Subject

Starting from the Tanagra's 1.4.11 version, a new EXCEL add-in is available. It enables to define a data mining analysis directly from EXCEL spreadsheet without closing the EXCEL session.

The main asset of this functionality is that we can perform all data preparation (data transformation, feature construction, etc.) and basic descriptive statistics (mean, standard deviation, pivot table, etc.) in the spreadsheet. Then we call TANAGRA, from EXCEL, only for advanced machine learning technique.

In this tutorial, we show how to install and use this EXCEL add-in.

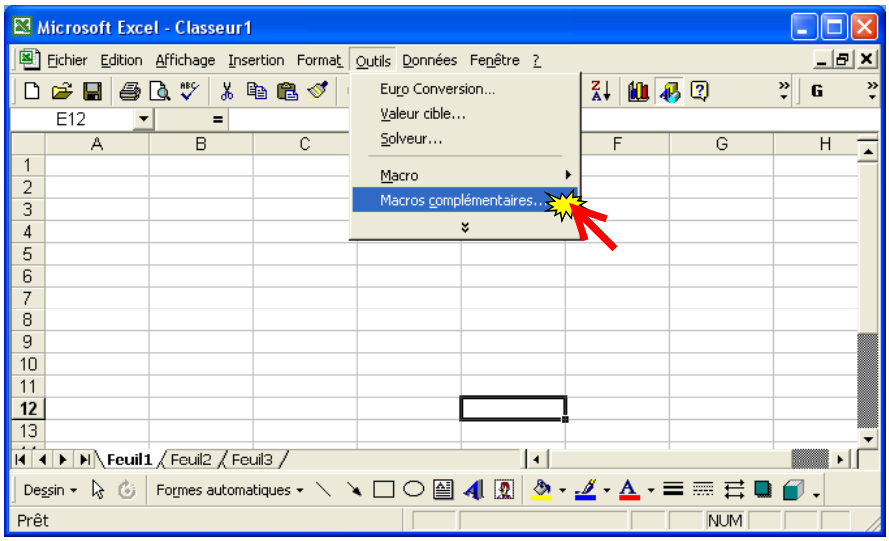
Installing the new EXCEL add-in

Checking presence of the Add-In

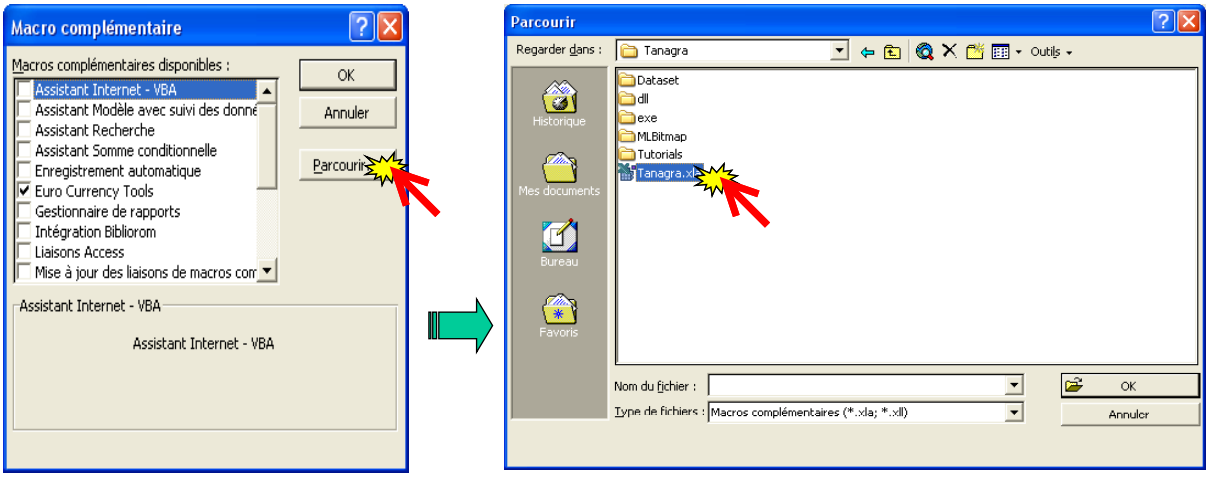
This new add-in is available from the 1.4.11 version. See in the TANAGRA directory if the **TANAGRA.XLA** really exists (with the standard installation, the TANAGRA's directory is usually « *c:\program files\tanagra* »). **Do not move this file**.

