Subject

In this tutorial, we try to build a roc curve from a logistic regression.

Regardless the software we used, even for commercial software, we have to prepare the following steps when we want build a ROC curve.

- Import the dataset in the soft;
- Compute descriptive statistics;
- Select target and input attributes;
- Select the "positive" value of the target attribute;
- Split the dataset into learning (e.g. 70%) and test set (30%);
- Choose the learning algorithm. Be careful, the softwares can have different implementation and present a slightly different results;
- Build the prediction model on the learning set and visualize the results;
- Build the ROC curve on the test set.

According the softwares, the progression can be different but it is clear that we must, explicitly or not, process these steps.

Dataset

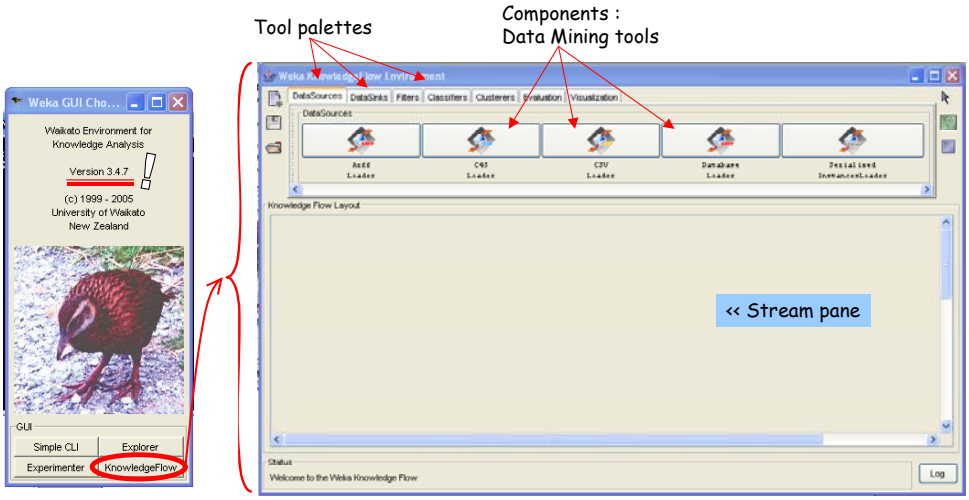
We use the dataset from the Komarek's website which implement the LR-TRIRLS logistic regression library (http://komarix.org/ac/lr). The DS1-10 dataset contains 26733 examples, 10 continuous descriptors; the frequency of the positive value of the target attribute is 5%. We use ARFF file format for WEKA and TANAGRA, TXT for ORANGE (TANAGRA can also handle TXT file format).

Building a ROC curve with WEKA

The number of methods is impressive in WEKA, but it is also the main weakness of this software, a through initiation is necessary. The needed components for the construction of a roc curve are not obvious.

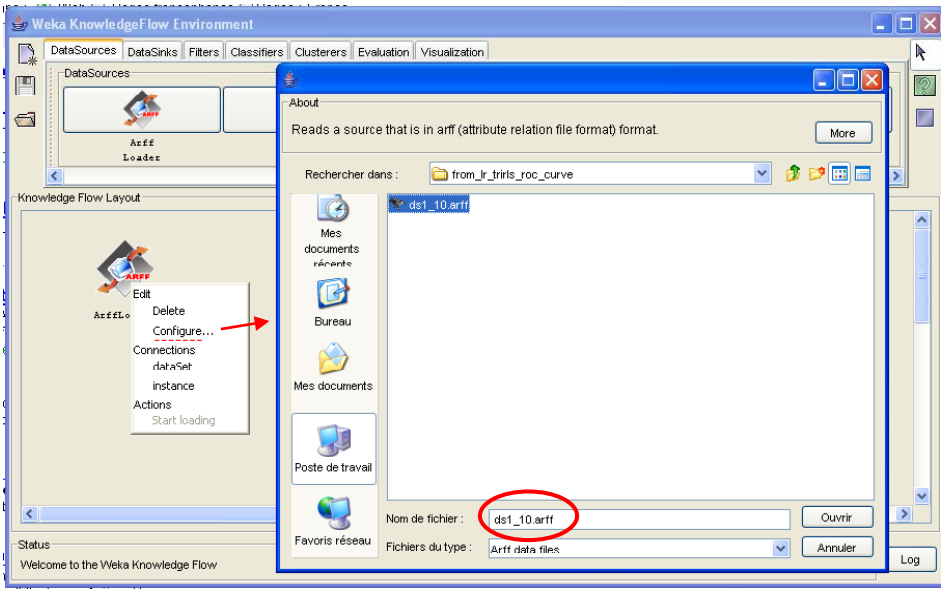
Execution of WEKA

When we execute WEKA, a dialog box enables to choose the execution mode. We select the KNOWLEDGE FLOW option. We have used the 3.4.7 version in this tutorial; the results can be slightly different on the others versions. The organization of the software is classic. In the top of the window, we find the tools, machine learning components, in some palettes. The KNOWLEDGE FLOW layout allows us to define the succession of data treatments.



