Subject

Building regression trees with TANAGRA.

The aim of inducing regression trees is to predict the values of a continuous target variable (endogenous) from input variables (exogenous), continuous or discrete.

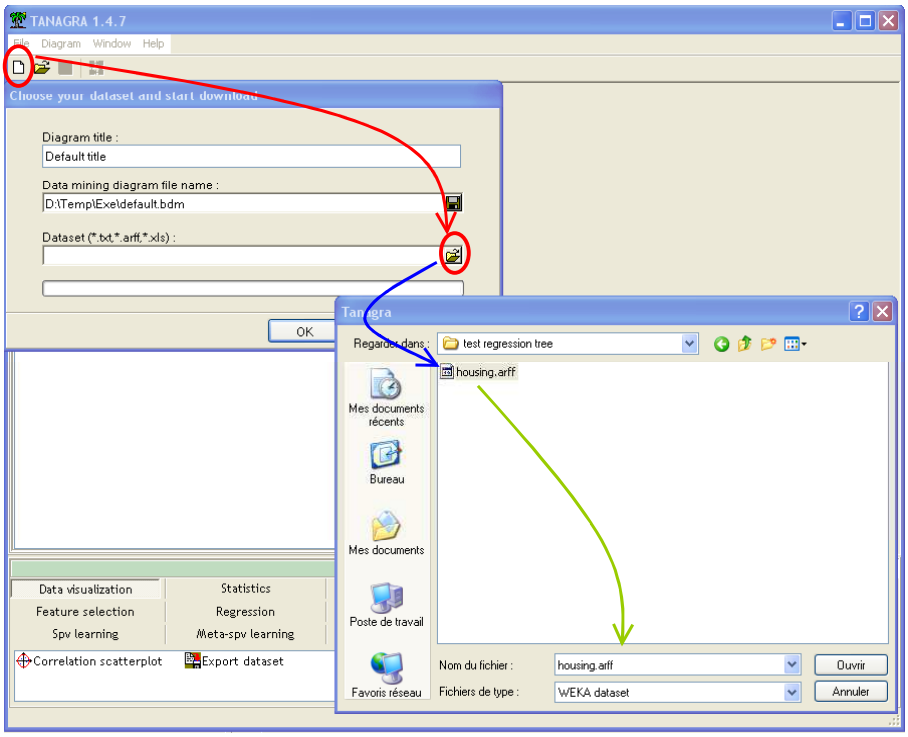
Dataset

We use the famous HOUSING dataset (STATLIB Library). We want to predict the median of housing values in suburbs of Boston.