Installing the Add-In

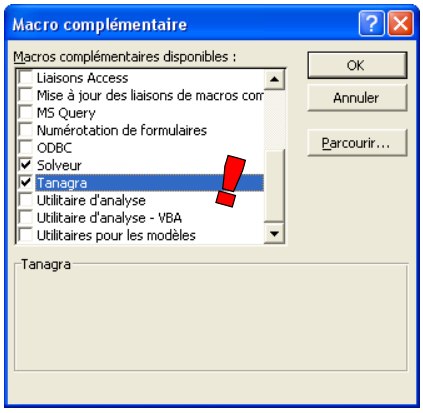
In the next step, we must install the add-in in the spreadsheet. We click on the following menu.



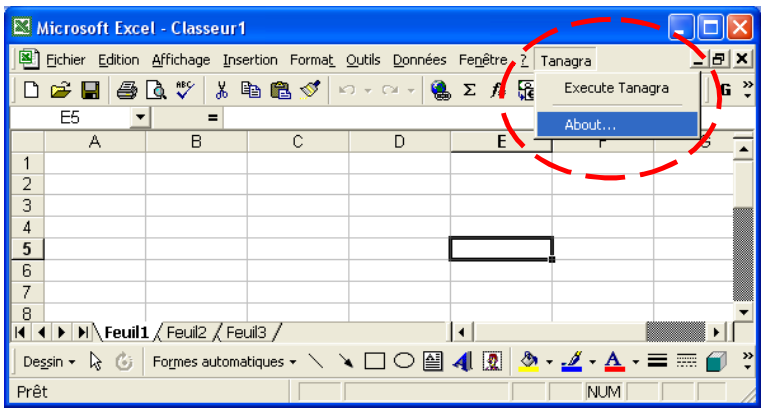
A dialog box appears, we select the XLA file in TANAGRA's directory.



The add-in is downloaded, we check if it is really activated.



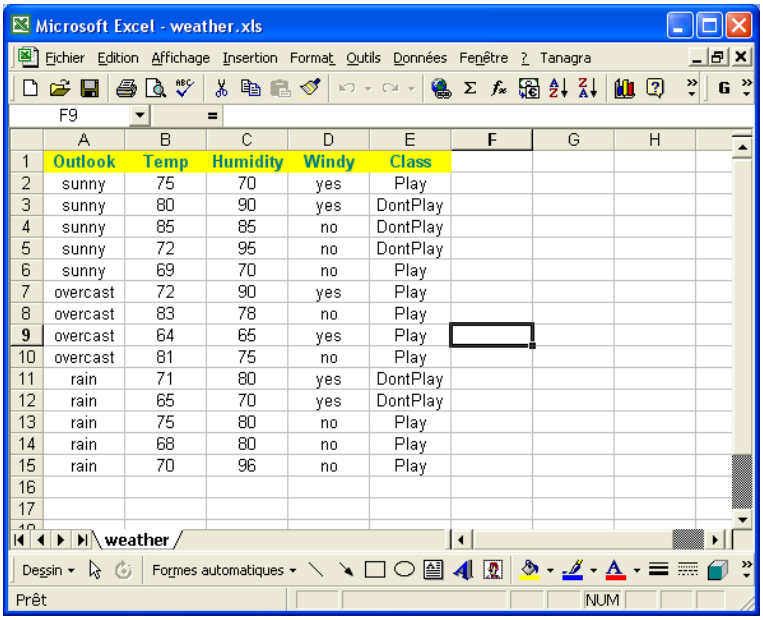
A new menu is now available in the EXCEL spreadsheet.



From now, until we remove this add-in, this menu is always available when we start the spreadsheet.

Working on a dataset

We use the Quinlan's WEATHER.XLS dataset (1993) in order to show the utilization of this add-in.

