Subject

Coding categorical predictive attributes for logistic regression.

When we want to use predictive categorical attributes in a logistic regression or a linear discriminant analysis, we must recode them. The most used strategy is certainly dummy variables. The coding scheme is the following: we create a dummy variable for each category of the original attribute. If there are K categories, we build (K-1) dummy variables; the last category is deduced from the other variables.

In this tutorial, we display how to use the 0_1_BINARIZE component in order to transform categorical predictive attributes for logistic regression.

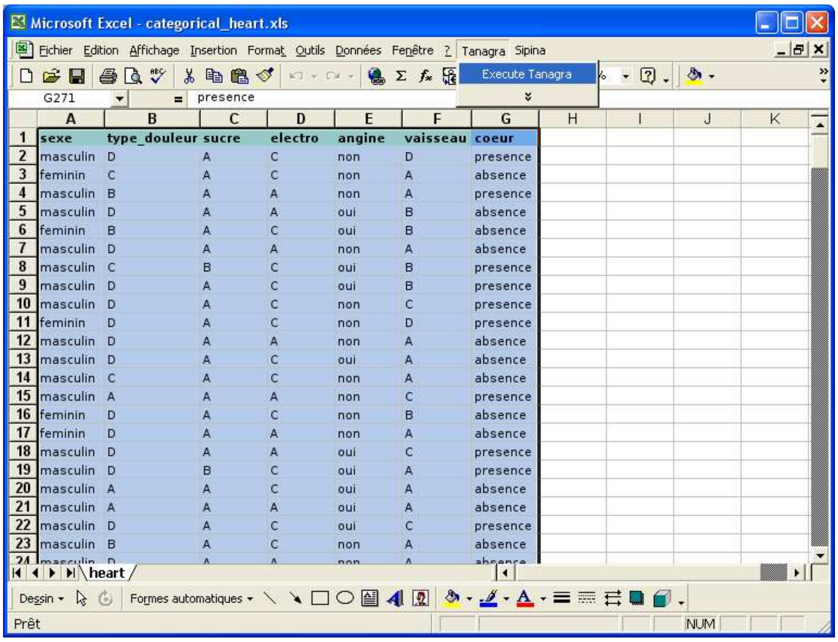
Dataset

We use CATEGORICAL_HEART.XLS in this tutorial. We want to predict the values of COEUR based on several predictive attributes (SEXE,…,VAISSEAU). The dataset contains 270 examples.

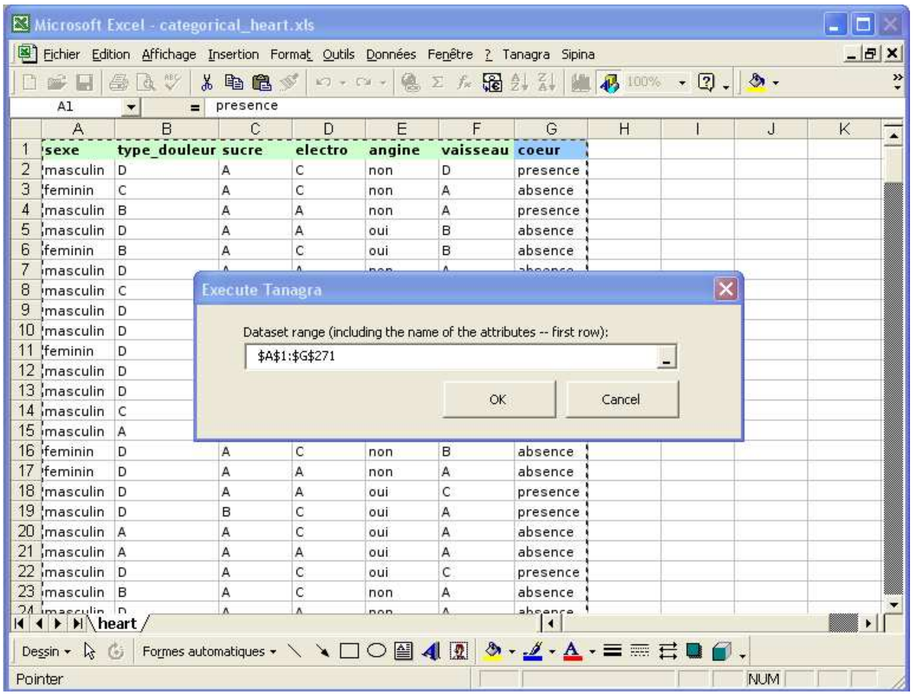
Dummy variables for categorical predictive attributes