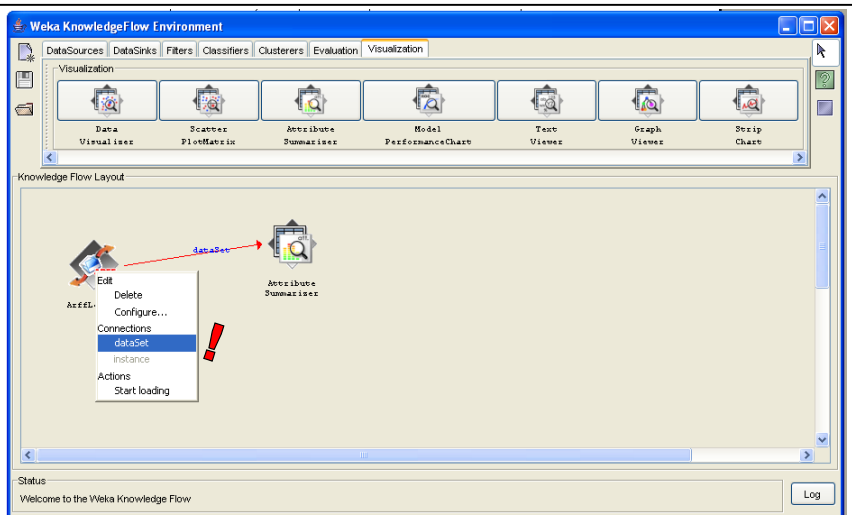
Load the dataset

The ARFF LOADER component (DATASOURCES palette) enables to load a dataset. We add it in the workspace, we can select the file with the CONFIGURE menu of the component.

