Subject

Implementing RANDOM FOREST with TANAGRA.

RANDOM FOREST is a combination of an ensemble method (BAGGING) and a particular decision tree algorithm ("Random Tree" into TANAGRA).

Dataset

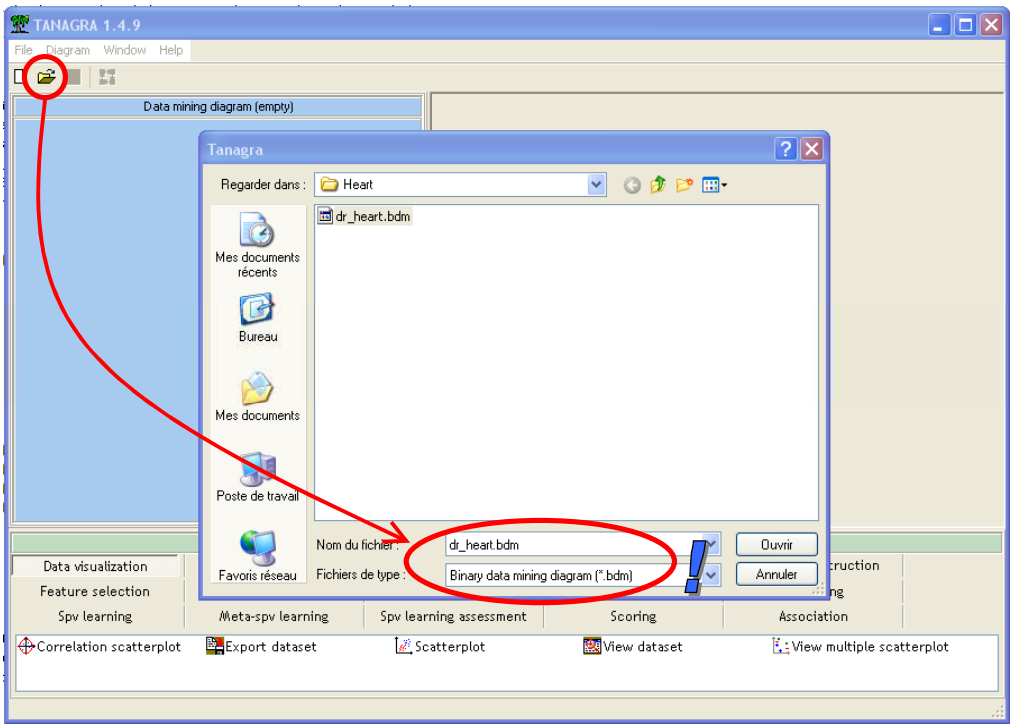
We use the HEART dataset from UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html).

We aim to predict a heart disease from various descriptors such as the age of the patient, etc. We have already used this dataset in other tutorials (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/dr_comparer_spv_learning.pdf).

RANDOM FOREST

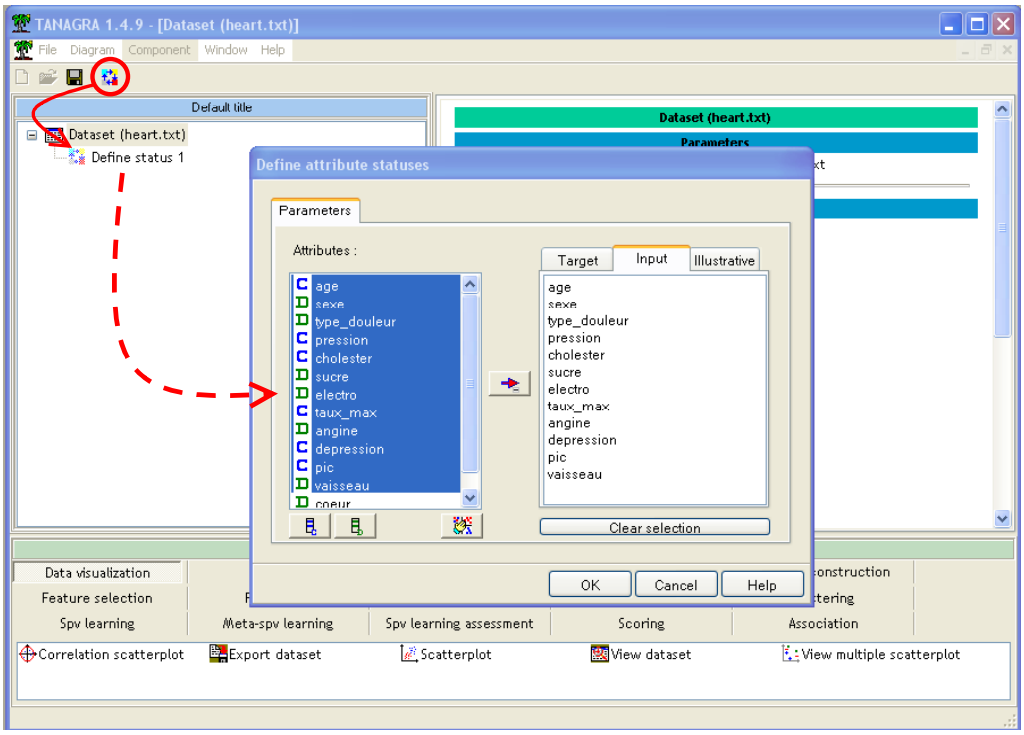
Create a diagram and load the dataset

We open (FILE/OPEN) the file DR_HEART.BDM. BDM is a binary format of TANAGRA.



