Goal

Show how to import a spreadsheet file format dataset (EXCEL 97 & 2000).

The main advantage of this approach is that it is possible to modify on the fly the data source without having to rebuild the stream diagram (!).

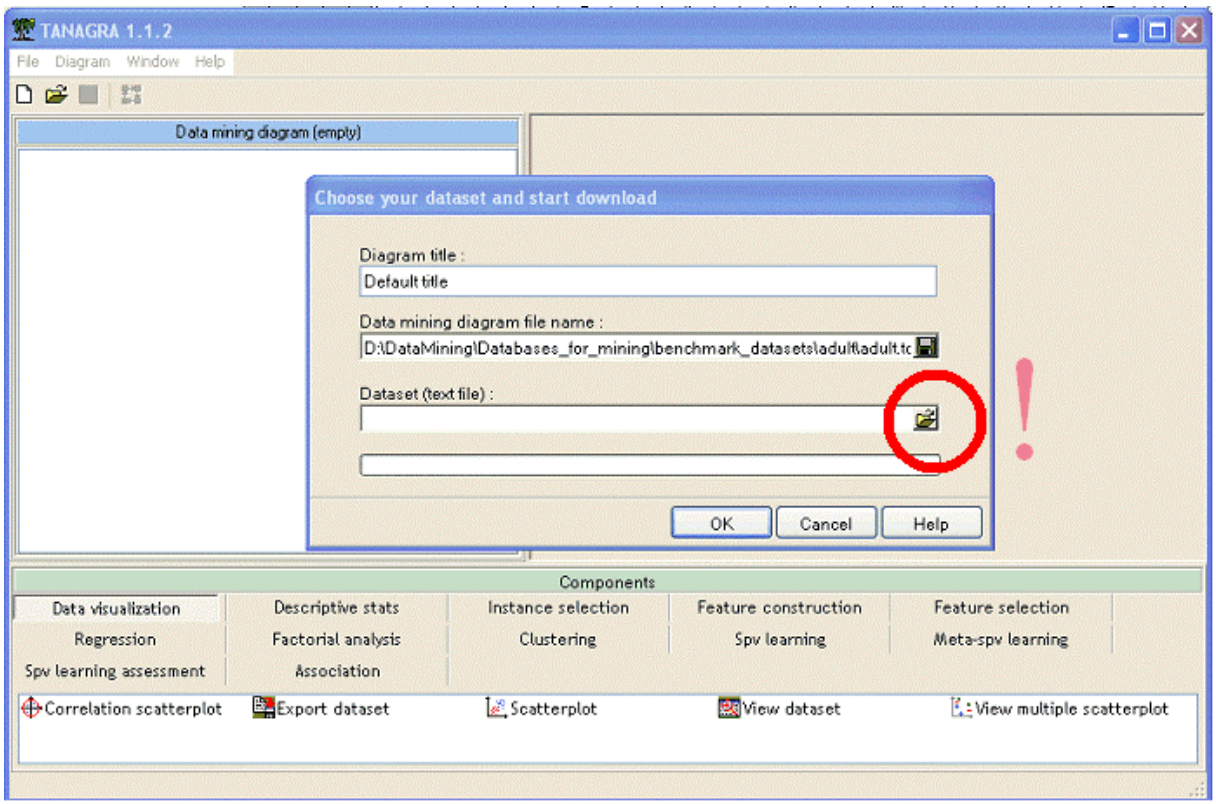
Dataset

ADULT dataset with 48842 examples and 15 attributes. We want to characterize the people who have a high income (more than 50K$).