Diagram creation

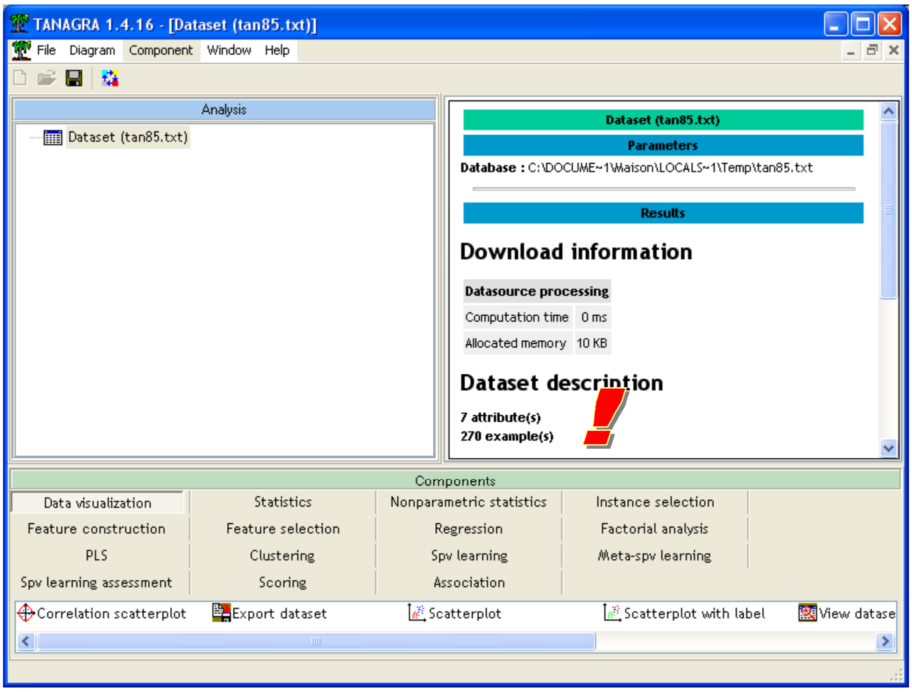
The simplest way to create a diagram is to open the dataset in the EXCEL spreadsheet. If you have installed the TANAGRA.XLA add-in, a new menu is now available (this add-in is come from 1.4.11 version). We select the range of cells and activate the TANAGRA / EXECUTE TANAGRA menu.



A dialog box appears. We check that the range selection is right and we validate.



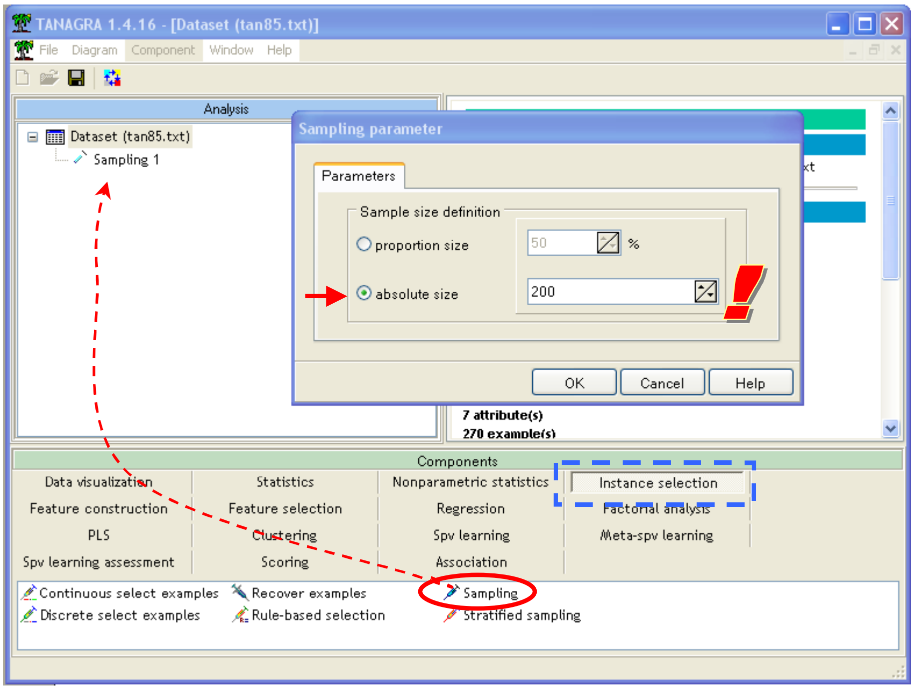
TANAGRA is automatically executed. We check that 270 observations and 7 variables are available.