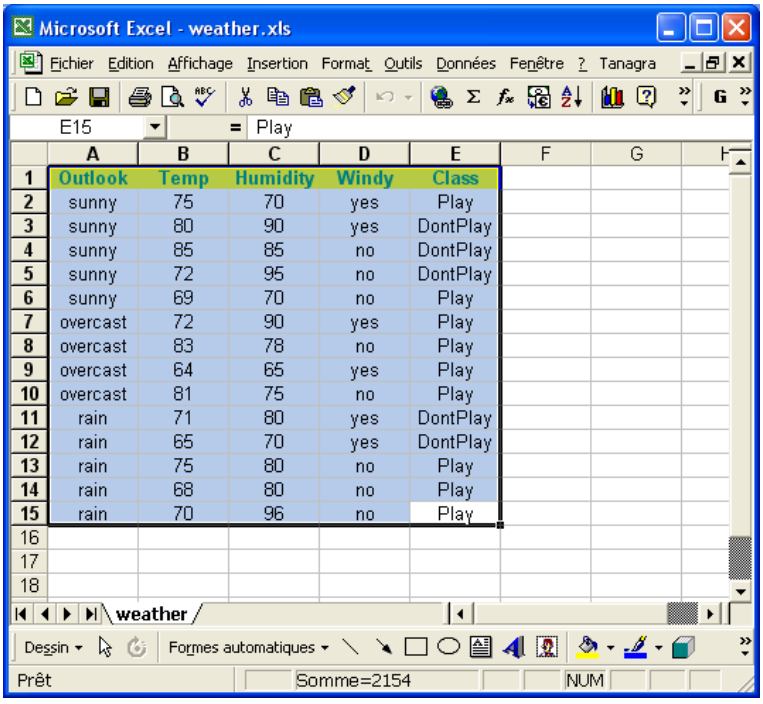


	A	B	C	D	E
1	Outlook	Temp	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	80	90	yes	DontPlay
4	sunny	85	85	no	DontPlay
5	sunny	72	95	no	DontPlay
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	65	yes	Play
10	overcast	81	75	no	Play
11	rain	71	80	yes	DontPlay
12	rain	65	70	yes	DontPlay
13	rain	75	80	no	Play
14	rain	68	80	no	Play
15	rain	70	96	no	Play

Selecting the dataset range

Before we activate the menu, we must **select the dataset range**.

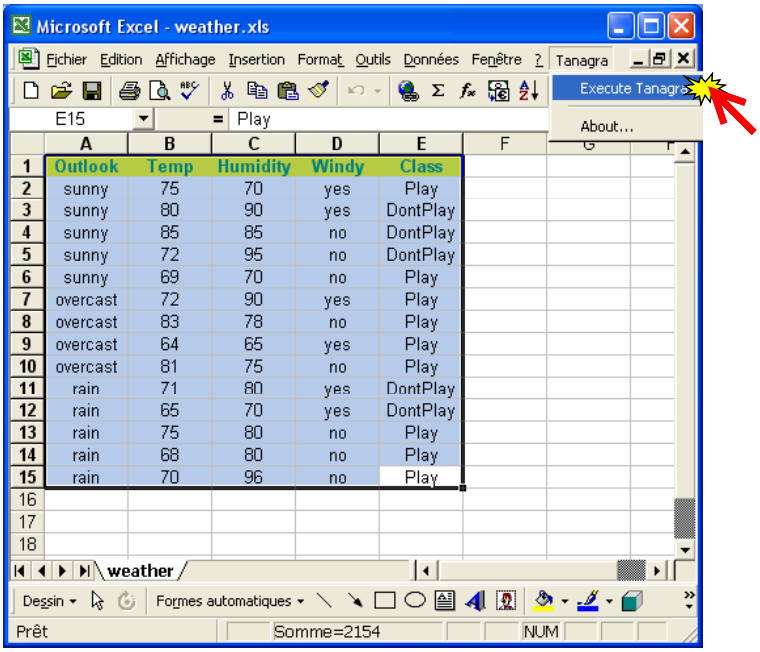
The first row stands for the attribute name. The determination of the data type relies on the first row of the data: if it can be transformed in a numeric value, the variable is turned into a continuous attribute; it is defined as a discrete one otherwise.