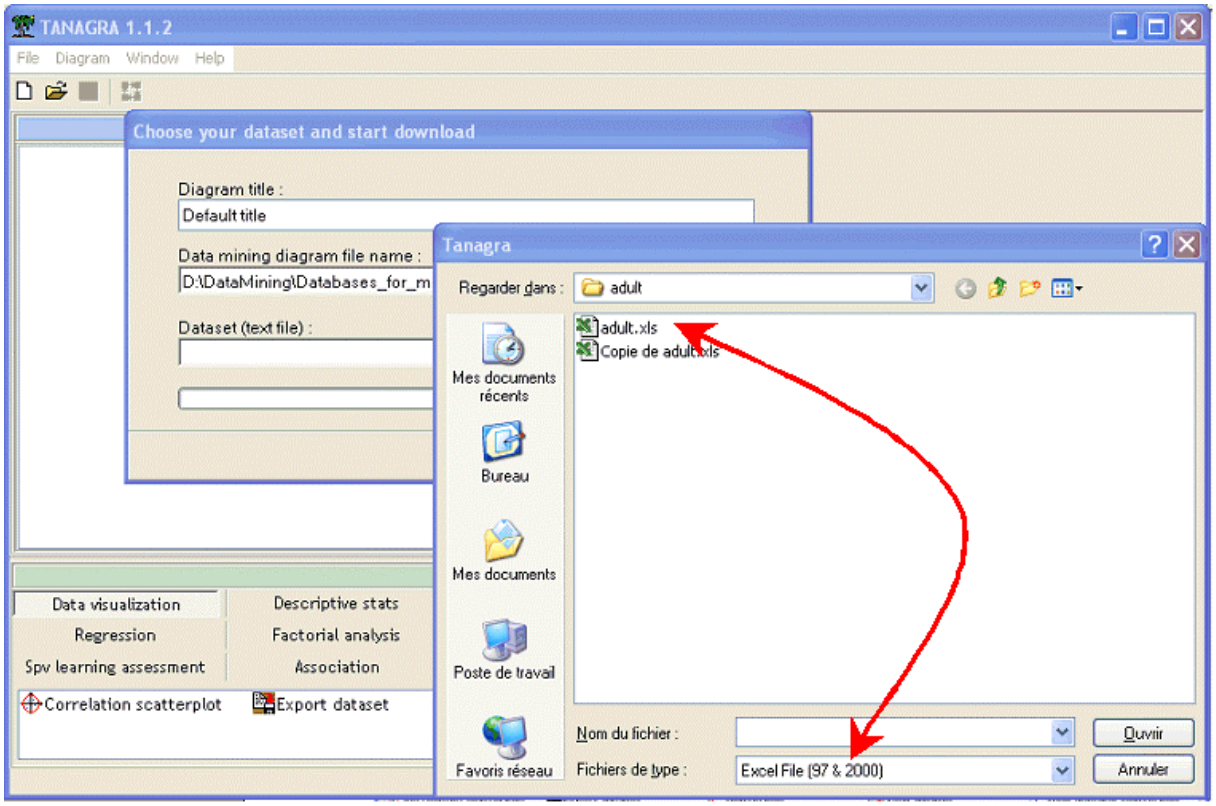
Handle a spreadsheet EXCEL file

Import

The first step is to import the dataset in TANAGRA. Build a new stream diagram according the usual method (File / New).



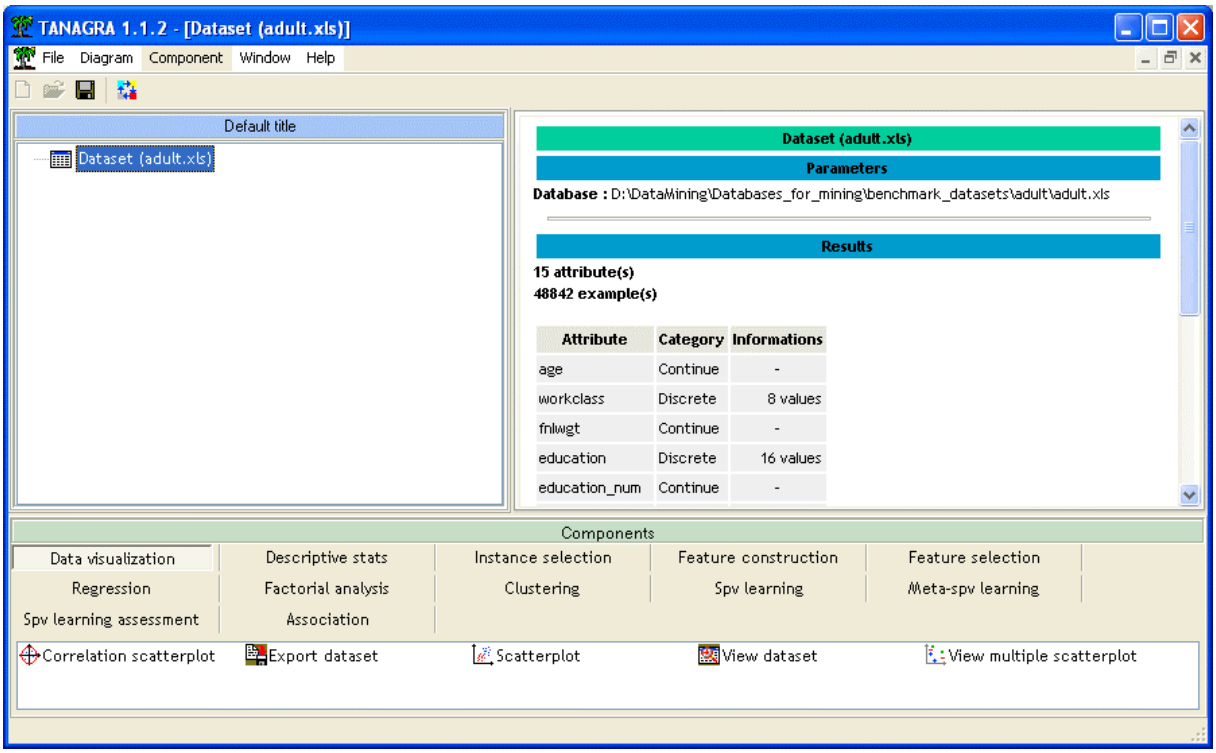
In the dialog box enabling to select the file, several file format is now available, especially the EXCEL file format. Select this one and download the ADULT.XLS dataset.