Dividing the dataset into learning and test set

In order to obtain an honest error rate estimate, we must evaluate the classifier on a test set i.e. a dataset that is not used in the learning phase. So we divide the dataset in two parts using a random sampling: 200 examples for the training phase, and 70 examples for the test phase.

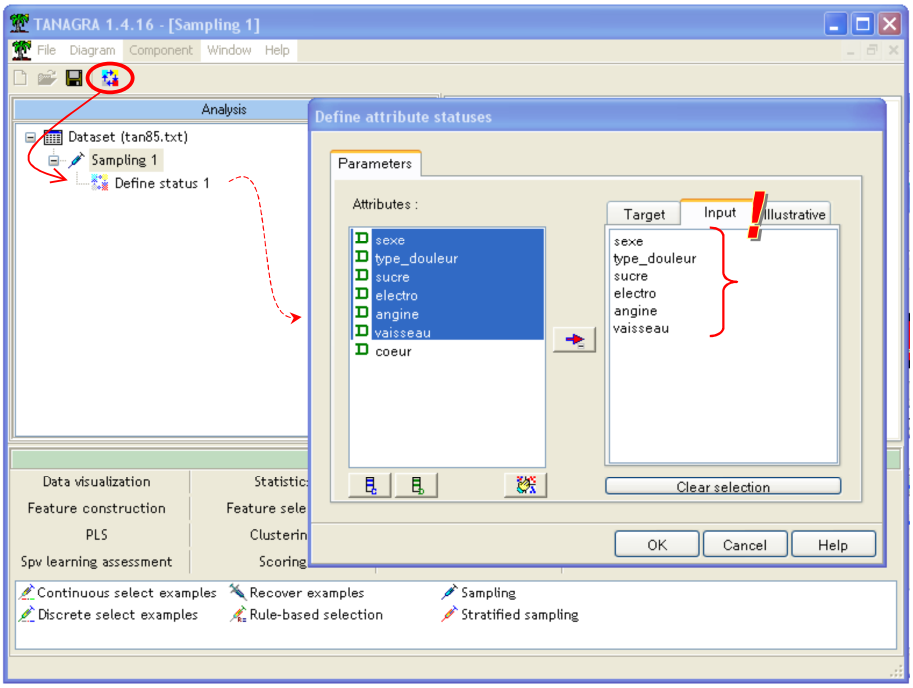
We insert the SAMPLING component (INSTANCE SELECTION tab) into the diagram. We activate the PARAMETERS menu and we select 200 examples.