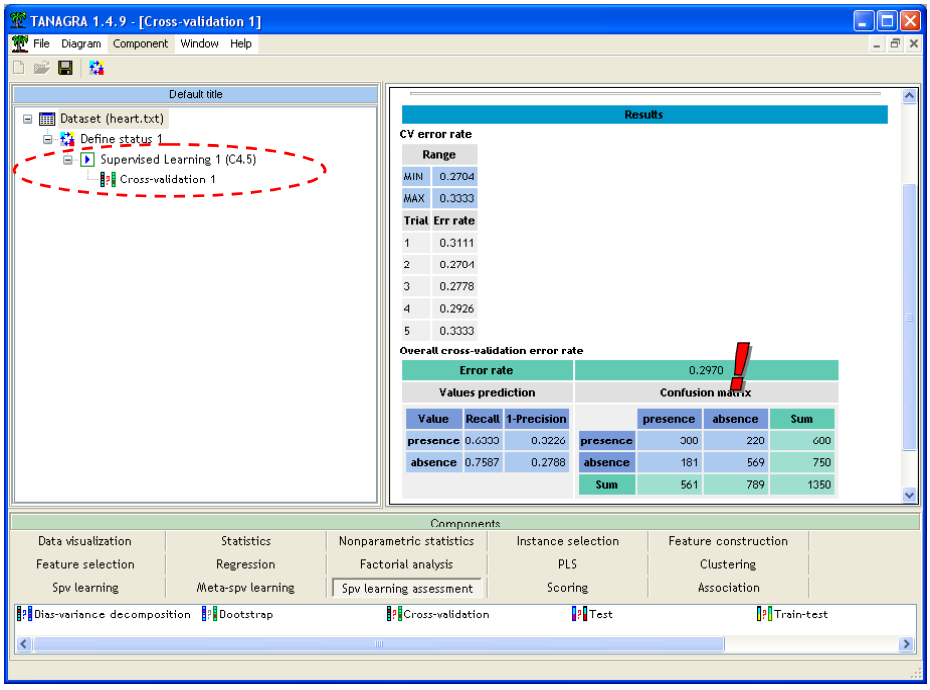
Define the task

In the next step, we use the DEFINE STATUS component in order to choose the target attribute (COEUR) and the input attributes (all the others).



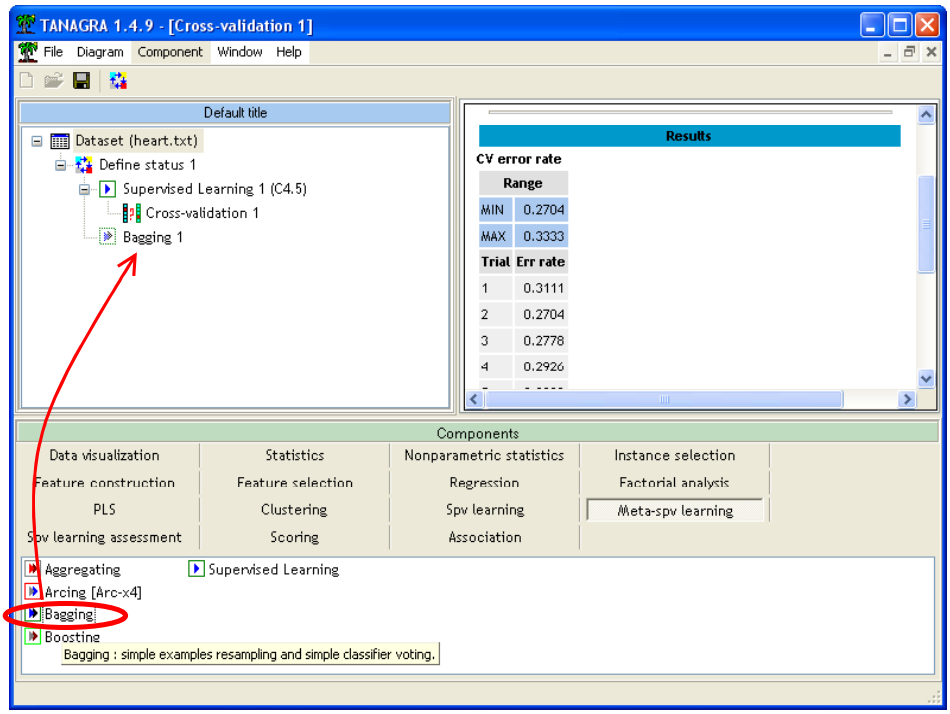
C4.5 decision tree algorithm

We use C4.5 and cross-validation in order to evaluate the accuracy of a standard (individual) decision tree algorithm. The error rate is 29.7%