Some precautions are necessary to import the file:
(1) only EXCEL 97 & 2000 file format are accepted;
(2) if there are several sheets in the workbook, the dataset must be in the first one;
(3) the dataset must be aligned in the left corner of the sheet i.e. first value (name of the first attribute) must be in the A1 cell;
(4) the first row corresponds to the name of attributes;
(5) there is no examples label;
(6) there is no missing values or column/row empty in the dataset.

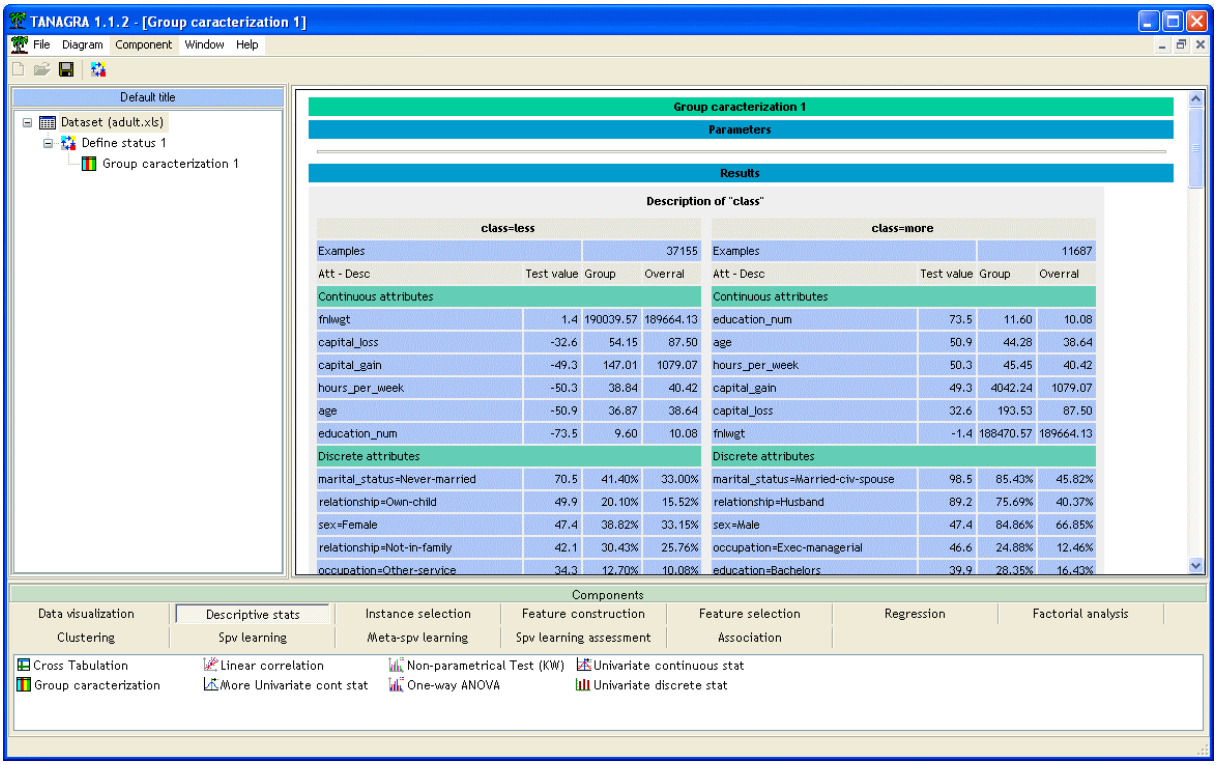
Maximum number of examples and attributes that we can handle is limited by EXCEL capacities: 65534 examples (first row contains attributes name) and 256 attributes.