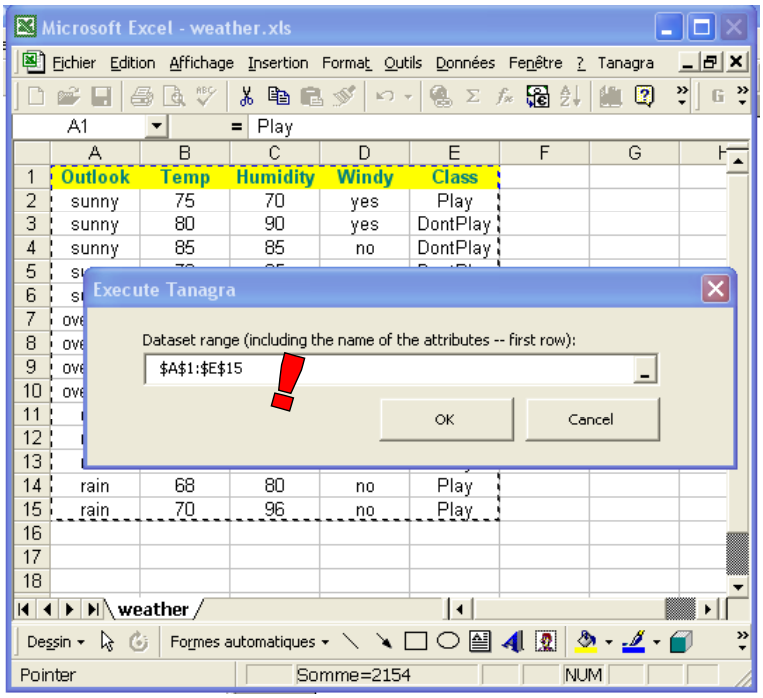
	A	B	C	D	E
1	Outlook	Temp	Humidity	Windy	Class
2	sunny	75	70	yes	Play
3	sunny	80	90	yes	DontPlay
4	sunny	85	85	no	DontPlay
5	sunny	72	95	no	DontPlay
6	sunny	69	70	no	Play
7	overcast	72	90	yes	Play
8	overcast	83	78	no	Play
9	overcast	64	65	yes	Play
10	overcast	81	75	no	Play
11	rain	71	80	yes	DontPlay
12	rain	65	70	yes	DontPlay
13	rain	75	80	no	Play
14	rain	68	80	no	Play
15	rain	70	96	no	Play

The TANAGRA / EXECUTE TANAGRA menu

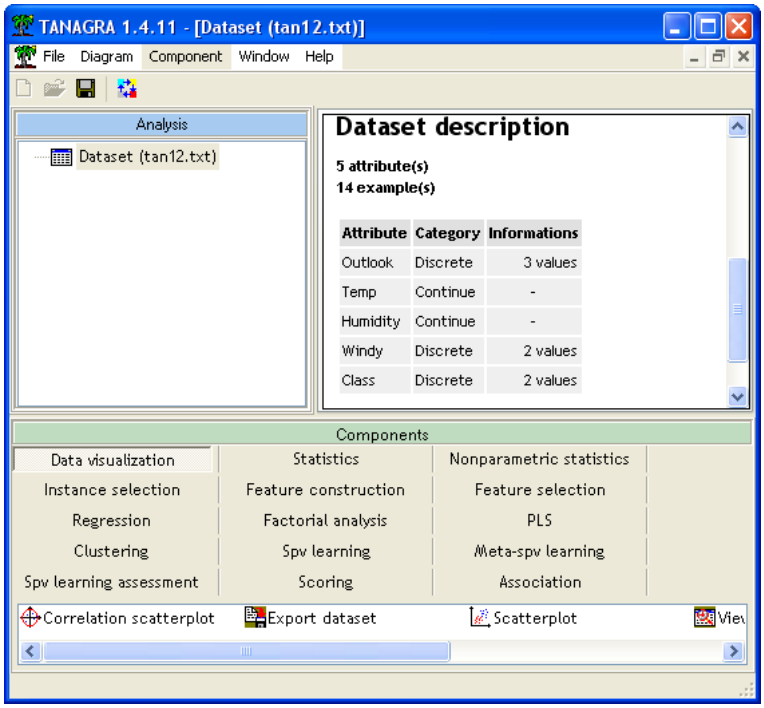
We click on the TANAGRA / EXECUTE TANAGRA menu in order to perform a data mining analysis.



In the next dialog box, we can check and reset the range selection.