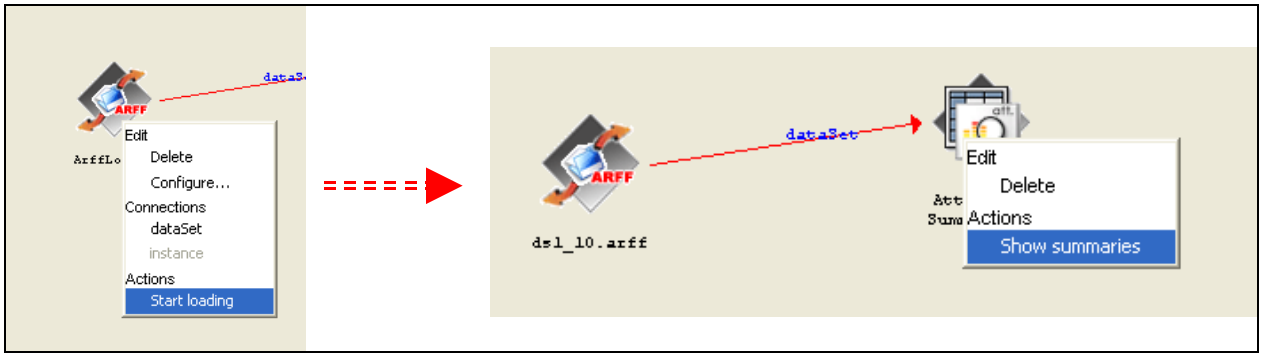


Descriptive statistics

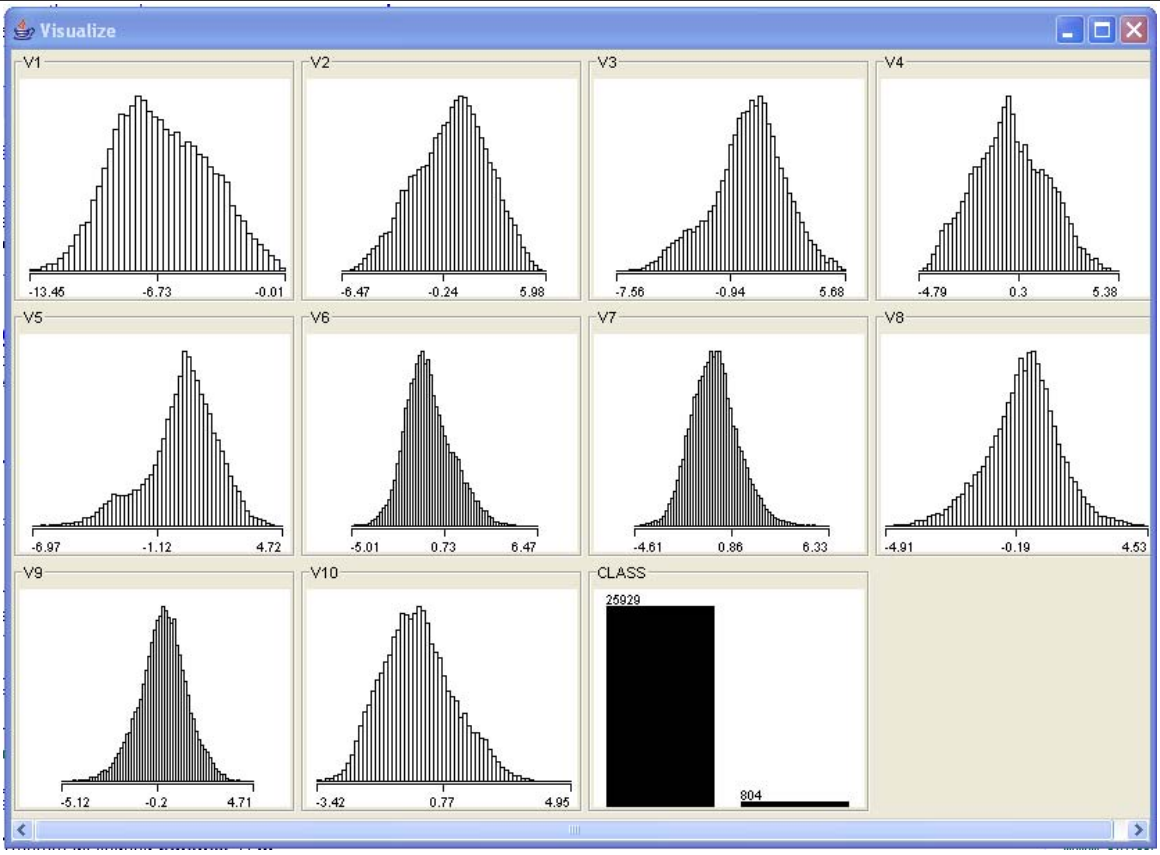
The ATTRIBUT SUMMARIZER component (VISUALIZATION) shows the data distribution; we can quickly detect abnormalities in the dataset. We add this component and we to connect ARFF LOADER to this new component (DATASET connection).



The treatment will be executed when we select the START LOADING menu of the ARFF LOADER component. To see the results, we select the SHOW SUMMARIES option of ATTRIBUTES SUMMARIZER.

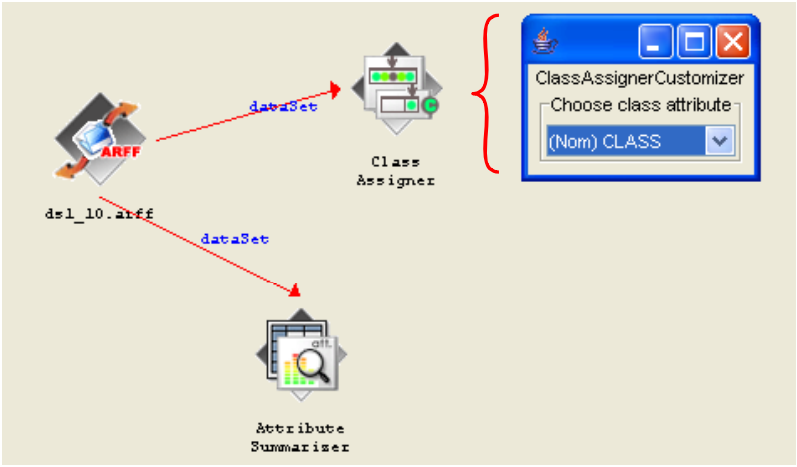


We see that there are not abnormalities on the descriptors; we see also that we have imbalanced dataset (804 "positive" vs. 25929 "negative").