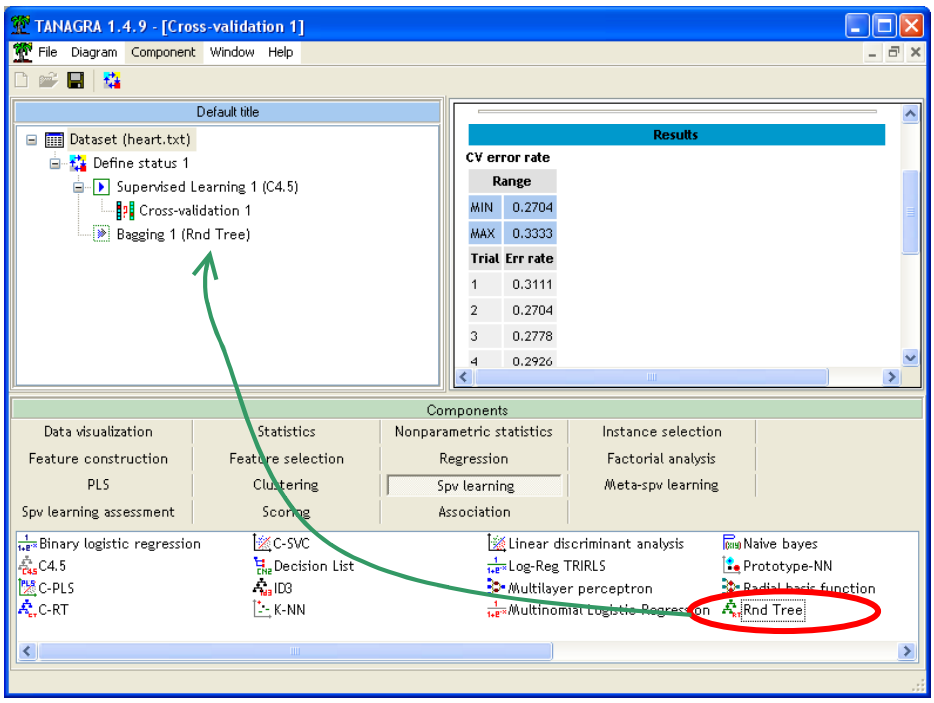
RANDOM FOREST

We want to implement the Random Forest algorithm now. There are two steps in order to insert the Random Forest method in the diagram: (1) Insert the BAGGING ensemble method (META-SPV LEARNING tab)

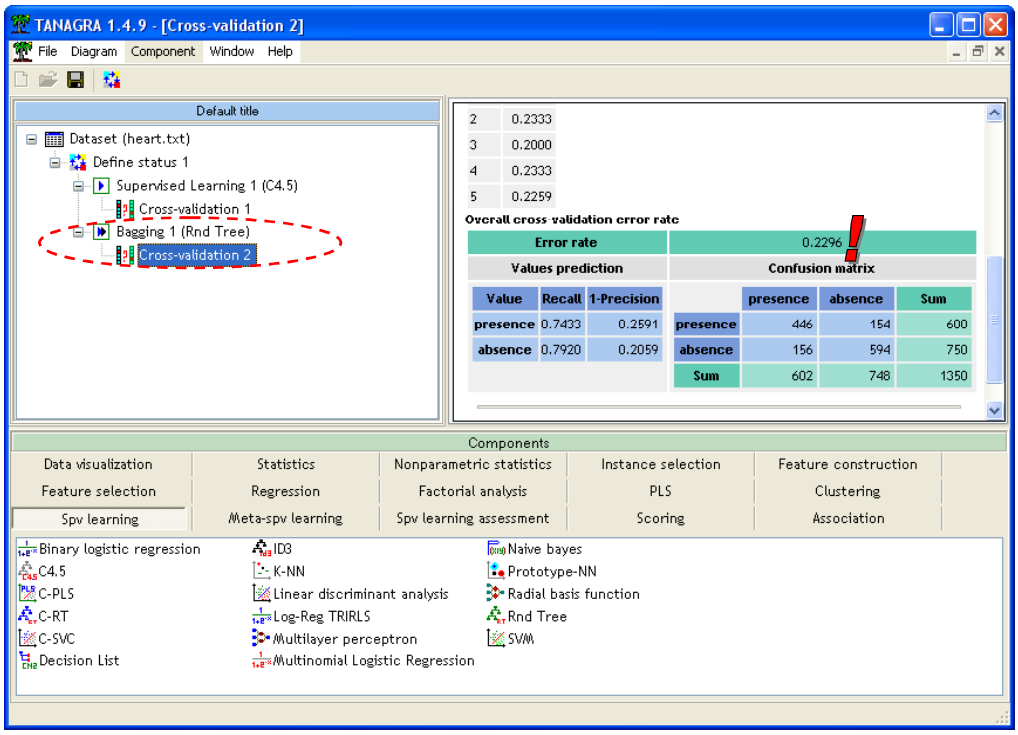


(2) Embed in this component the RANDOM TREE algorithm (SPV LEARNING tab).

It is obvious that using this learning algorithm without an aggregating framework often gives poor performances.



So, we evaluate the learning accuracy with a cross-validation component.



The reduction of the error rate is impressive, we obtain 22.9% now.