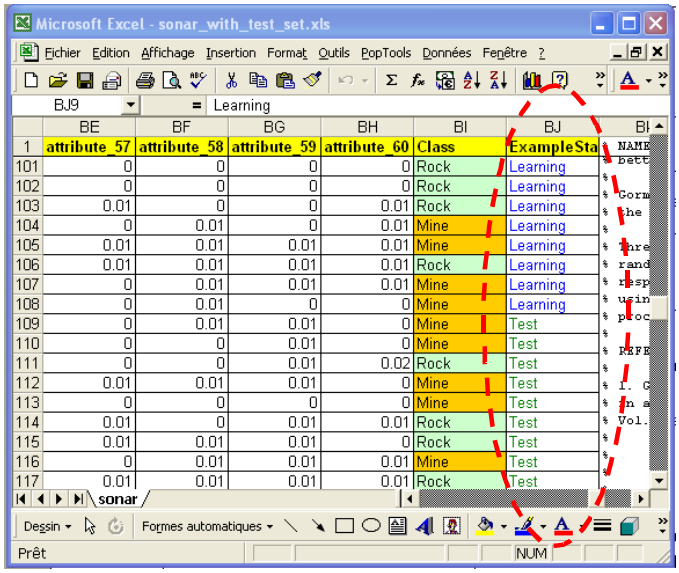
Subject

In order to evaluate a supervised learning algorithm, we often split the dataset into training set, which is used in the training process, and test set, which is used to obtain an unbiased error rate evaluation. There are sampling components in TANAGRA, which enable to subdivide randomly the dataset, but in some circumstances, the user want use a predefined test set for their comparisons.

Dataset

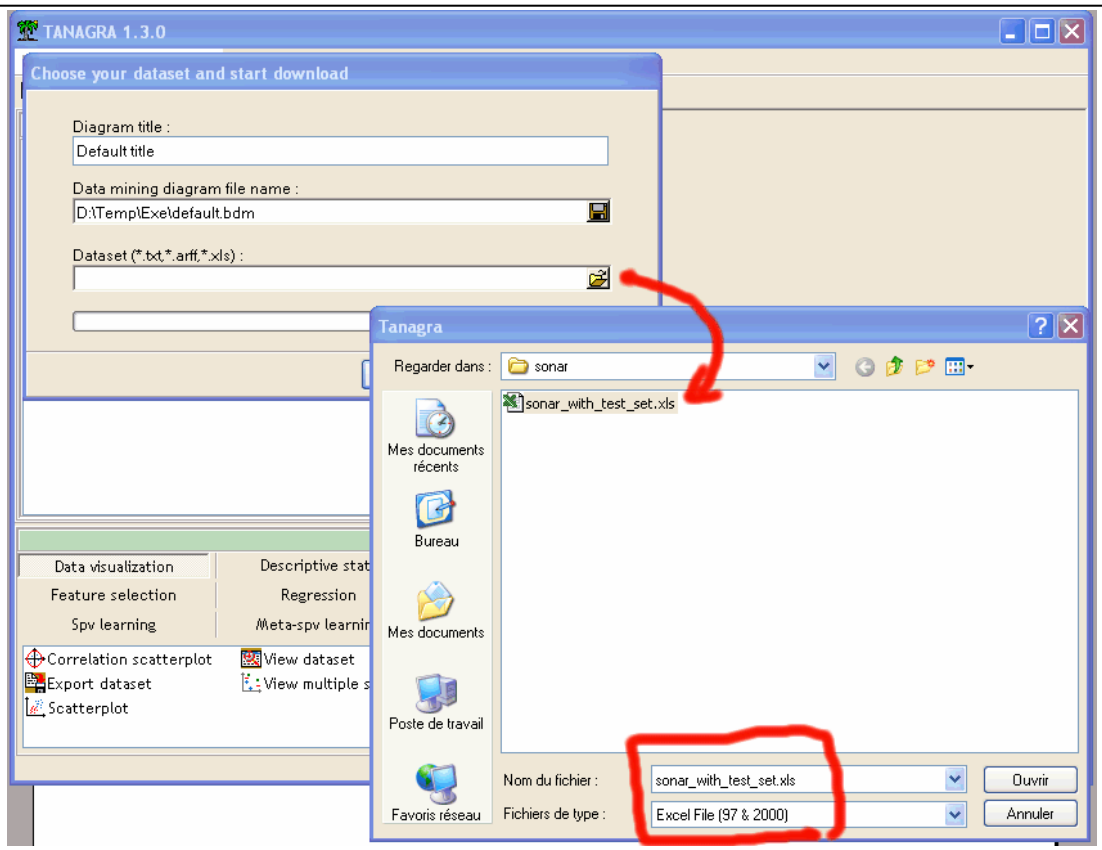
We use the SONAR dataset from the UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html -- **sonar_with_test_set.xls**). The dataset contains 208 examples with 60 descriptors (ATTRIBUTE_1 to ATTRIBUTE_60) and a binary class attribute (CLASS). Rather than to use two distinct files for the learning and the test set, we prefer join them together in a single file and create an additional column indicating the role that each observation must play (EXAMPLESTATUS).