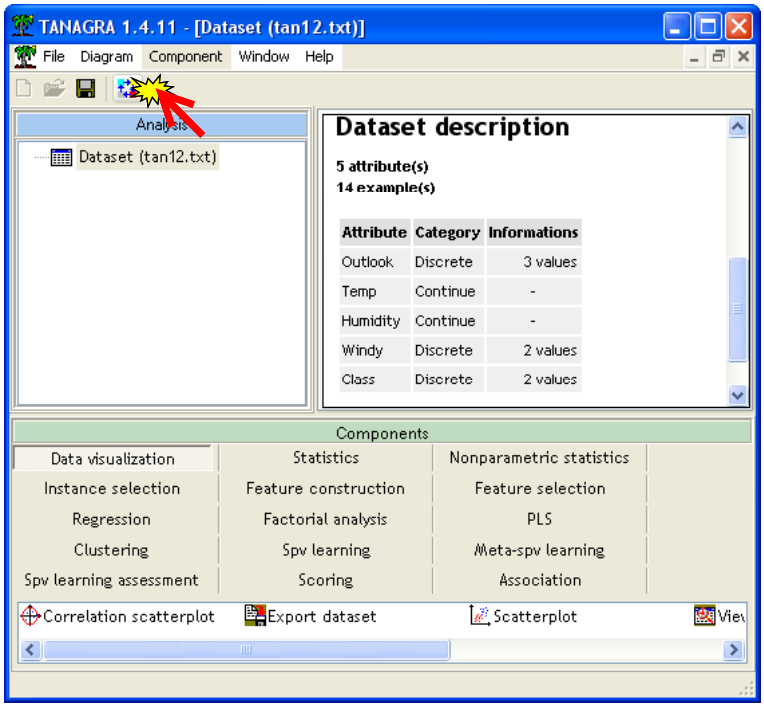
To **validate** the parameter settings, we click on the **OK button**. TANAGRA is automatically executed with the appropriate dataset.



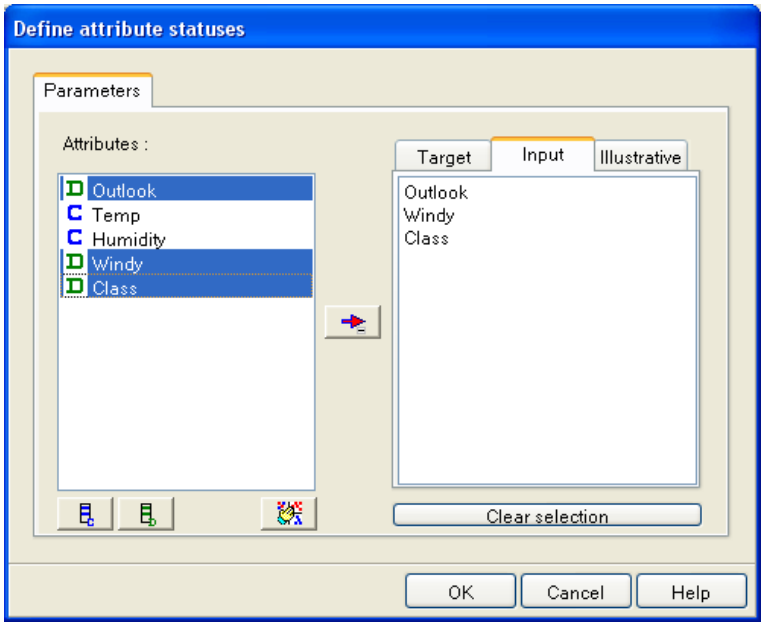
We see that the whole dataset (14 examples and 5 attributes) is really exported. The type of the variables is automatically defined according to the rules described above.