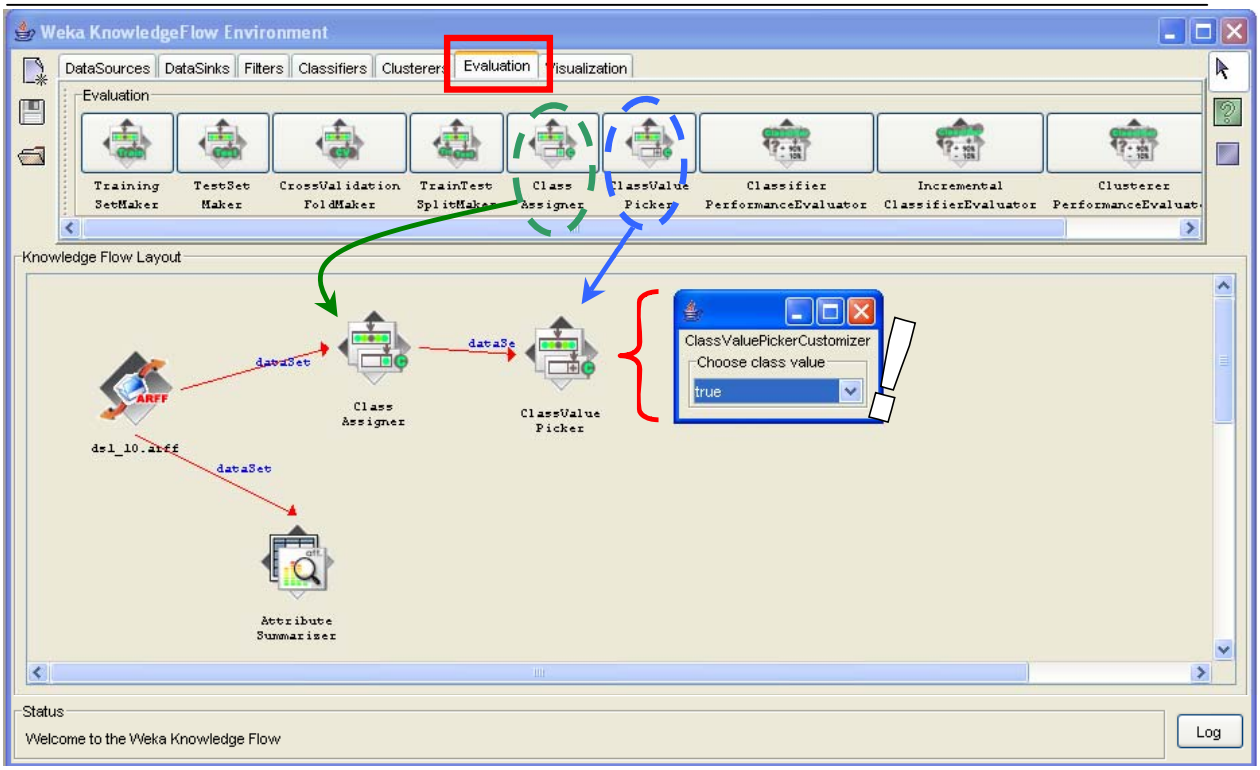


Define target and input attributes

The last column is the default class attribute for WEKA, but we can also explicitly select the column of the class attribute. In this case, we use the CLASS ASSIGNER component (EVALUATION palette). We add this component in our diagram; we connect ARFF LOADER to this new component (DATASET connection). We select the CONFIGURE menu in order to select the right class attribute.