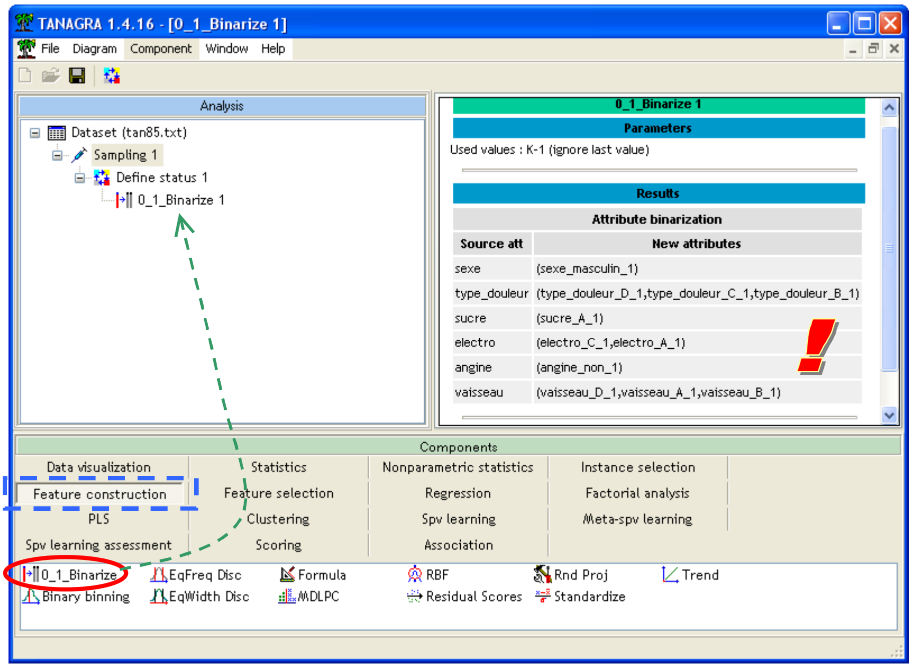
Coding categorical attributes

The 0_1_BINARIZE component enables to transform categorical attributes into dummy (binary) variables. By default, (K-1) attributes are created from a categorical variable with K values. But we can modify the parameters settings for some statistical methods.

We insert DEFINE STATUS component into the diagram, the simplest way is to use the shortcut in the toolbar. We set as INPUT the 6 categorical predictive variables.



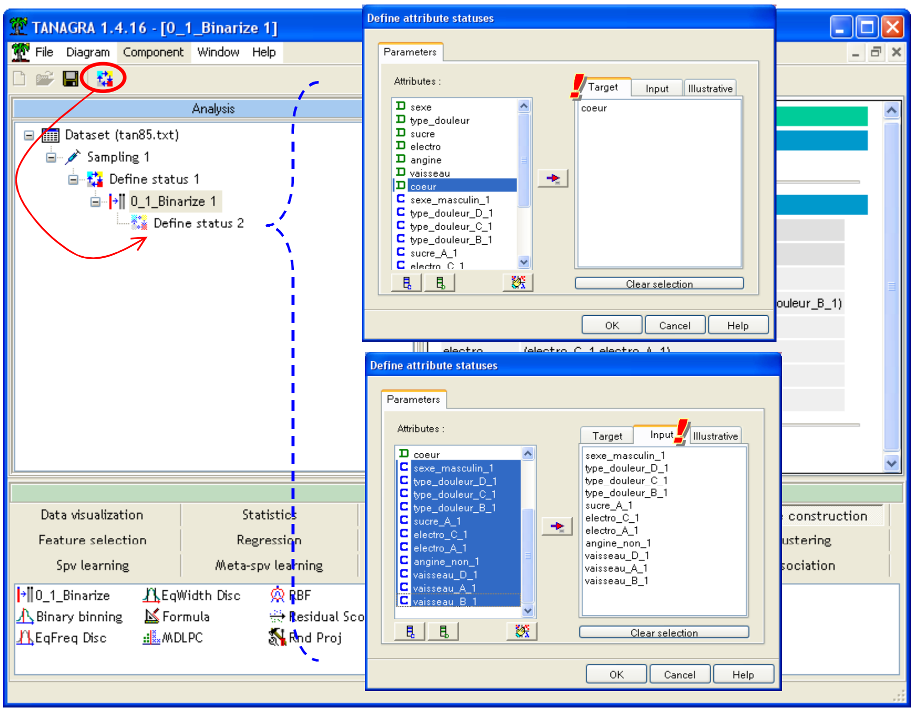
We add the 0_1_BINARIZE component into the diagram; we activate the VIEW menu in order to visualize the results.