Working with TANAGRA

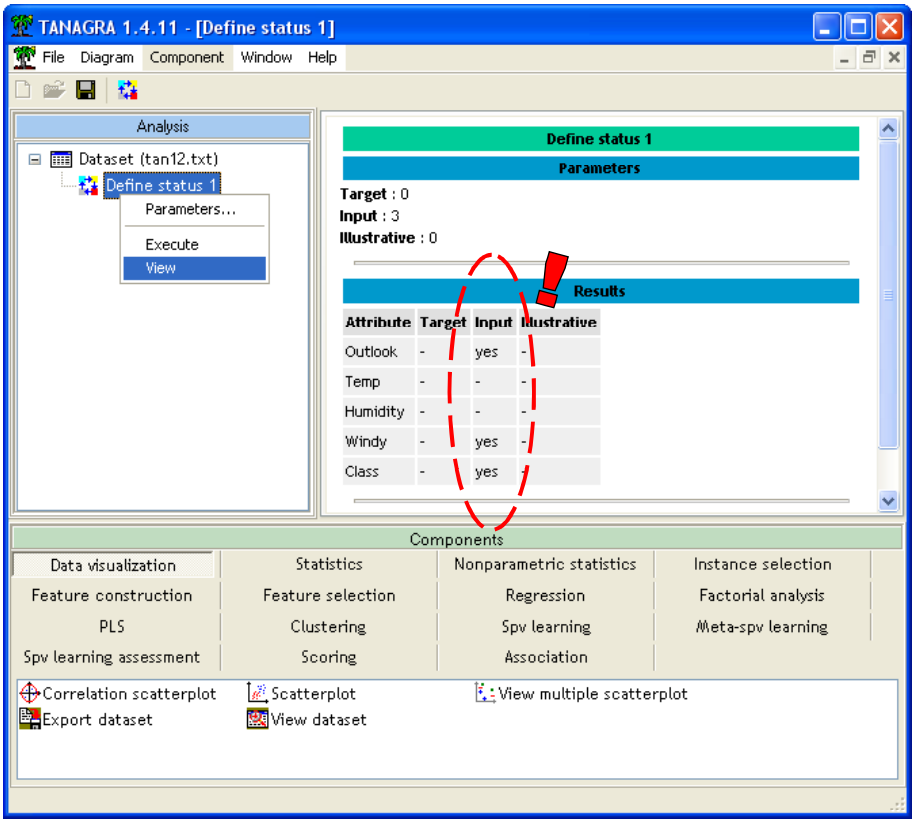
In this tutorial, we want to perform a basic descriptive statistic. First, we must **define the INPUT attributes**. We add the DEFINE STATUS component in the diagram using the short cut in the toolbar.



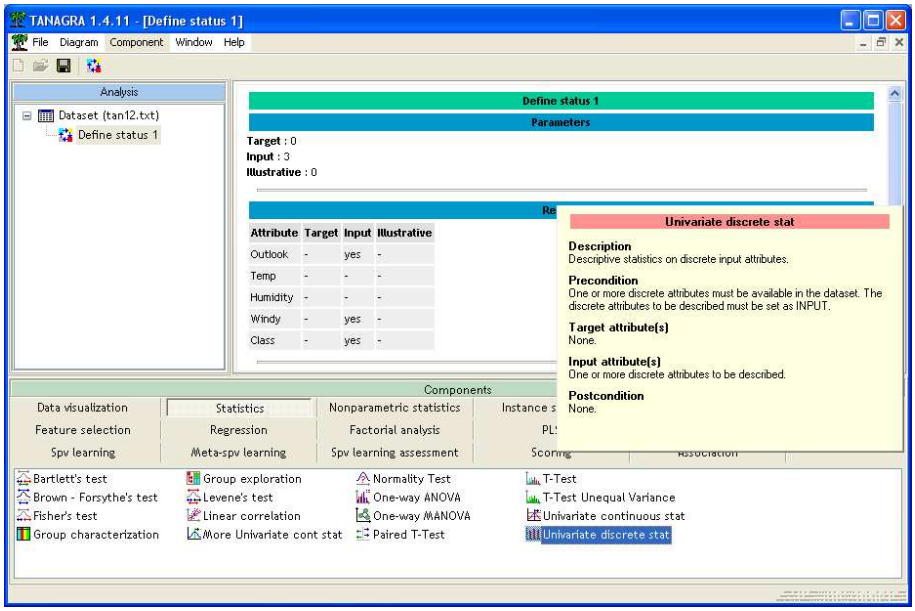
We select the INPUT attributes in the following dialog box.