Regression tree with TANAGRA

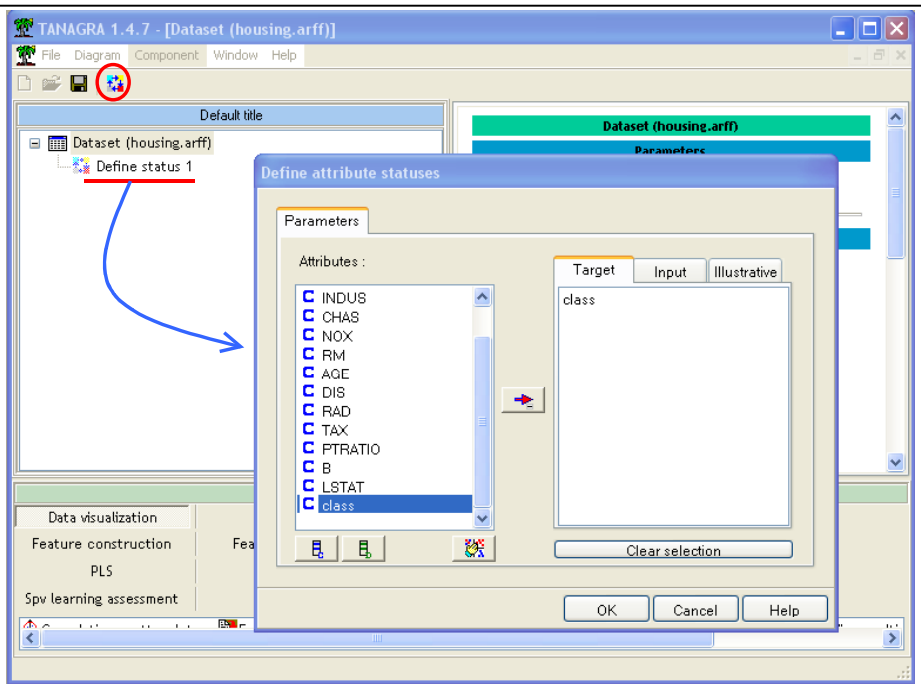
Download the dataset

We must create a diagram and import the dataset. We click on the FILE/NEW Menu and select the HOUSING.ARFF file.



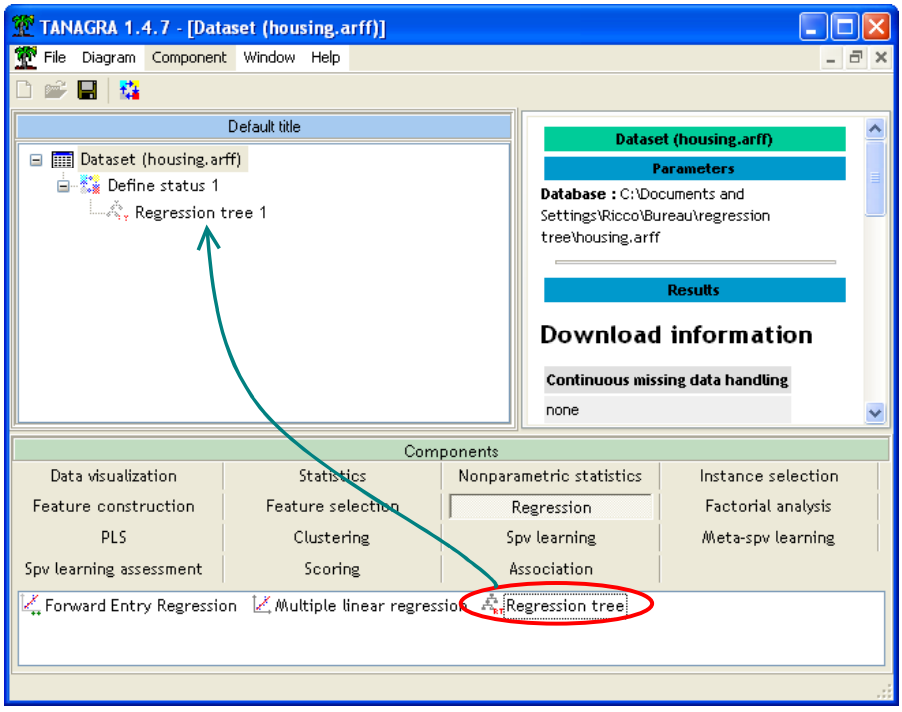
Defining the target and input attributes

We add a DEFINE STATUS component with the toolbar button. We set CLASS as TARGET and all the other attributes as INPUT. Contrary to the linear regression, the input attributes are allowed to be continuous and/or discrete.