Comparing learning methods

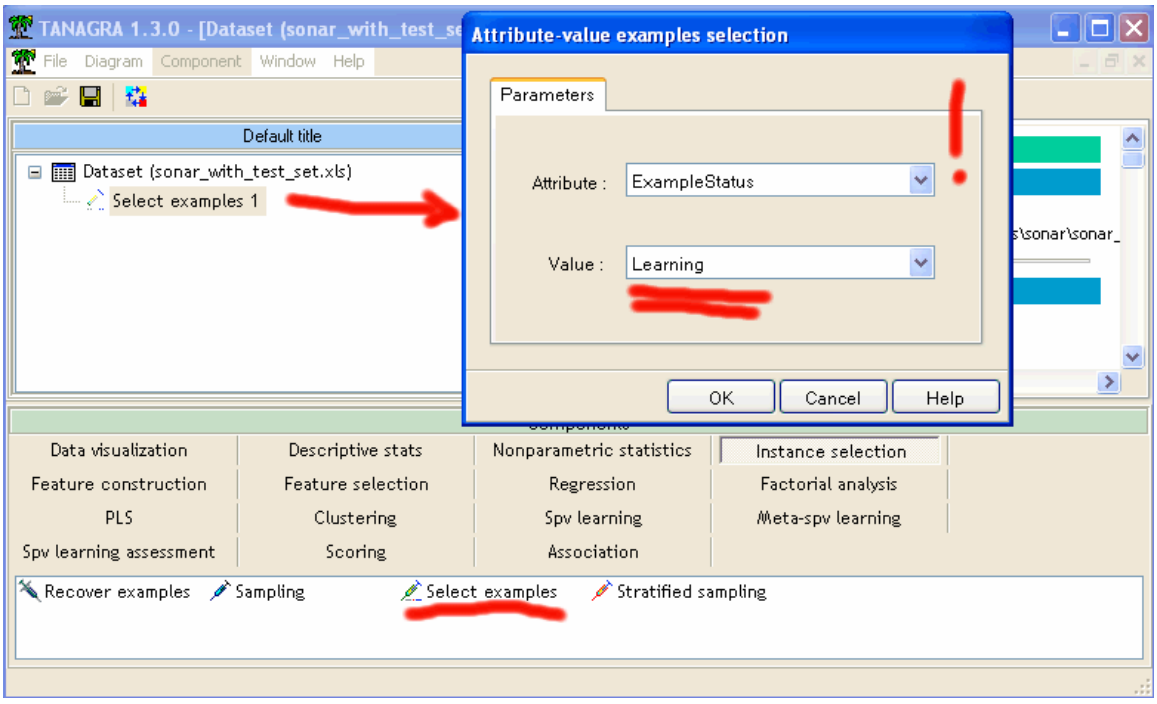
Download the dataset

First of all, we have to create a new stream diagram and select the dataset: SONAR_WITH_TEST_SET.XLS (FILE / NEW).