Importation is rather fast (7 seconds on a P4 at 3Ghz).



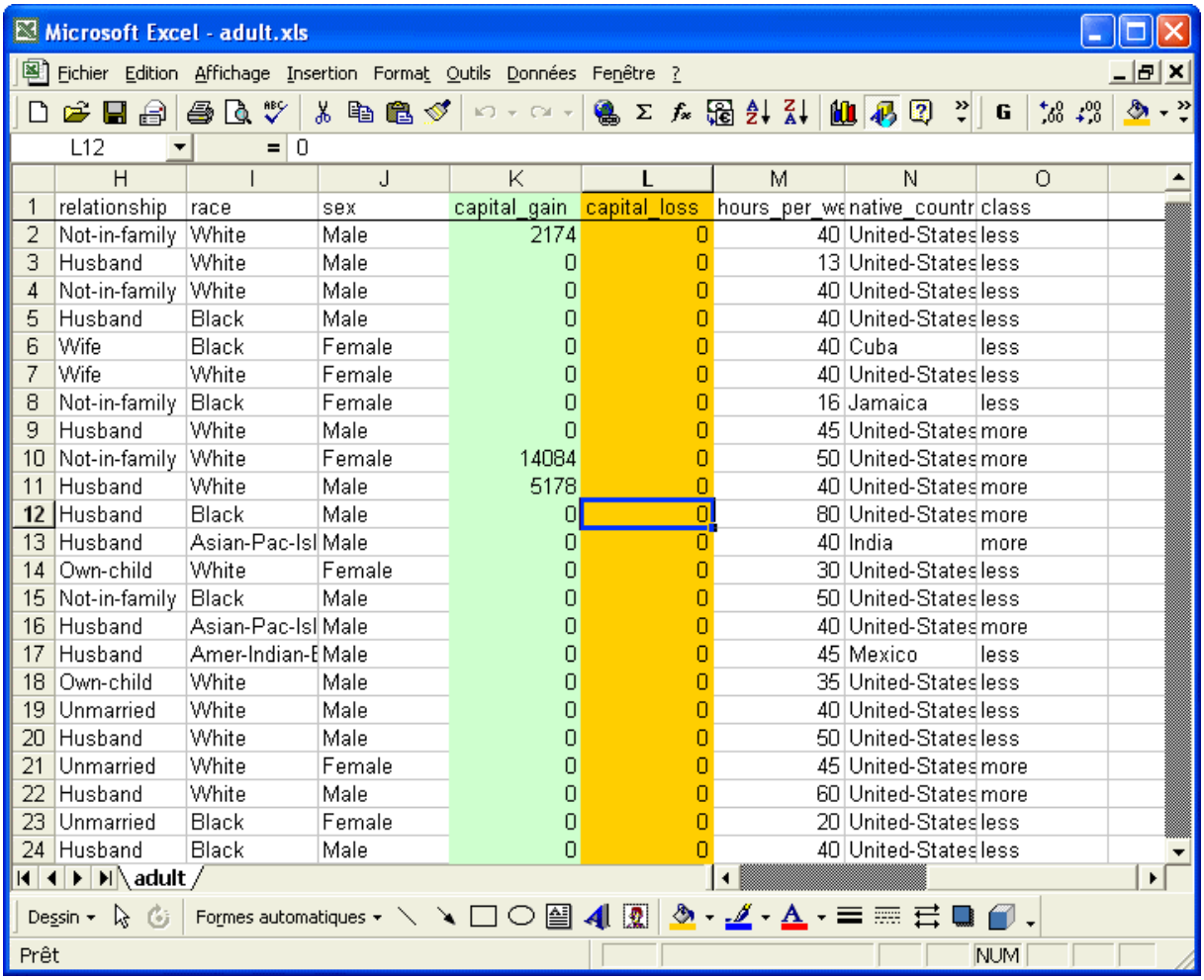
Characterization of the attribute CLASS

We use the GROUP CHARACTERIZATION component to describe CLASS. Before, insert the DEFINE STATUS in the diagram and set CLASS to TARGET, and the other one to INPUT. **Save the diagram.**