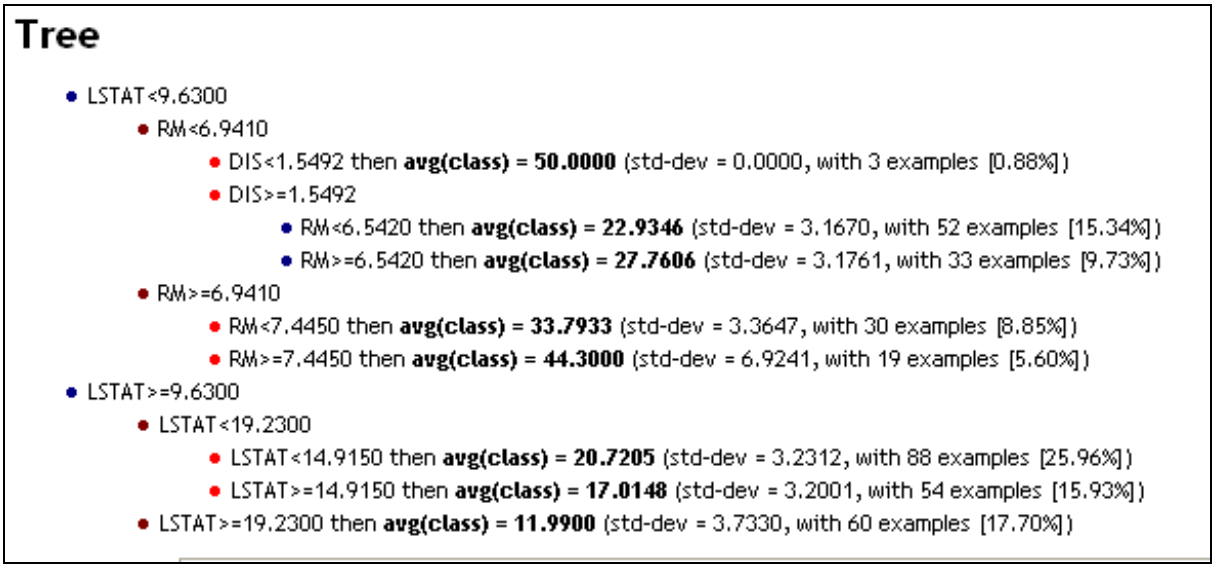
Regression tree

We add the REGRESSION TREE component (REGRESSION palette).



We click on the VIEW menu. The results are displayed in the main window. There are several areas.

The TREE area displays the induced regression tree. We see that the root node of the tree is split with the LSTAT attribute with the cut point 9.63. The other attributes that appear in the tree are DIS and RM.



The characteristics of the problem are summarized in the GLOBAL RESULTS area. We note especially the R^2. It indicates the variance explained with the regression. If we obtain $R^2 = 1$, we have a perfect regression; the regression is very poor if we obtain $R^2 = 0$. In this last case, the best tree is the tree reduced to the root node.

Global results

Endogenous attribute	**class**
Examples	506
R^2	0.8376

The TREE SEQUENCES area points out the growing and the pruning error reduction (RE = 1 – R^2) according the number of leaves of the tree.

Our regression tree algorithm is very similar to CART. The dataset is split into growing and pruning set. We use a two steps algorithm. In the first time, we build a maximal tree that fits as possible the growing set. In the second time, we test nested sub-trees according to the cost-complexity principle, and evaluate the RE on the pruning set. We do not select the optimal tree on the pruning set but the simplest sub-tree that have a performance near to the optimal tree.

Trees sequence (# 46) -- Inertia Within-Groups

N°	# Leaves	Inertia (growing set)	Inertia (pruning set)
46	1	1.0000	1.0000
39	8	0.1489	0.1909
9	40	0.0571	0.1757
1	50	0.0502	0.1813

