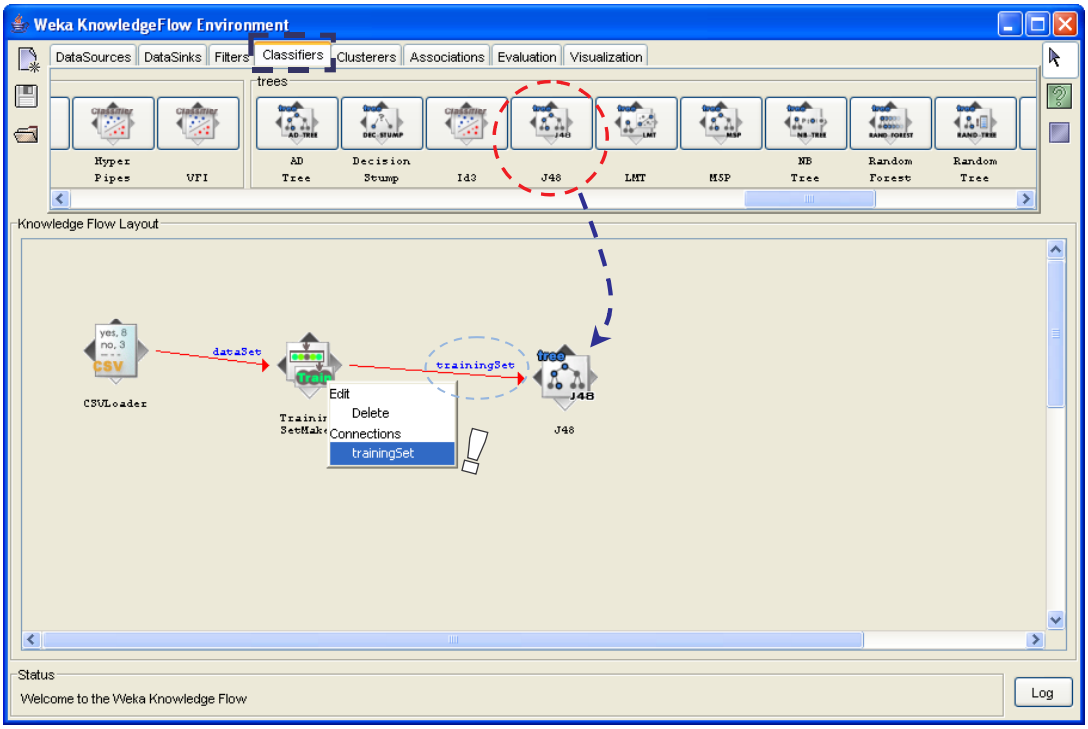
Learning process

By default, the target attribute is the last column; the others are the input attributes. We have the right configuration in our dataset. On the other hand, we must explicitly select the learning set in the WEKA diagram. We add the TRAINING SET MAKER (EVALUATION tab) in the diagram; all examples are used for the construction of the decision tree. We choose the DATASET connection when we connect the LOADER component to this last component.