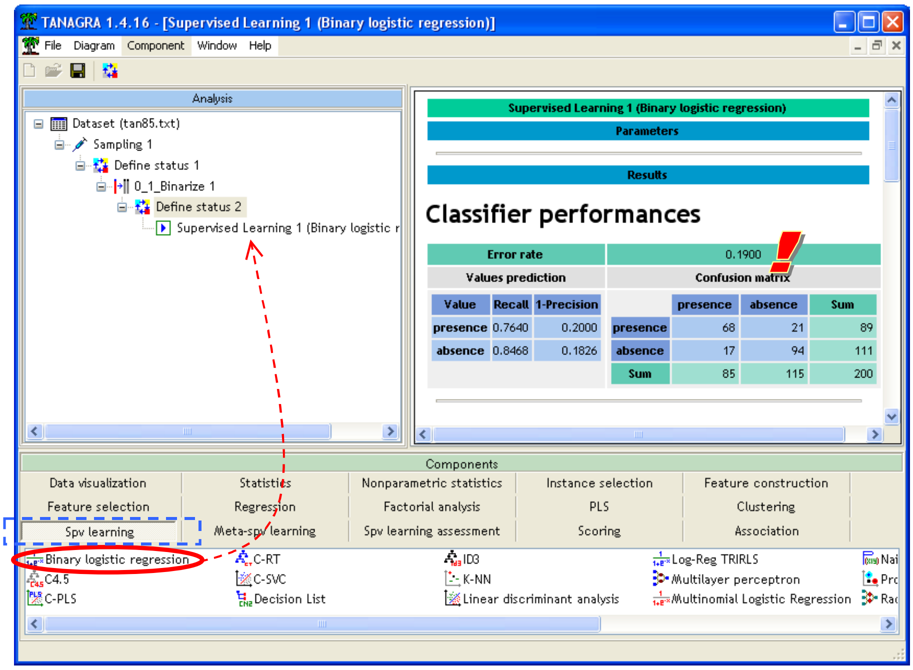
For each categorical attribute, a set of dummy variables is created.

Logistic regression

These dummy variables can be used in a logistic regression now. We add again a DEFINE STATUS component into the diagram. We set as INPUT these binary variables, and we set as TARGET the class attribute (COEUR).



Then we add the BINARY LOGISTIC REGRESSION (SPV LEARNING tab) into the diagram. We activate the contextual VIEW menu.



The resubstitution error rate is 19.0%. For a 0.05 significance level, only 4 binary attributes are significant. We use a very rough variable selection strategy. We simply remove all the variables that are not significant. Perhaps, more sophisticated approaches give better results but the goal is to display the utilization of the variable transformation component in this tutorial.