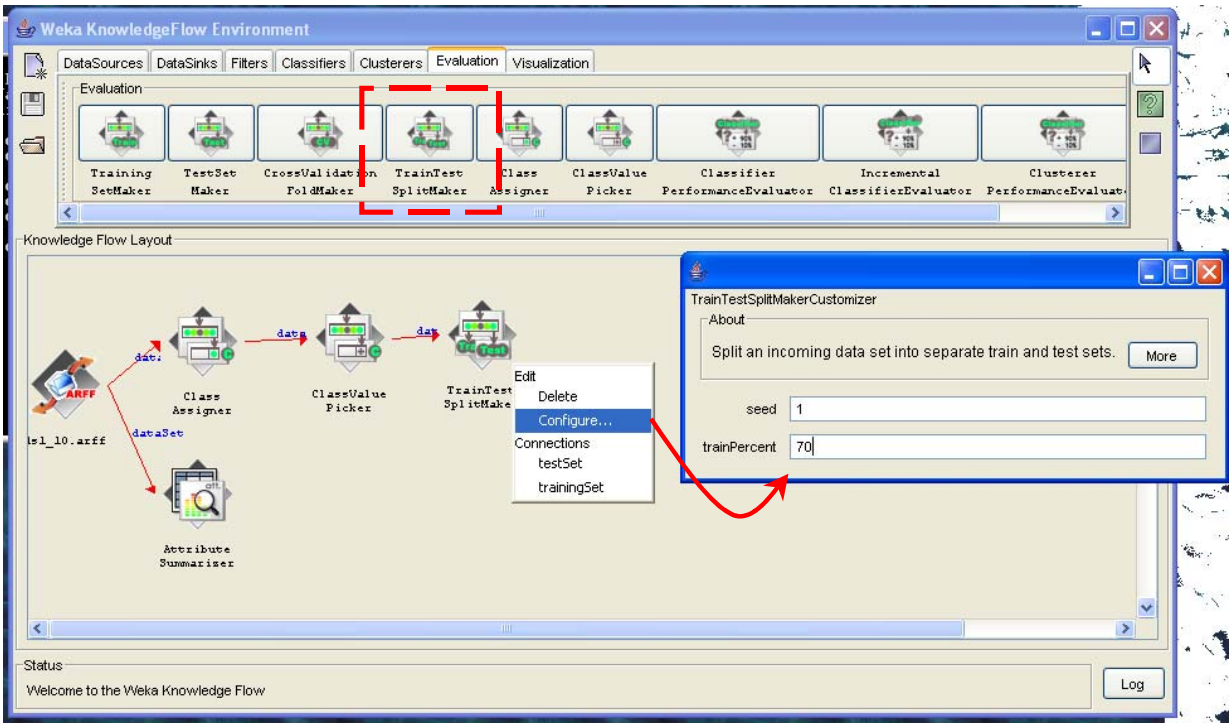


In the following step, we must specify the "positive" value of the class attribute. We use the CLASSVALUE PICKER (EVALUATION palette) component and select the "TRUE" value in the parameter dialog box.



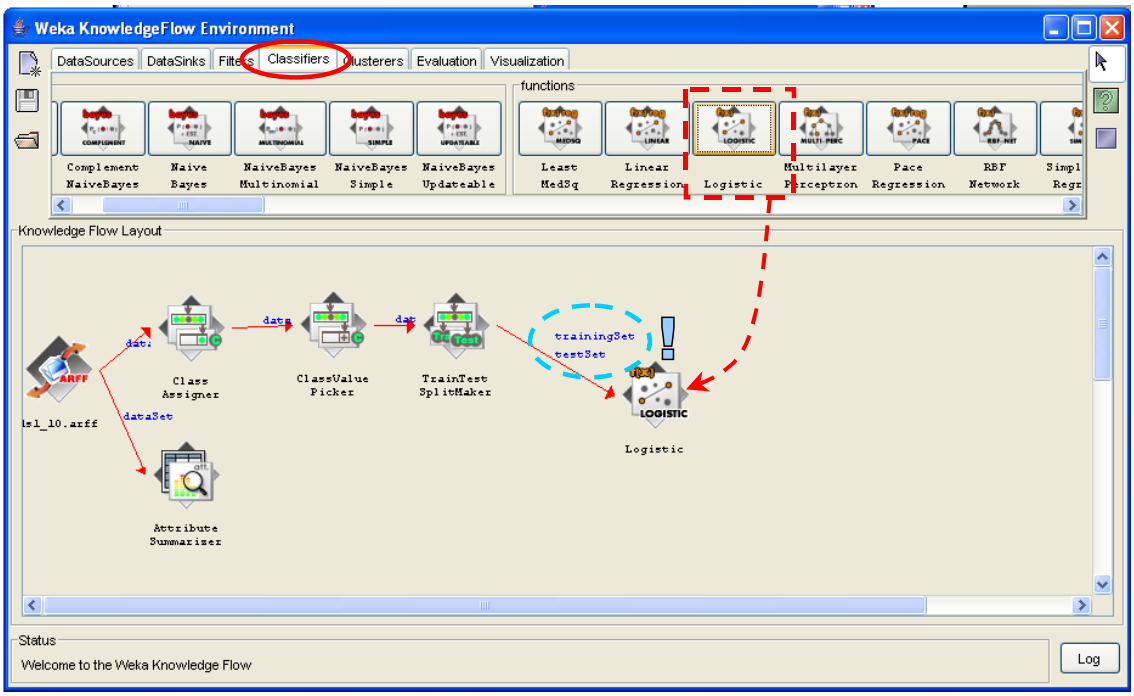
Subdivision of the dataset into "learning" and "test" set

We want to build our prediction model on the 70% of the whole dataset, and compute the ROC curve on the remaining. So, we set the TRAINTEST SPLIT MAKER (EVALUATION) in the diagram and configure its parameters.