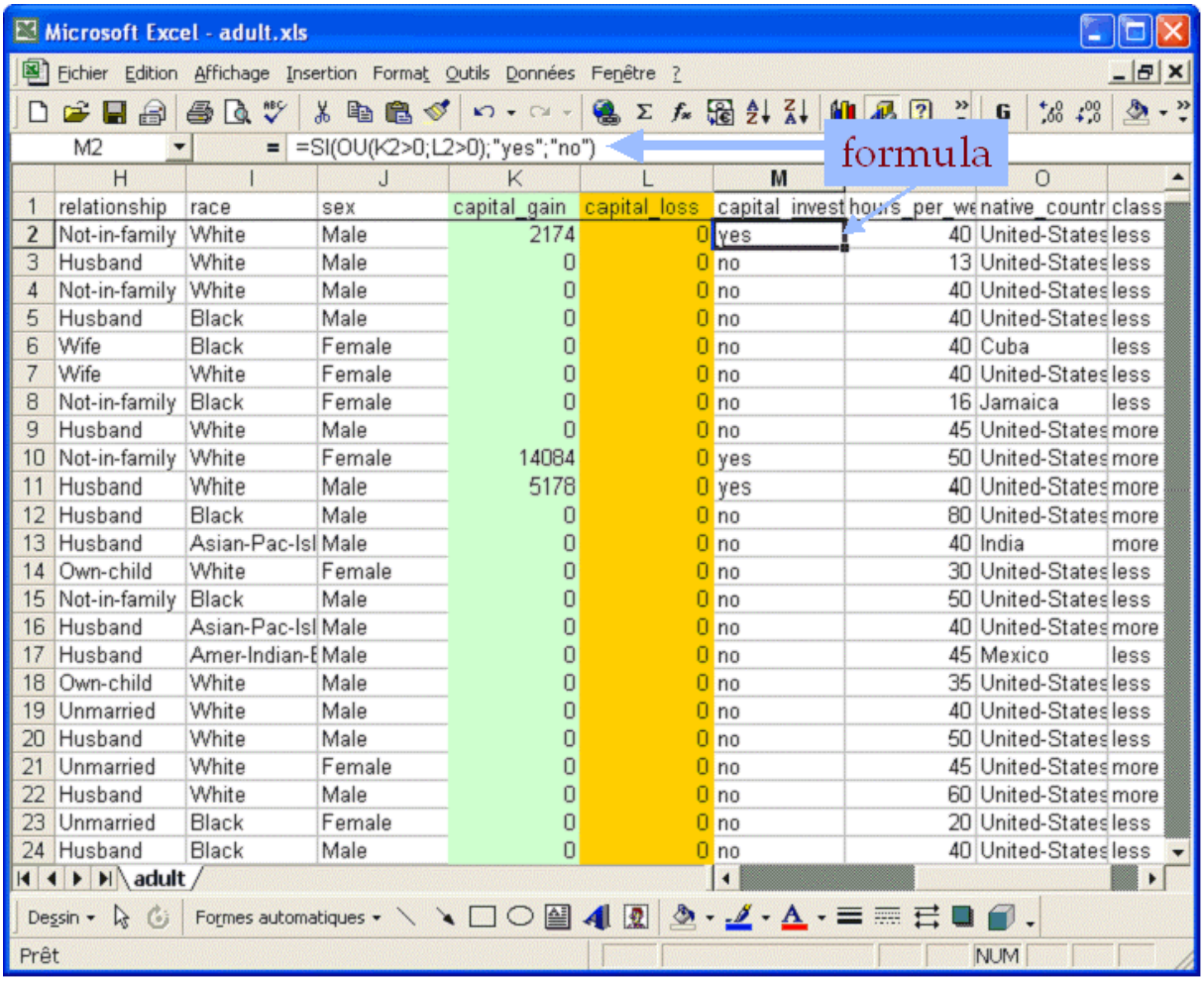
We see that the people with high income are married men, rather old, with a high level of education. These persons have a high capital gain (ok) and high capital loss (???).

While returning on the initial file (ADULT.XLS), we note that these two variables gather two types of information: if the person does not place are money, the profit (the loss) is equal to zero; and if they place their money, the variable indicates the amount of the profit (of the loss). Moreover, one individual having carried out a profit cannot carry out loss, and conversely.