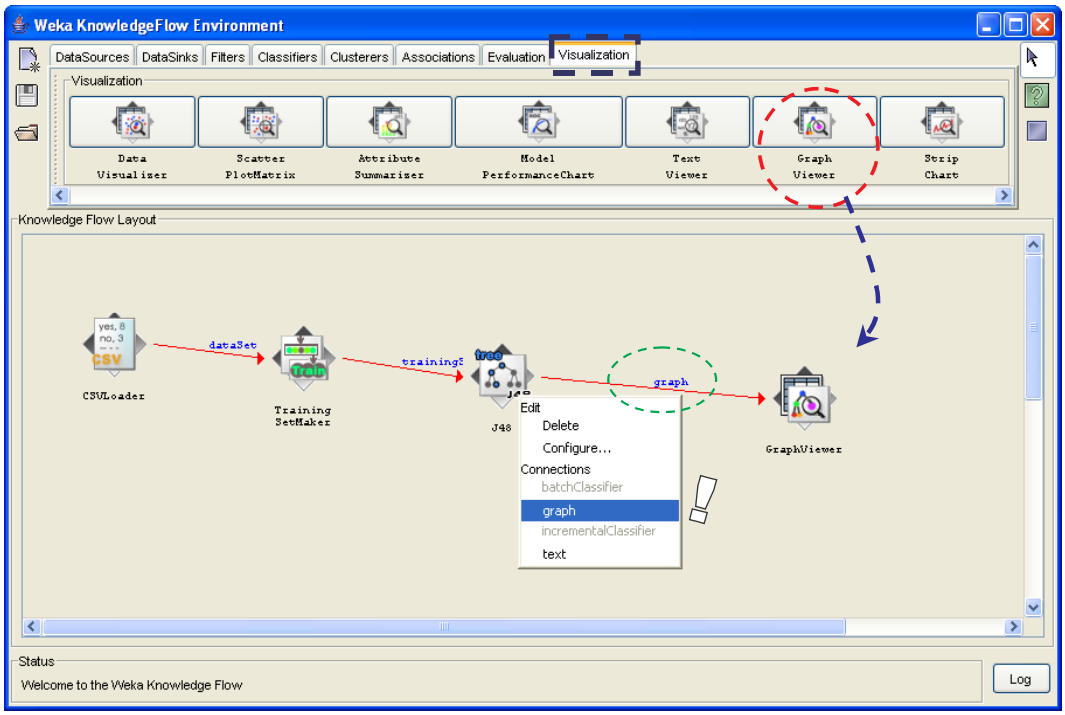


We add the J48 component (decision tree algorithm such as C4.5, CLASSIFIERS tab). We set the connection between TRAINING SET MAKER and J48 (training set connection).

Decision tree visualization

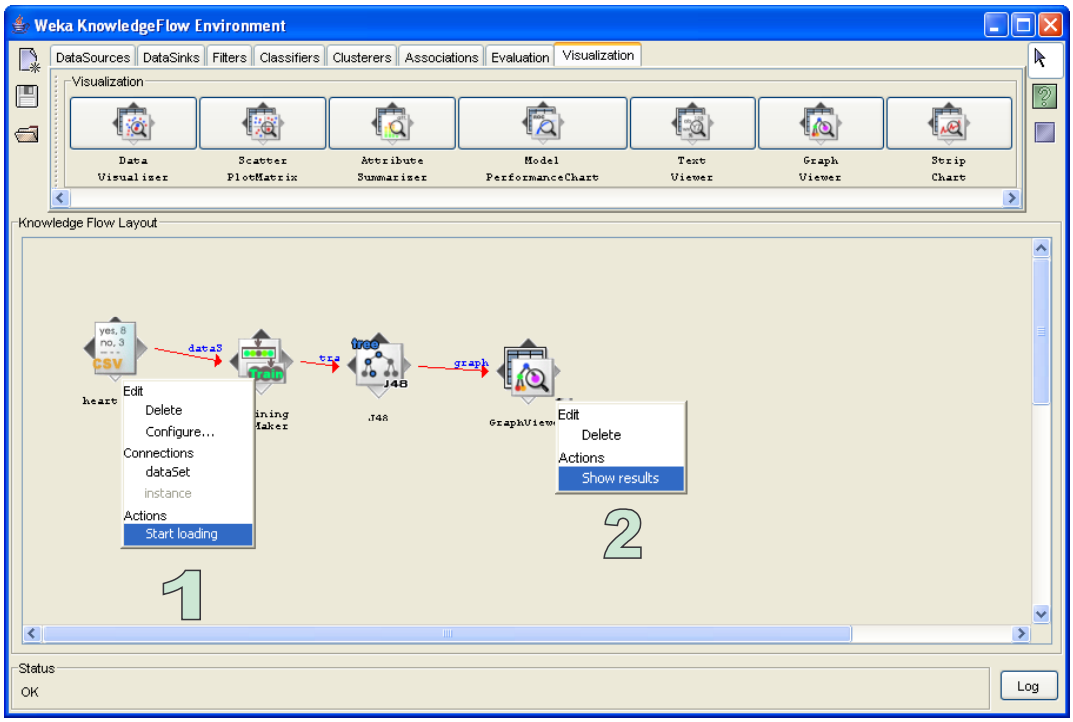


We have two representations in WEKA: a textual representation, suggested when we have a lot of nodes in the tree; a graphical representation that is more pleasant. We select this last one (GRAPH VIEWER – VISUALIZATION tab) and use the GRAPH connection.