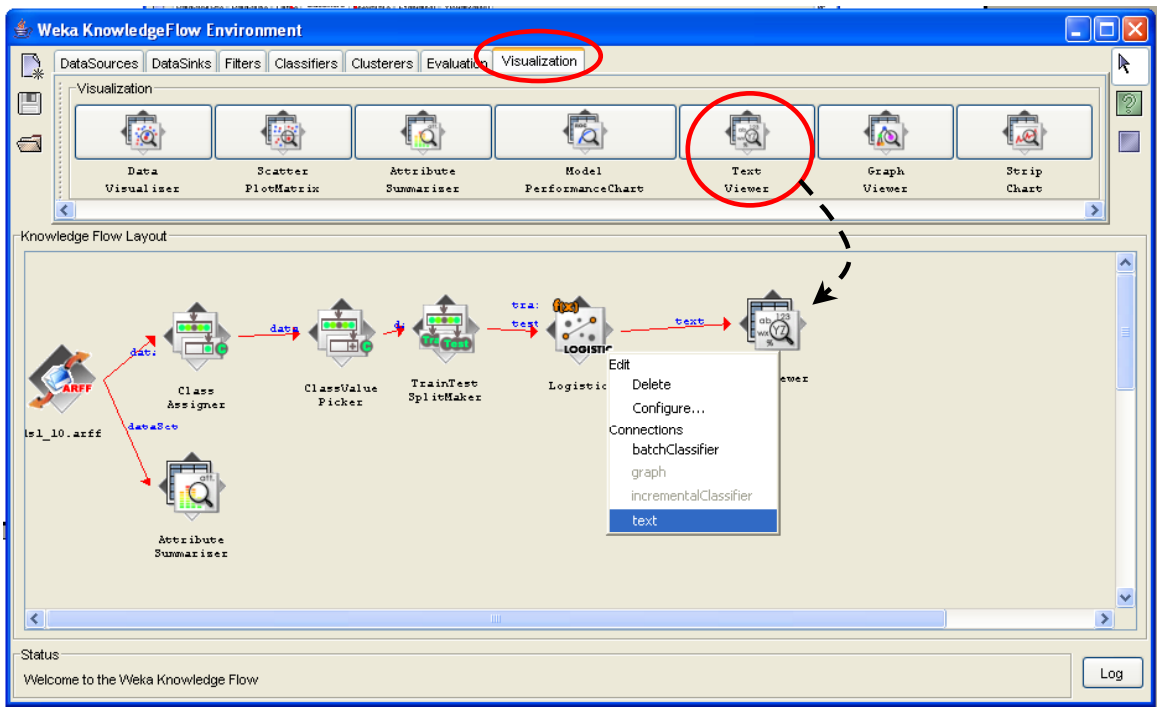
Logistic regression

The logistic regression component is in the CLASSIFIERS palette. We set it in the diagram, we connect **twice** the TRAIN TEST SPLIT MAKER to this new component: twice because we must use together the training and the test set which are produced by the same component.

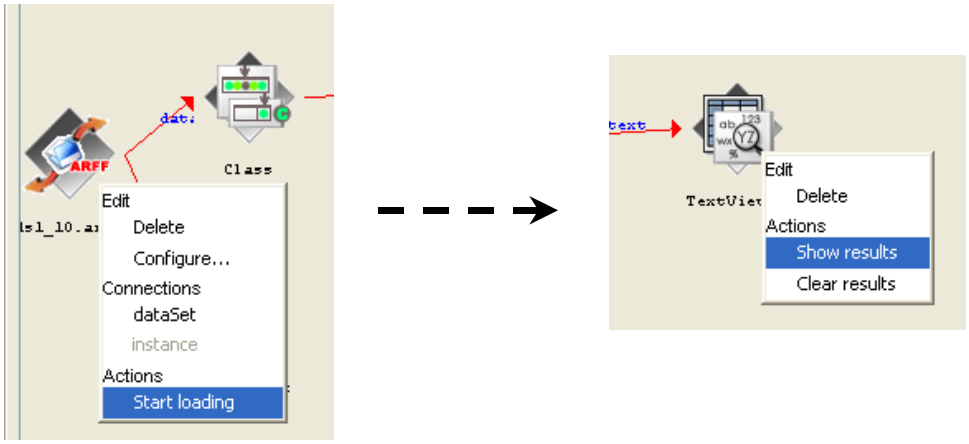


The results

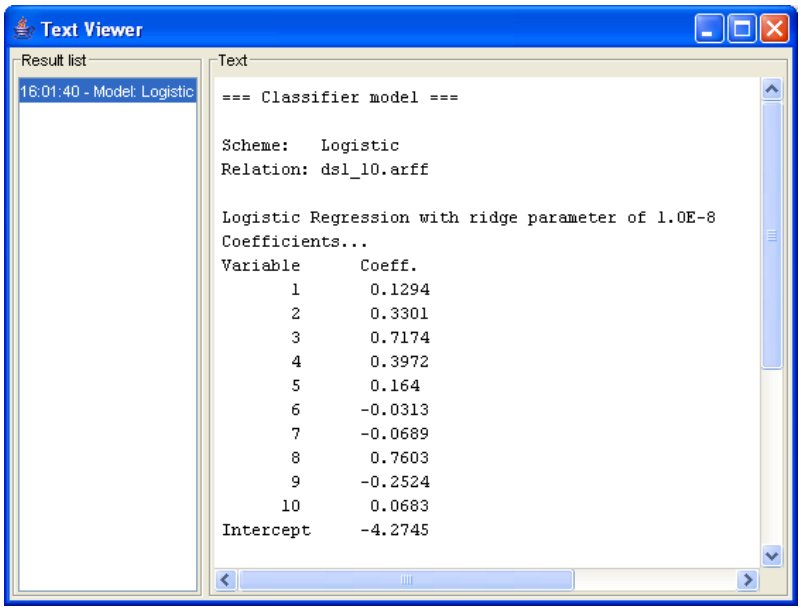
To see the results of the regression, we connect the LOGISTIC component to TEXT VIEWER (VISUALIZATION palette) that we set in the diagram.