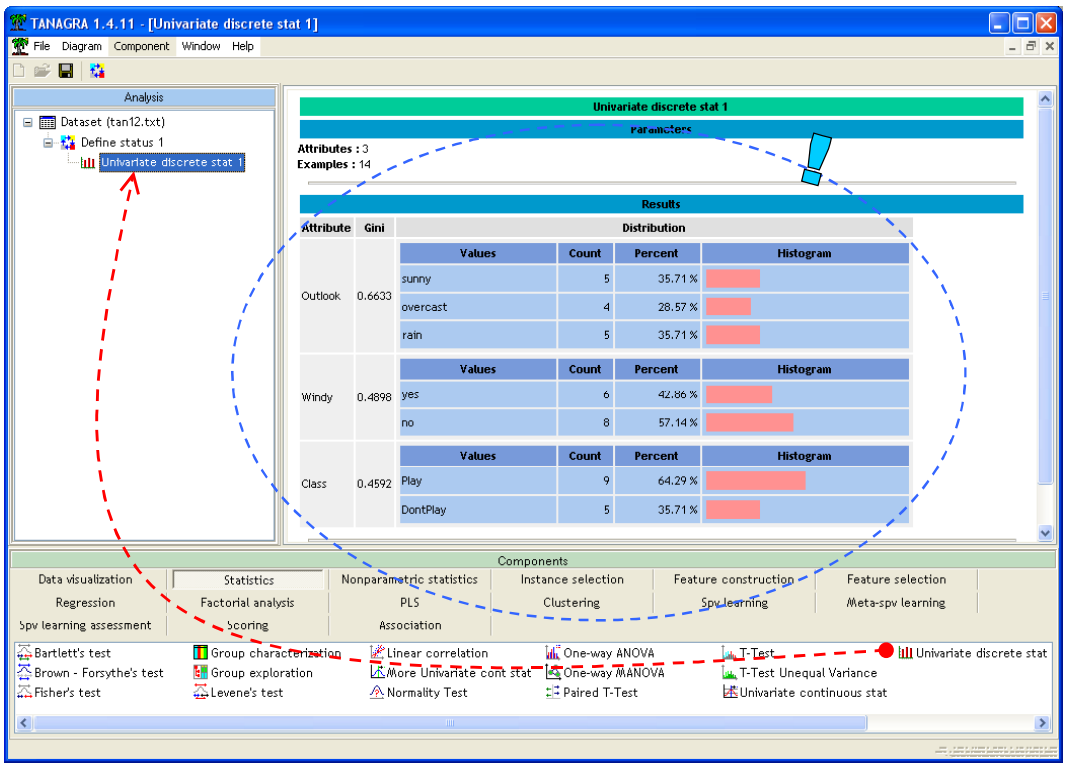
We click on the contextual menu VIEW in order to display the results.



The next component to insert into the diagram is UNIVARIATE DISCRETE STAT from the STATISTICS tab.