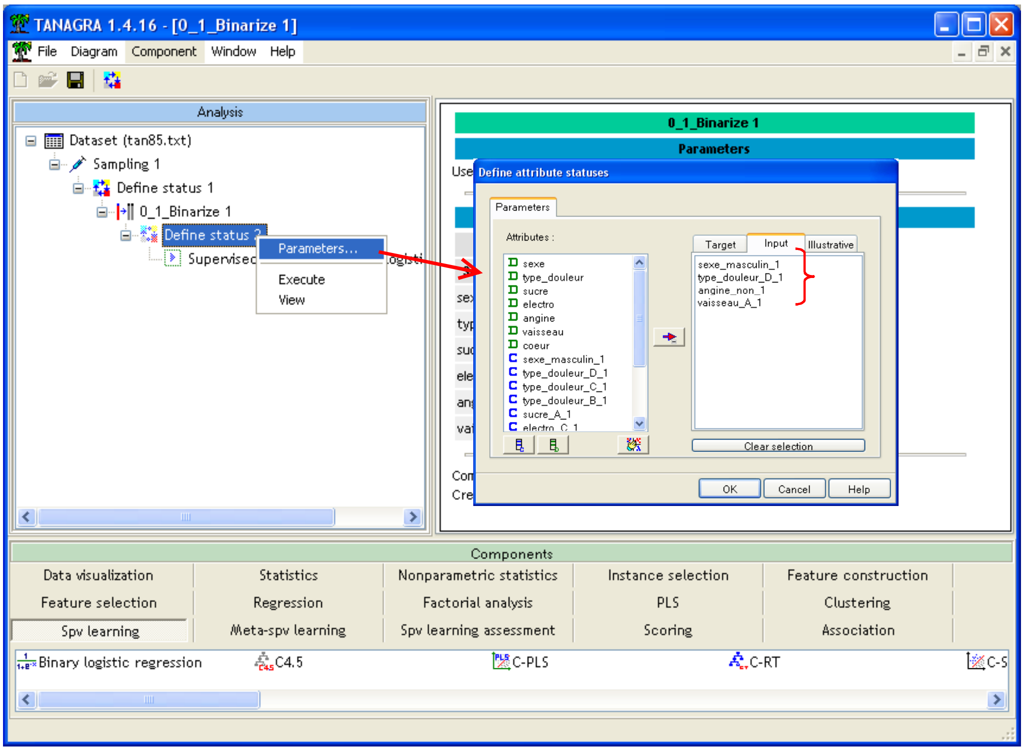
Adjustement quality

Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	159.4172
Chi-2	115.4167
P(>Chi-2)	0.0000

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	1.216147	-	-	-
sexe_masculin_1	1.812569	0.4704	14.8506	0.0001
type_douleur_D_1	2.452703	0.8692	7.9628	0.0048
type_douleur_C_1	1.057939	0.8852	1.4283	0.2320
type_douleur_B_1	0.663332	0.9715	0.4662	0.4947
sucre_A_1	0.384675	0.5761	0.4459	0.5043
electro_C_1	-1.776383	1.8446	0.9274	0.3355
electro_A_1	-2.006760	1.8413	1.1878	0.2758
angine_non_1	-1.389320	0.4613	9.0719	0.0026
vaisseau_D_1	-0.174881	1.1084	0.0249	0.8746
vaisseau_A_1	-2.661483	0.7256	13.4557	0.0002
vaisseau_B_1	-0.902098	0.7982	1.2771	0.2584

To do that, we select the DEFINE STATUS 2 component and activate the parameters menu. We clear the current selection and select only the 4 significant predictive attributes.



We activate the VIEW menu of the logistic regression. We obtain the following results.

Classifier performances

Error rate			0.185			
Values prediction			Confusion matrix			
Value	**Recall**	**1-Precision**		**presence**	**absence**	**Sum**
presence	0.7416	0.175	presence	66	23	89
absence	0.8739	0.1917	absence	14	97	111
			Sum	80	120	200

Classifier characteristics

Data description

Target attribute	values)
# descriptors	4

Adjustement quality

Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	164.1061
Chi-2	110.7279
P(>Chi-2)	0

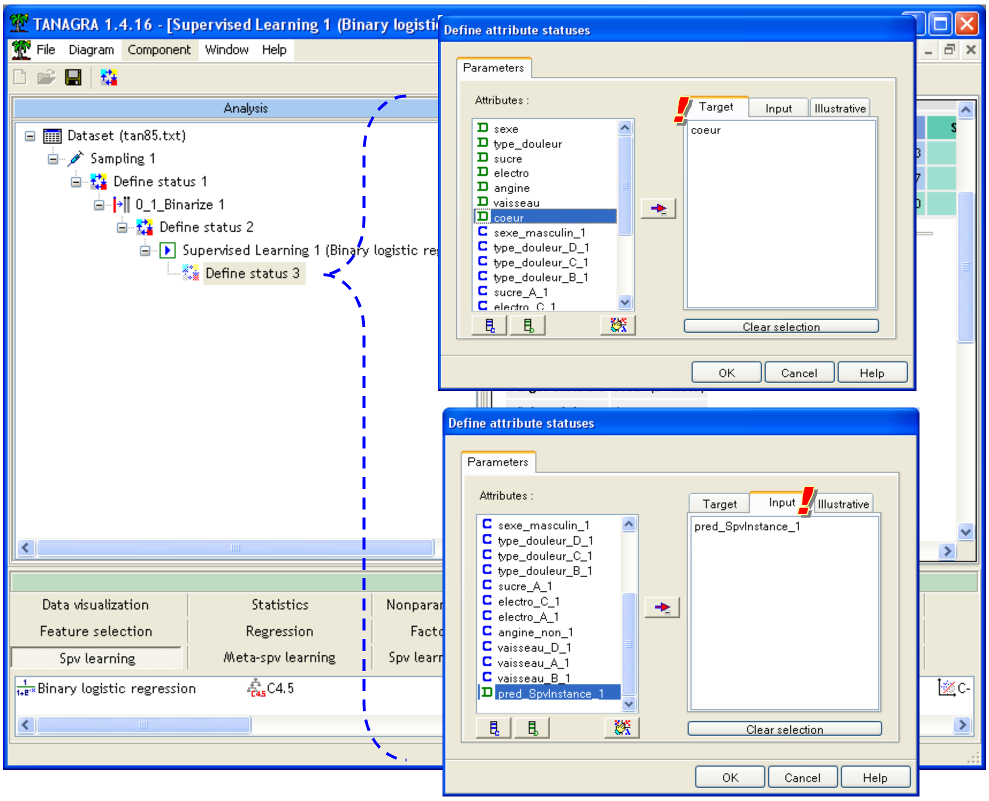
Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.127911	-	-	-
sexe_masculin_1	1.659938	0.4354	14.5359	0.0001
type_douleur_D_1	1.772698	0.4165	18.1156	0
angine_non_1	-1.317853	0.4465	8.7096	0.0032
vaisseau_A_1	-2.024704	0.4146	23.8492	0