We execute again the diagram (START LOADING of ARFF LOADER component). The SHOW RESULTS of TEXT VIEWER opens a new window with the results of the learning process.

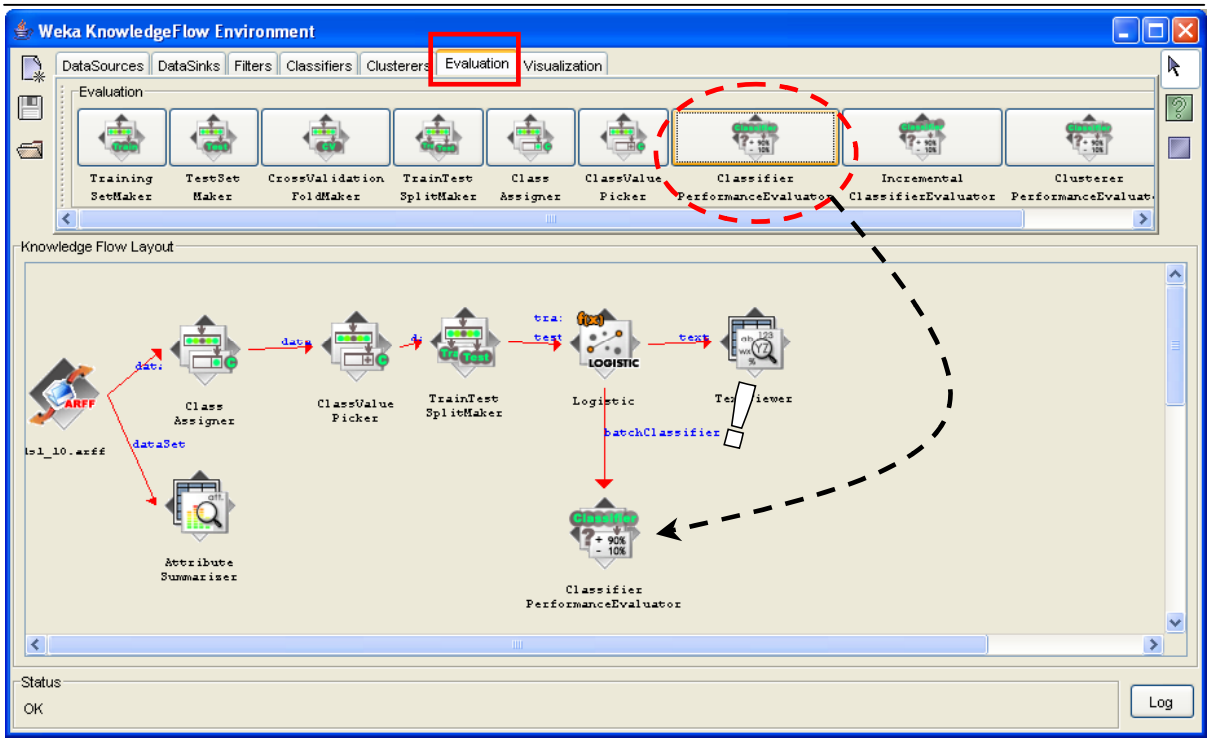


We obtain the regression coefficients.