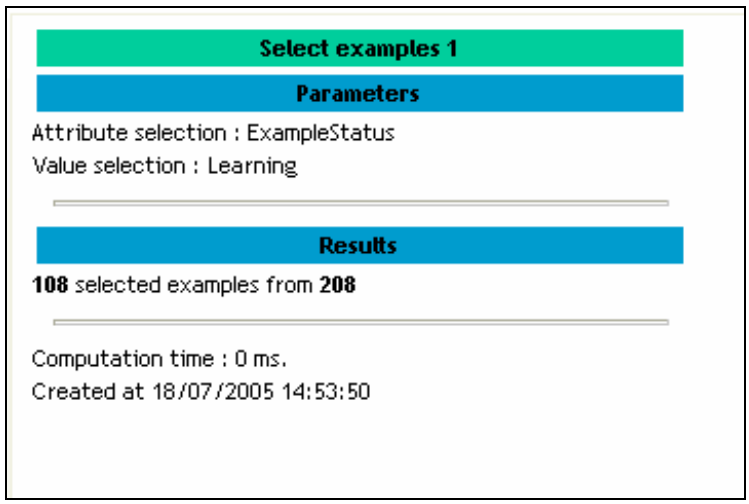


Select learning and test set

In the next step, we use the EXAMPLESTATUS column in order to define the learning and the test set. Add the SELECT EXAMPLES component in the diagram, and set the right parameters (Attribute: EXAMPLESTATUS, Value: LEARNING).

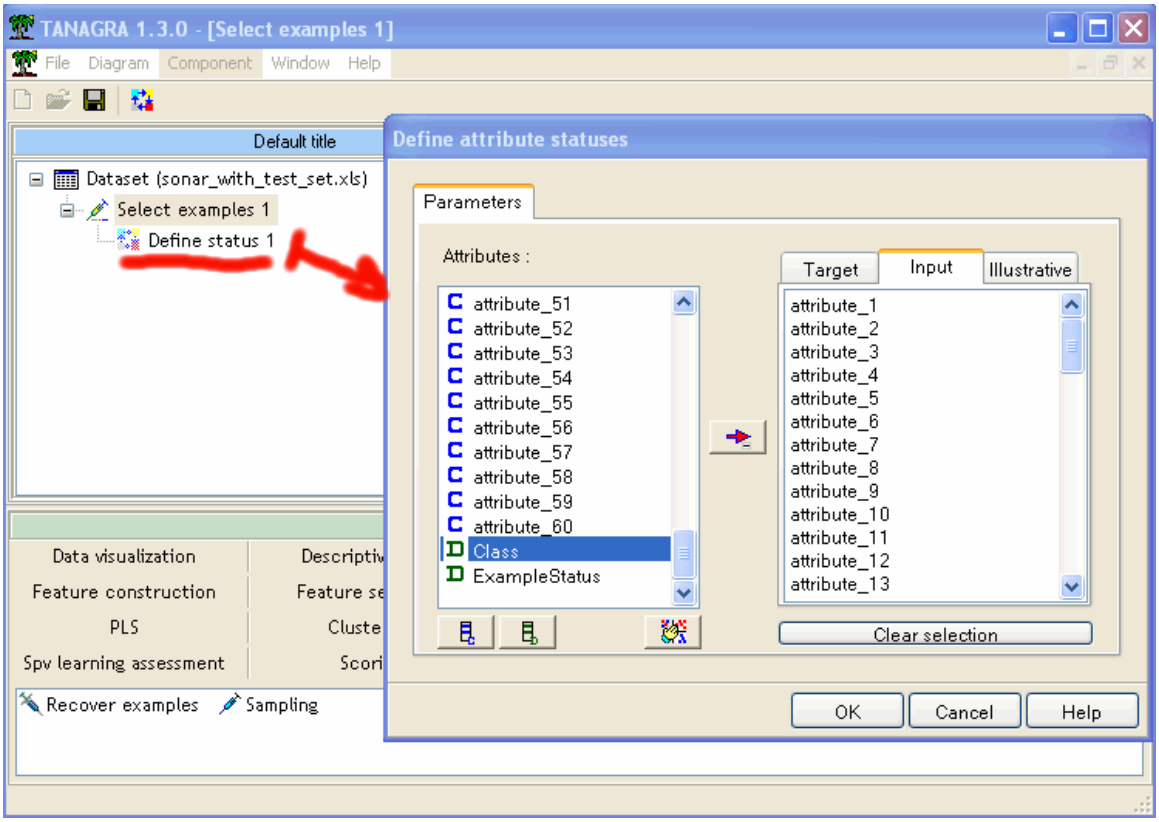


We see that the learning set size is 108 (thus, the test set size is 100).