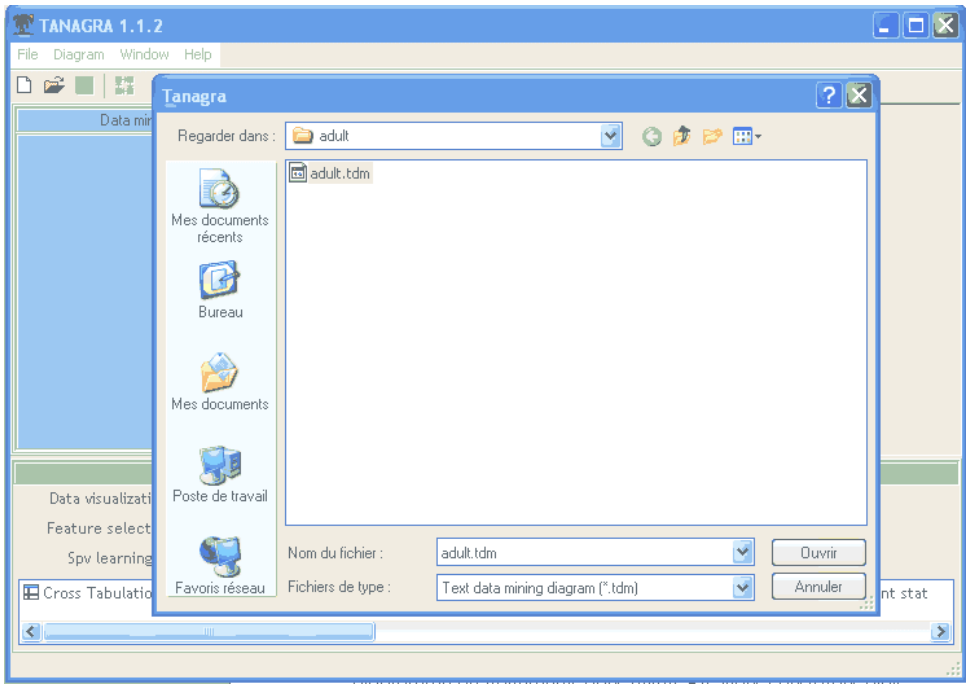
	H	I	J	K	L	M	N	O
1	relationship	race	sex	capital_gain	capital_loss	hours_per_we	native_countr	class
2	Not-in-family	White	Male	2174	0	40	United-States	less
3	Husband	White	Male	0	0	13	United-States	less
4	Not-in-family	White	Male	0	0	40	United-States	less
5	Husband	Black	Male	0	0	40	United-States	less
6	Wife	Black	Female	0	0	40	Cuba	less
7	Wife	White	Female	0	0	40	United-States	less
8	Not-in-family	Black	Female	0	0	16	Jamaica	less
9	Husband	White	Male	0	0	45	United-States	more
10	Not-in-family	White	Female	14084	0	50	United-States	more
11	Husband	White	Male	5178	0	40	United-States	more
12	Husband	Black	Male	0	0	80	United-States	more
13	Husband	Asian-Pac-Isl	Male	0	0	40	India	more
14	Own-child	White	Female	0	0	30	United-States	less
15	Not-in-family	Black	Male	0	0	50	United-States	less
16	Husband	Asian-Pac-Isl	Male	0	0	40	United-States	more
17	Husband	Amer-Indian-E	Male	0	0	45	Mexico	less
18	Own-child	White	Male	0	0	35	United-States	less
19	Unmarried	White	Male	0	0	40	United-States	less
20	Husband	White	Male	0	0	50	United-States	less
21	Unmarried	White	Female	0	0	45	United-States	more
22	Husband	White	Male	0	0	60	United-States	more
23	Unmarried	Black	Female	0	0	20	United-States	less
24	Husband	Black	Male	0	0	40	United-States	less

We want to create a new attribute which points out the fact of investing or not its money.