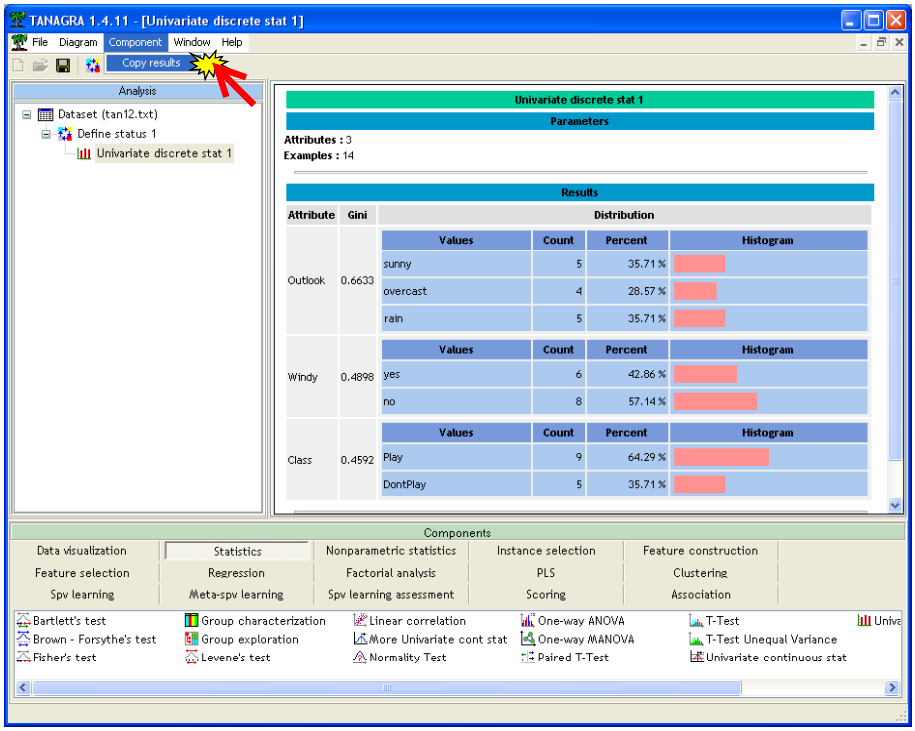


We select this component and drag it under the DEFINE STATUS component into the diagram. We click on the VIEW menu in order to display the results.

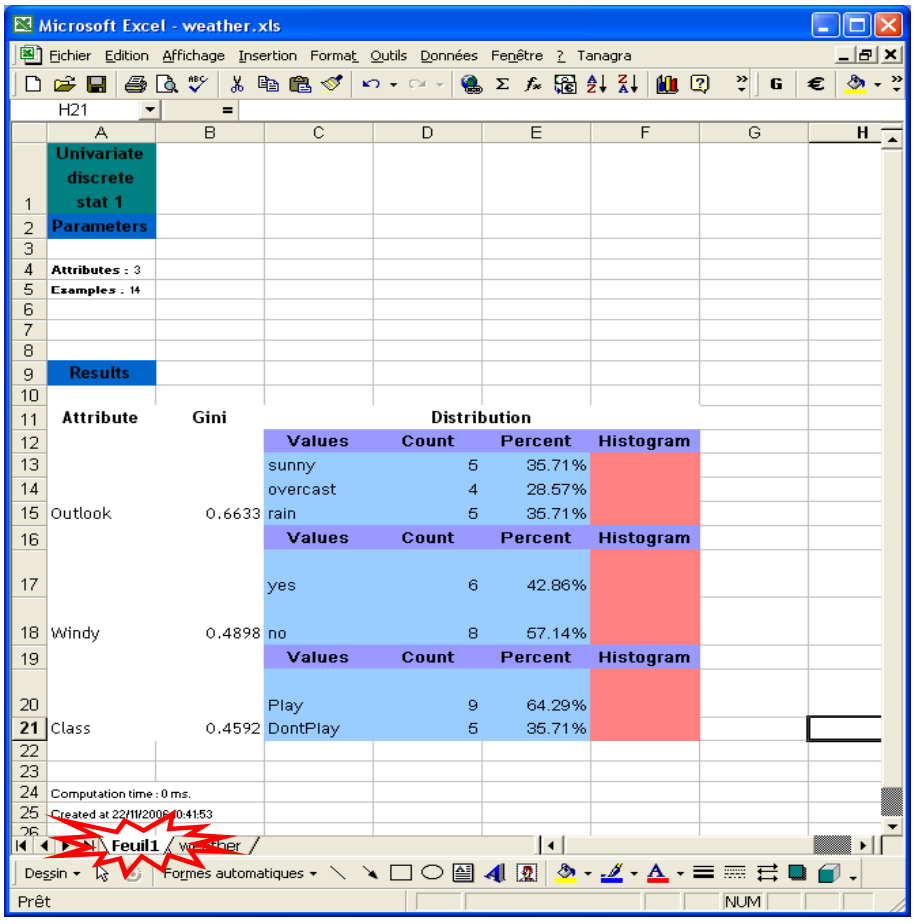


Recovering the results in EXCEL

Because all the results are in the HTML format, we can copy and paste them in the spreadsheet. We click on the COMPONENT / COPY RESULTS menu.



Then, in the spreadsheet, we insert a new sheet and paste the result. The appearance is kept more or less but the essential information is preserved.



Conclusion -- Performance evaluation

In this tutorial, we use a very small dataset in order to show the mechanism of this new functionality. The real question is: what is the quickness of this add-in when we treat a large dataset, knowing that the maximum size is anyway limited by the EXCEL capacities?

Another dataset (SHUTTLE.XLS) is distributed with this tutorial. It contains 58,000 examples and 10 attributes. By using the same method, we note that the computation time (data preparation and exportation towards TANAGRA) lasts only a few seconds.