SONAR prediction problem

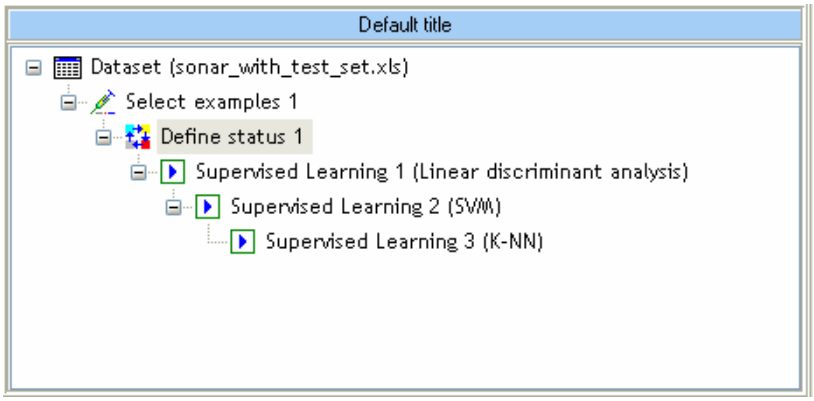
Add a DEFINE STATUS component in the stream diagram, set CLASS as TARGET, and the other descriptors as INPUT. Of course, we do not use the EXAMPLESTATUS column in the remaining.



Learning process

We want to compare three algorithms that have a very different representation and learning bias.

Add a linear discriminant analysis (LINEAR DISCRIMINANT ANALYSIS), a linear support vector machine (SVM) and a nearest neighbor algorithm (K-NN) in the diagram. Note that this kind of component must be embedded in meta-supervised component (META SPV LEARNING); we use a standard learning procedure (SUPERVISED LEARNING)[1].



The resubstitution error rates are:

- 2.78% for LDA;
- 12.04% for Linear SVM;
- 11.11% for K-NN (5-NN).

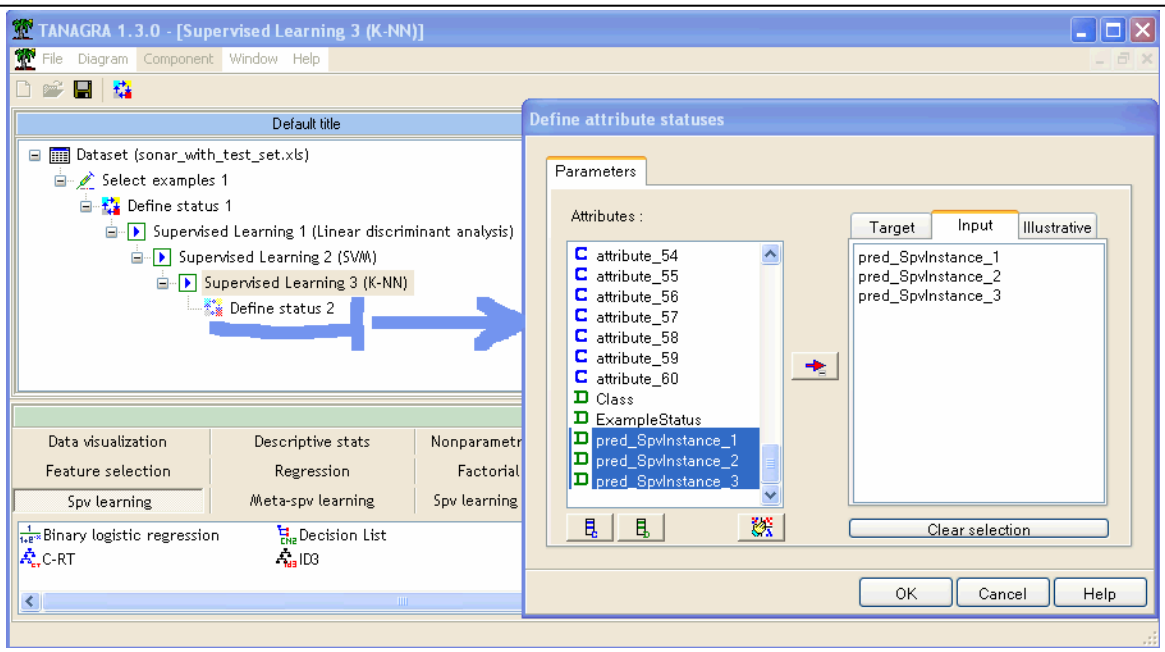
LDA seems the best learning algorithm on this dataset?

Test error rate

We must use the test set in order to obtain an unbiased error rate evaluation.

Add a DEFINE STATUS component in the diagram and set as TARGET the real class attribute (CLASS), the 3 projections of the methods are defined as INPUT.