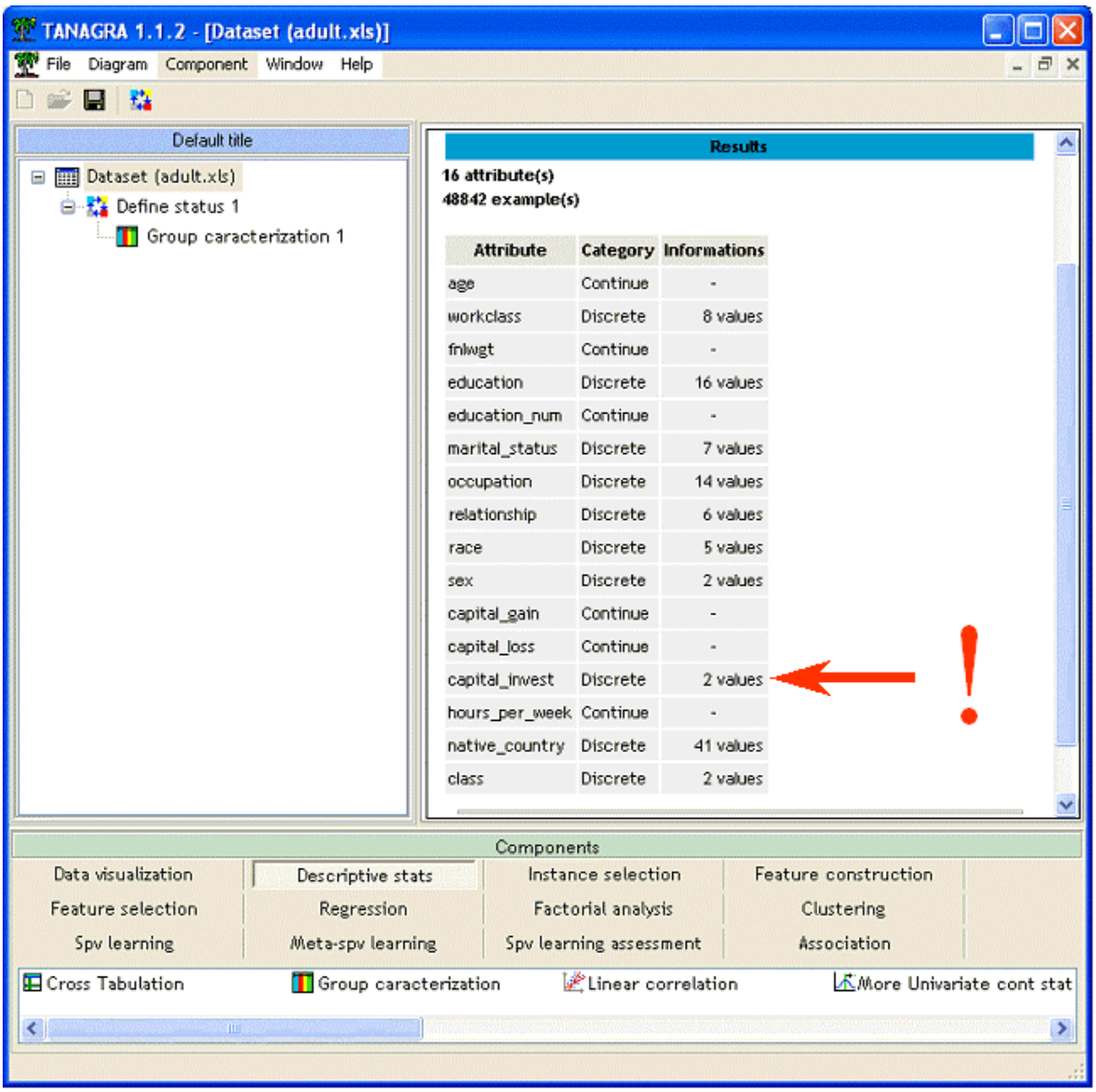


	H	I	J	K	L	M	O	
1	relationship	race	sex	capital_gain	capital_loss	capital_invest	native_countr	class
2	Not-in-family	White	Male	2174	0	yes	United-States	less
3	Husband	White	Male	0	0	no	United-States	less
4	Not-in-family	White	Male	0	0	no	United-States	less
5	Husband	Black	Male	0	0	no	United-States	less
6	Wife	Black	Female	0	0	no	Cuba	less
7	Wife	White	Female	0	0	no	United-States	less
8	Not-in-family	Black	Female	0	0	no	Jamaica	less
9	Husband	White	Male	0	0	no	United-States	more
10	Not-in-family	White	Female	14084	0	yes	United-States	more
11	Husband	White	Male	5178	0	yes	United-States	more
12	Husband	Black	Male	0	0	no	United-States	more
13	Husband	Asian-Pac-Isl	Male	0	0	no	India	more
14	Own-child	White	Female	0	0	no	United-States	less
15	Not-in-family	Black	Male	0	0	no	United-States	less
16	Husband	Asian-Pac-Isl	Male	0	0	no	United-States	more
17	Husband	Amer-Indian-E	Male	0	0	no	Mexico	less
18	Own-child	White	Male	0	0	no	United-States	less
19	Unmarried	White	Male	0	0	no	United-States	less
20	Husband	White	Male	0	0	no	United-States	less
21	Unmarried	White	Female	0	0	no	United-States	more
22	Husband	White	Male	0	0	no	United-States	more
23	Unmarried	Black	Female	0	0	no	United-States	less
24	Husband	Black	Male	0	0	no	United-States	less

After having saved the workbook and closed EXCEL, close and open the diagram, the new version of the dataset is parsed and downloaded.