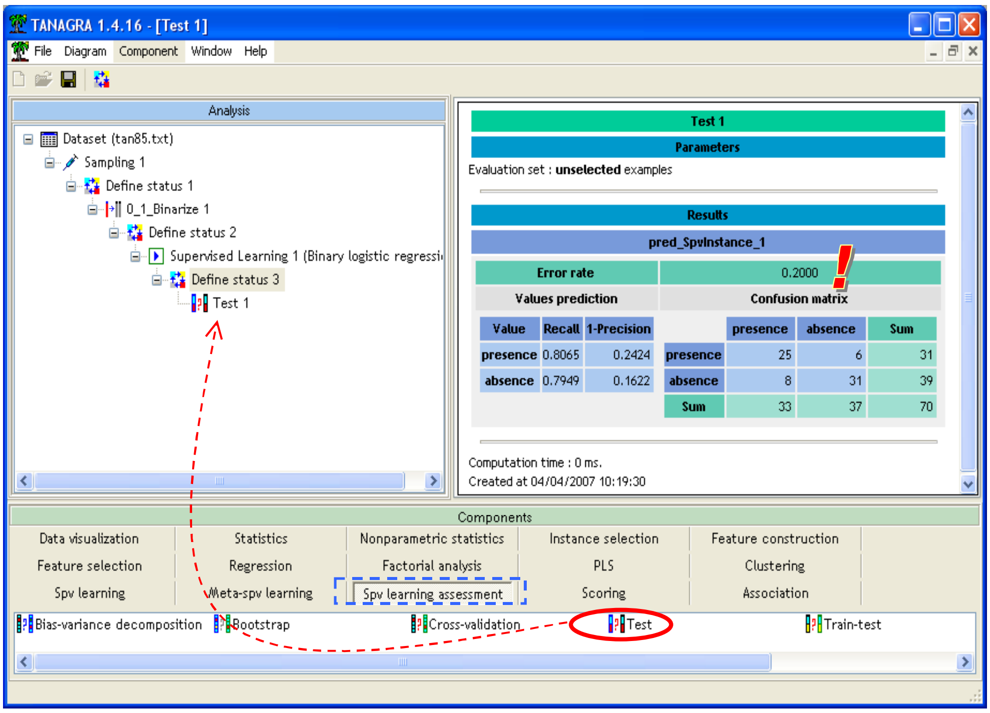
The resubstitution error rate is now 18.5%. All the predictive attributes are significant for a 5% significance level.

Measuring the error rate on the test set

We want now to compute the error rate on the test set. We insert the DEFINE STATUS component. We set as TARGET variable the class attribute (COEUR); the INPUT variable is PRED_SPV_INSTANCE_1. It is automatically generated by the supervised learning component.



We use the TEST component in order to compute the error rate on the test set. We note that the "true" error rate is 20%.



Conclusion

Predictive dummy variables can be used for logistic regression or linear discriminant analysis. In the literature, one advises to use logistic regression instead of discriminant analysis because the Gaussian assumption of this last approach seems not suited for binary variables. In practice, on the majority of the situations, we do not really observe a different behavior of these two supervised learning methods on predictive binary variables.

In our example, the linear discriminant analysis applied to the same variables gives the same performances: the test error rate is 20%.