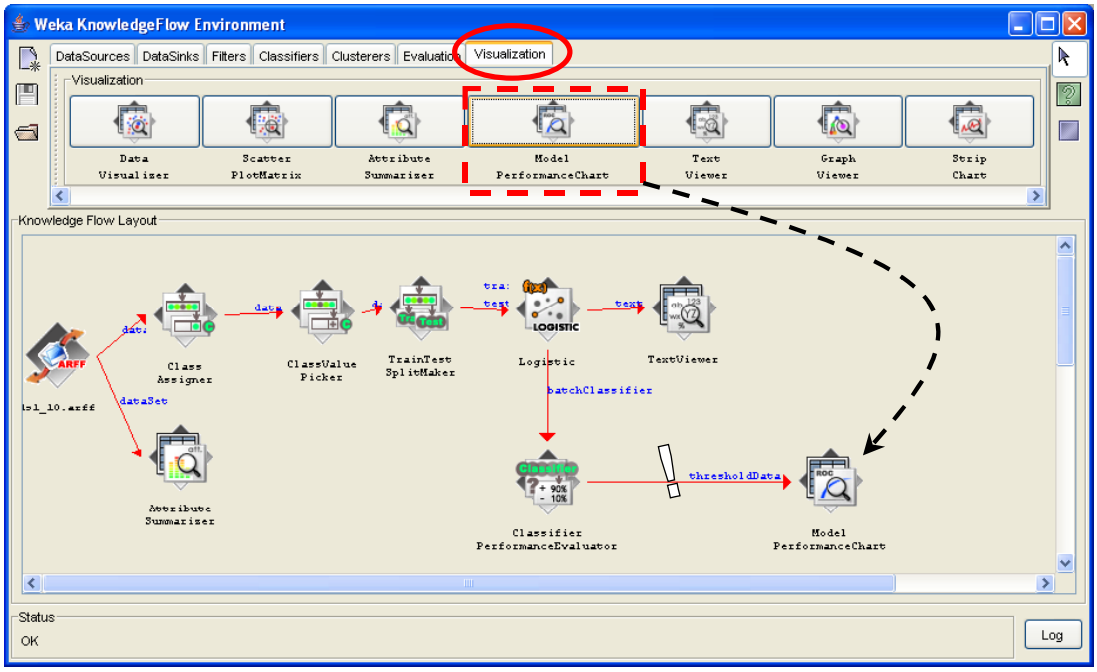


ROC curve

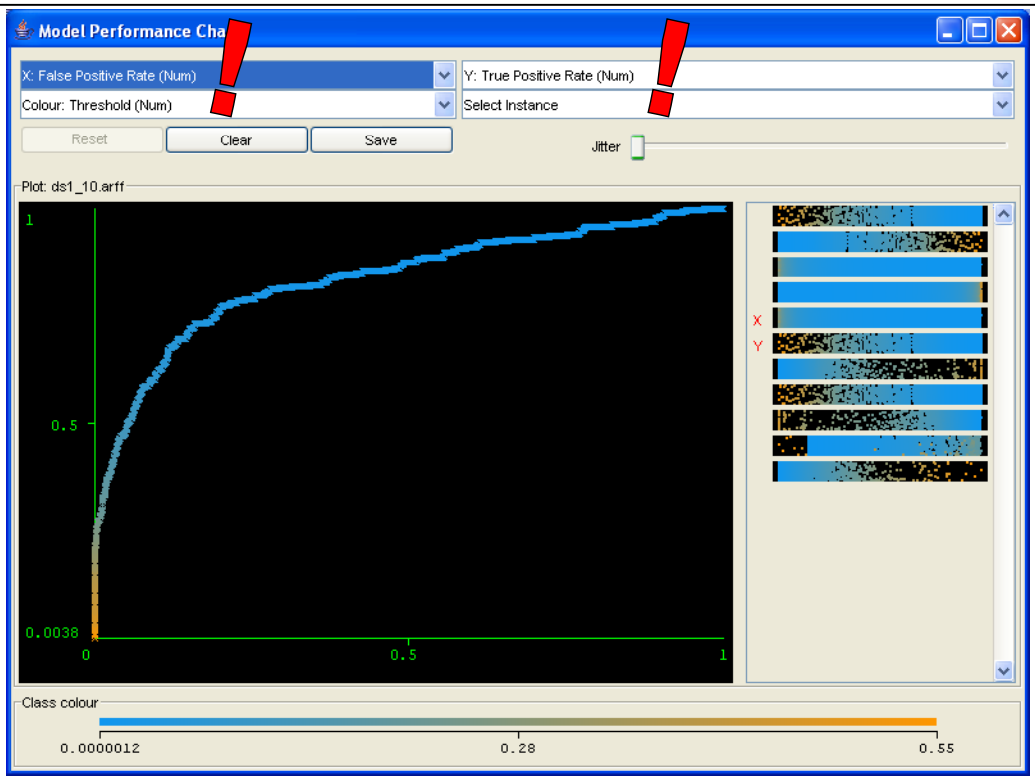
In order to evaluate the learning process, we add the CLASSIFIER PERFORMANCE EVALUATOR component (EVALUATION). We connect the BATCH CLASSIFIER output of LOGISTIC to this new component.



We set MODEL PERFORMANCE CHART (VISUALIZATION) in the diagram; we use the THRESOLD DATA (!) output of the CLASSIFIER PERFORMANCE EVALUATOR when we connect the two components.



We run again the diagram with the START LOADING menu of ARFF LOADER. We select the SHOW PLOT menu of the last component (MODEL PERFORMANCE CHART). We obtain the ROC curve.


Model Performance Cha
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Save
Jitter
Plot: ds1_10.arff
1
0.5
0.0038
0
0.5
1
X
Y
Class colour
0.0000012
0.28
0.55