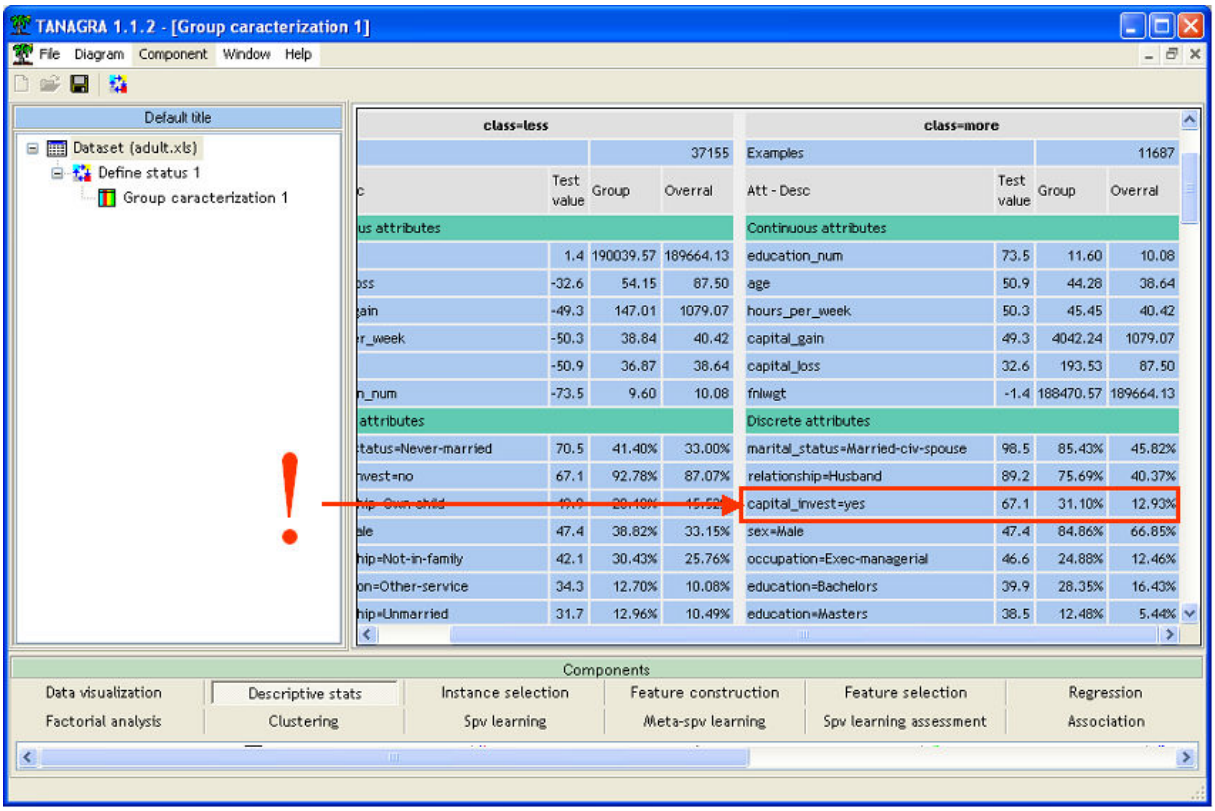


We note that:
(1) diagram is preserved;
(2) the new attribute was added in the dataset.



In the DEFINE STATUS component, add this new attribute among INPUT and run the diagram.

We note that CAPITAL rather indicates a behavior with respect to the saving: the people who have a higher income invest their money (12.9% in the whole dataset, 31% in this group).


TANAGRA 1.1.2 - [Group caracterization 1]
File Diagram Component Window Help
Default title
Dataset (adult.xls)
Define status 1
Group caracterization 1
class=less
class=more
Examples
37155
11687
Test value
Group
Overral
Att - Desc
Continuous attributes
education_num 73.5 11.60 10.08
age 50.9 44.28 38.64
hours_per_week 50.3 45.45 40.42
capital_gain 49.3 4042.24 1079.07
capital_loss 32.6 193.53 87.50
fnlwgt -1.4 188470.57 189664.13
Discrete attributes
marital_status=Married-civ-spouse 98.5 85.43% 45.82%
relationship=Husband 89.2 75.69% 40.37%
capital_invest=yes 67.1 31.10% 12.93%
sex=Male 47.4 84.86% 66.85%
occupation=Exec-managerial 46.6 24.88% 12.46%
education=Bachelors 39.9 28.35% 16.43%
education=Masters 38.5 12.48% 5.44%
Components
Data visualization
Descriptive stats
Instance selection
Feature construction
Feature selection
Regression
Factorial analysis
Clustering
Spv learning
Meta-spv learning
Spv learning assessment
Association