Tree with one leaf, the root node

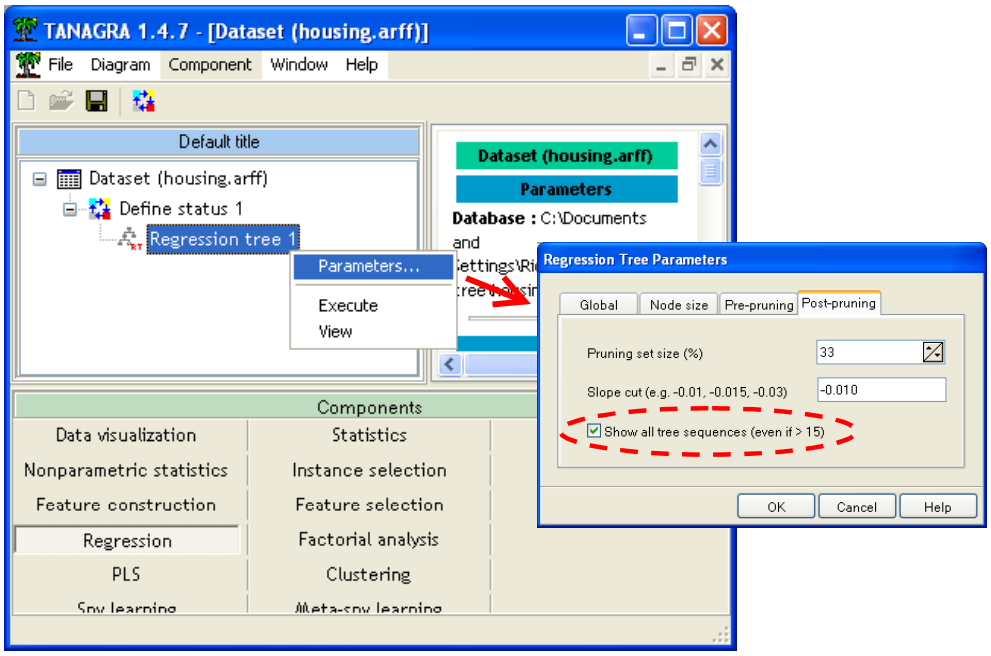
«Selected » tree

« Optimal » tree on the pruning set

Maximal tree -- « Optimal » tree on the growing set

Detailed results

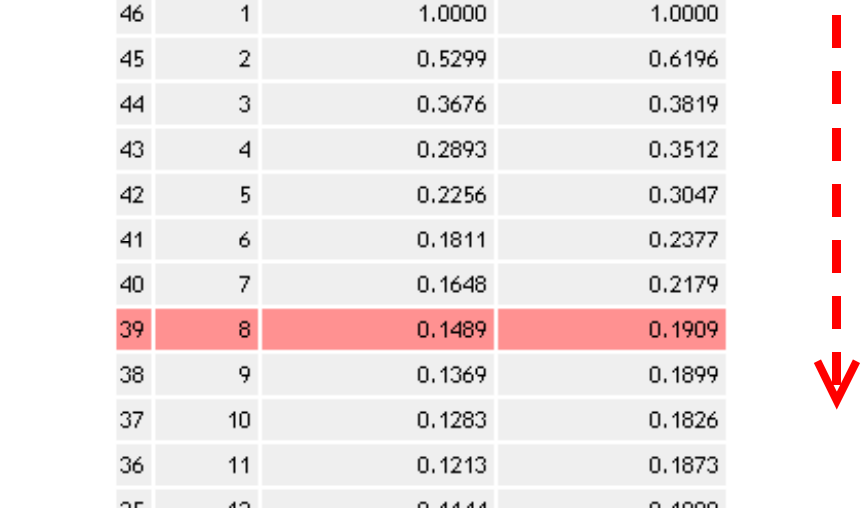
In order to visualize the detailed results, we modify the parameters of the regression tree algorithm (PARAMETERS menu) and check SEE ALL TREE SEQUENCES option.



We execute again the algorithm (VIEW menu). We obtain the detailed results.

Trees sequence (# 46) -- Inertia Within-Groups

N°	# Leaves	Inertia (growing set)	Inertia (pruning set)
46	1	1.0000	1.0000
45	2	0.5299	0.6196
44	3	0.3676	0.3819
43	4	0.2893	0.3512
42	5	0.2256	0.3047
41	6	0.1811	0.2377
40	7	0.1648	0.2179
39	8	0.1489	0.1909
38	9	0.1369	0.1899
37	10	0.1283	0.1826
36	11	0.1213	0.1873



These values can be copied in a spreadsheet. The error reduction curve is characteristic of the behavior of regression trees.