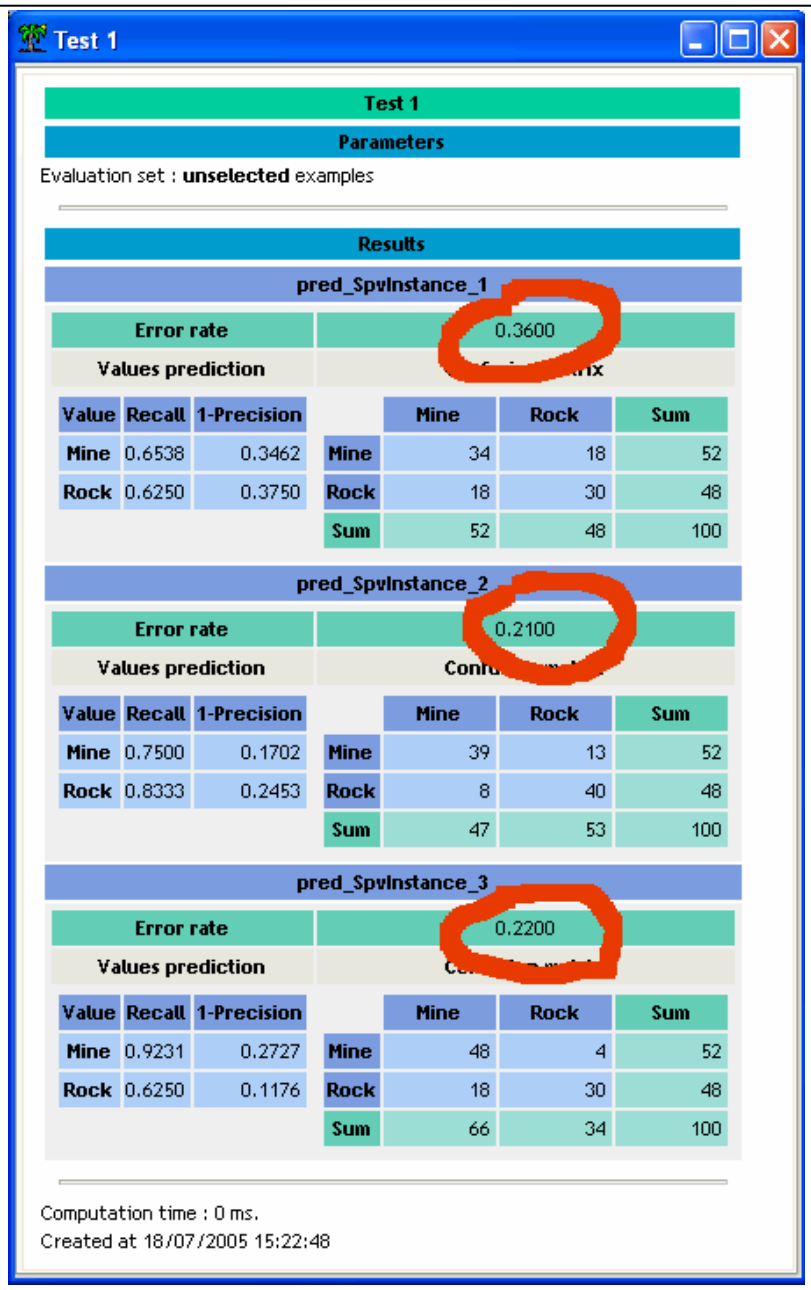
[1] In concrete terms, we operate in two steps: first, we put the SUPERVISED LEARNING component (META SPV LEARNING tab) into the diagram; then we embed in this one the learning algorithm component (e.g. LINEAR DISCRIMINANT ANALYSIS) from the SPV LEARNING tab.



We add the TEST component in the diagram, the default parameter enables to compute the error rate on the unselected examples, that is the test set.[2].

The results illustrate the overfitting phenomenon, which is foreseeable if we consider the characteristics of our dataset: many descriptors (60) compared the learning set size (108 examples).

[2] It is possible to modify the parameter so that the error rate is computed on the selected examples (the learning set). In this case, we obtain again the resubstitution error rate.



Test error rate are:

- 36% for LDA (SUPERVISED LEARNING 1);
- 21% for SVM (SUPERVISED LEARNING 2);
- 22% for K-NN (SUPERVISED LEARNING 3).

It appears that the LDA particularly suffered on the SONAR problem, the SVM and K-NN present on the other hand comparable performances.