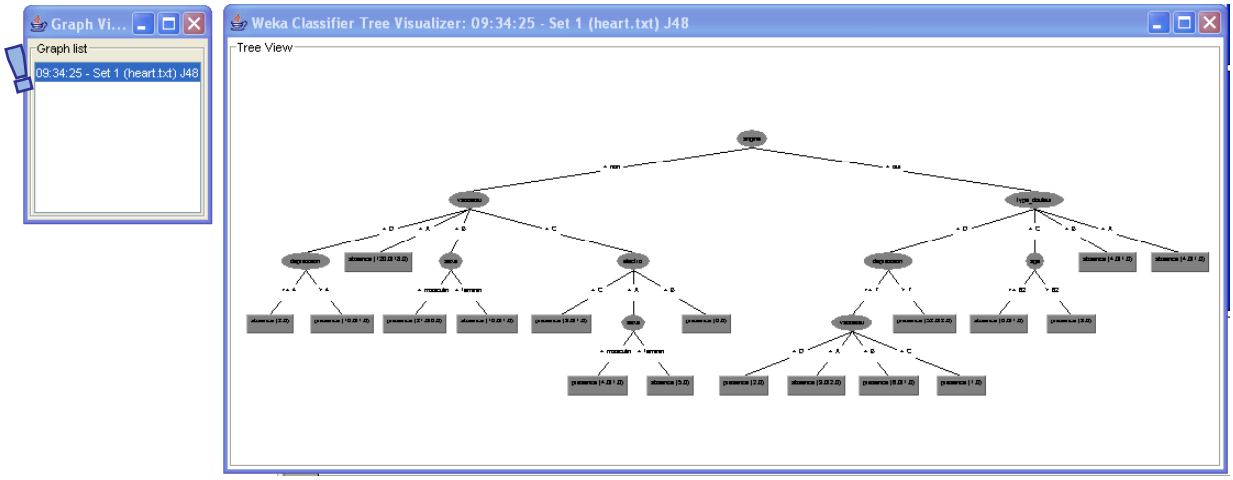


In order to start the execution, we select the first node of the diagram and click on the START LOADING contextual menu.

When the computation is achieved, we can select the last component (GRAPH VIEWER) and click on the SHOW GRAPH menu.

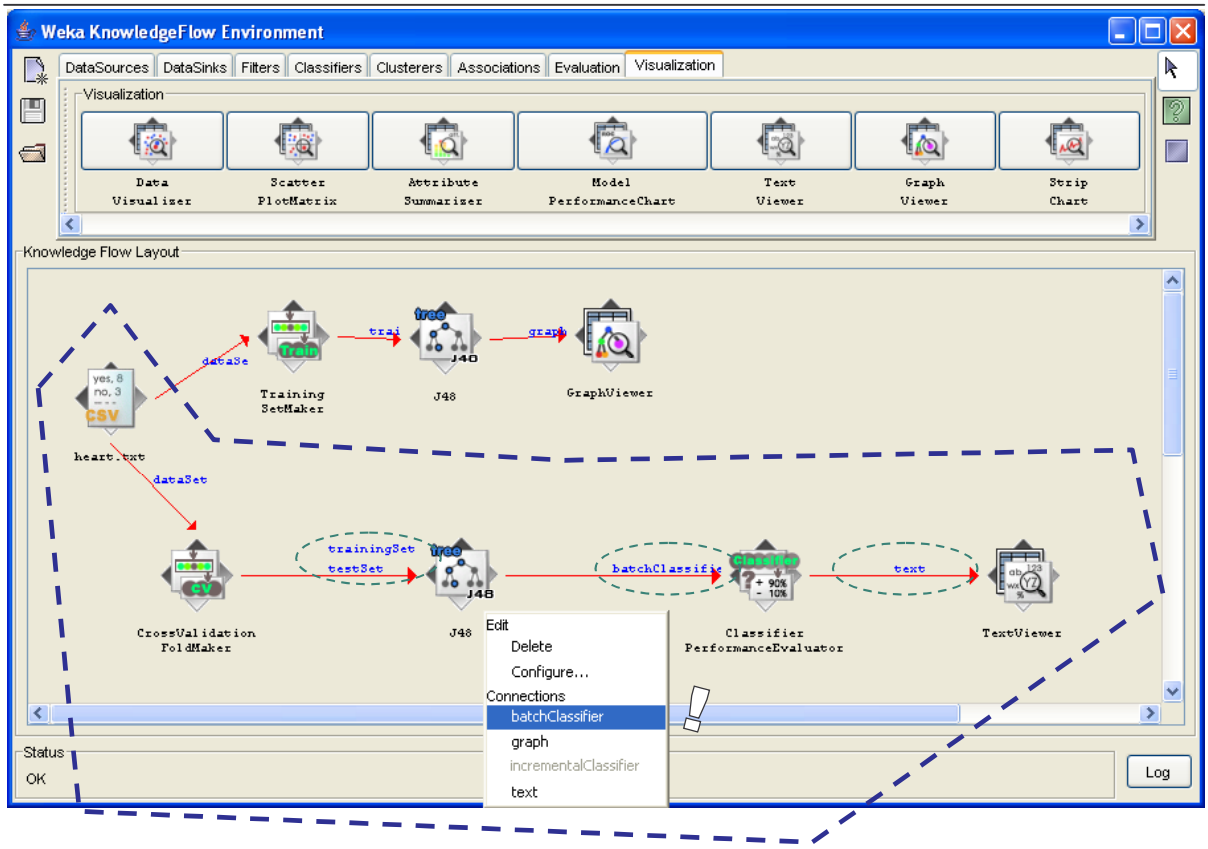


The decision tree has 18 leaves.



Cross validation

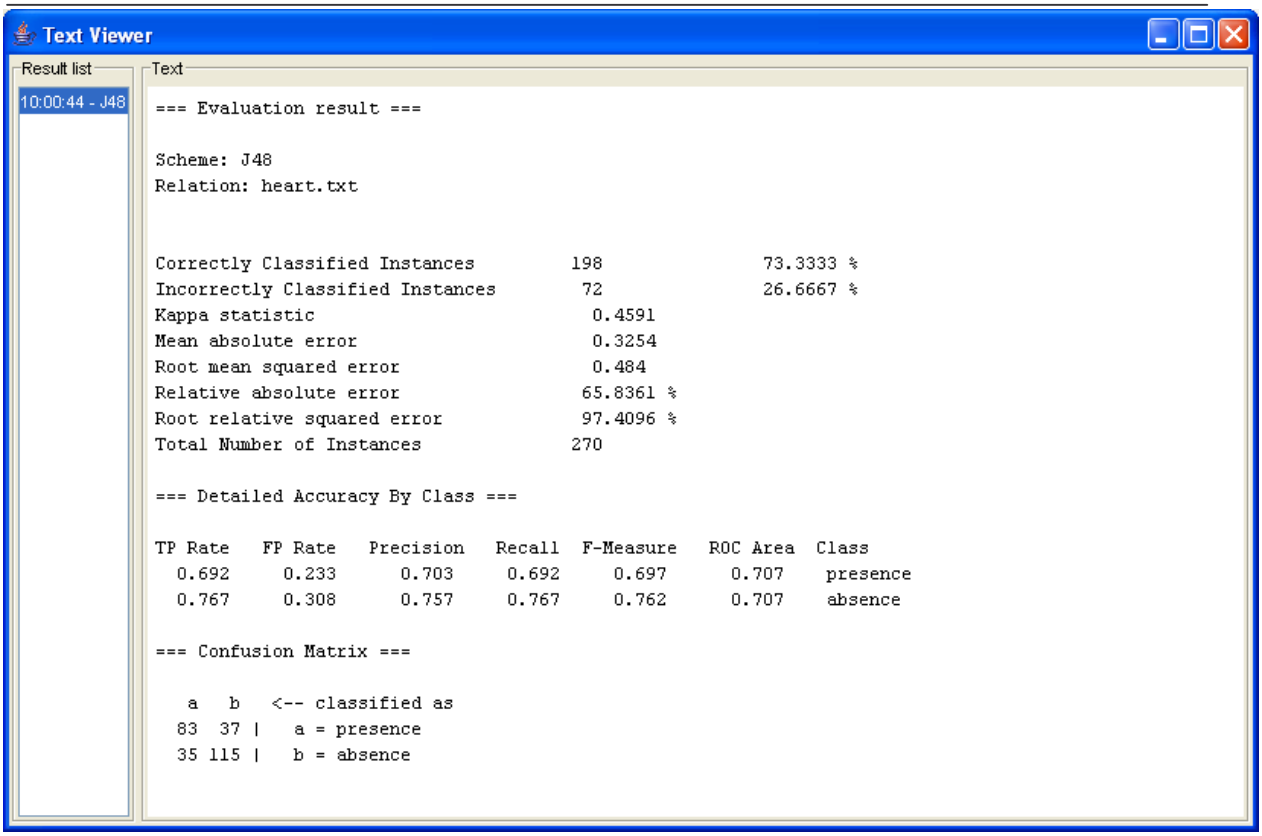
WEKA has at one's disposal sophisticated error rate estimation but needs to create a new sequence of components to do that.



We need to the following components:

- CROSS VALIDATION FOLD MAKER (EVALUATION), which builds folds (DATASET connection).
- Decision tree J48 component (CLASSIFY); be careful, we have to set the same parameters as the precedent J48 component. We must connect twice the CROSS VALIDATION FOLD MAKER to this component, for the training and the test sets.
- CLASSIFIER PERFORMANCE EVALUATOR (EVALUATION) computes the error rate in each fold. We use the BATCH CLASSIFIER output of J48.
- Last, the TEXT VIEWER component displays the results.

One again, we select the START LOADING of the CSV LOADER component in order to start the execution. The SHOW RESULTS menu of TEXT VIEWER displays the following results.



The computed error rate is 26.67%. Other statistics are available.

Let us note a very useful characteristic of WEKA; it is possible to visualize the 10 decision trees computed during the cross validation process. It would be necessary for that to connect a component TEXT VIEWER at the output of the J48 component, we can see the possible differences between the trees and judge the stability of computations.

Conclusion

Cross-validation is a very popular method of error rate estimation, especially when we have a few examples in our dataset. We see in this tutorial that ORANGE, TANAGRA and WEKA, can handle easily this process.