# Goal

How to handle WEKA file format (.ARFF) ?

***Missing data treatment is very basic here. If you want wide options, use the external module DATANAMORF.***

# Dataset

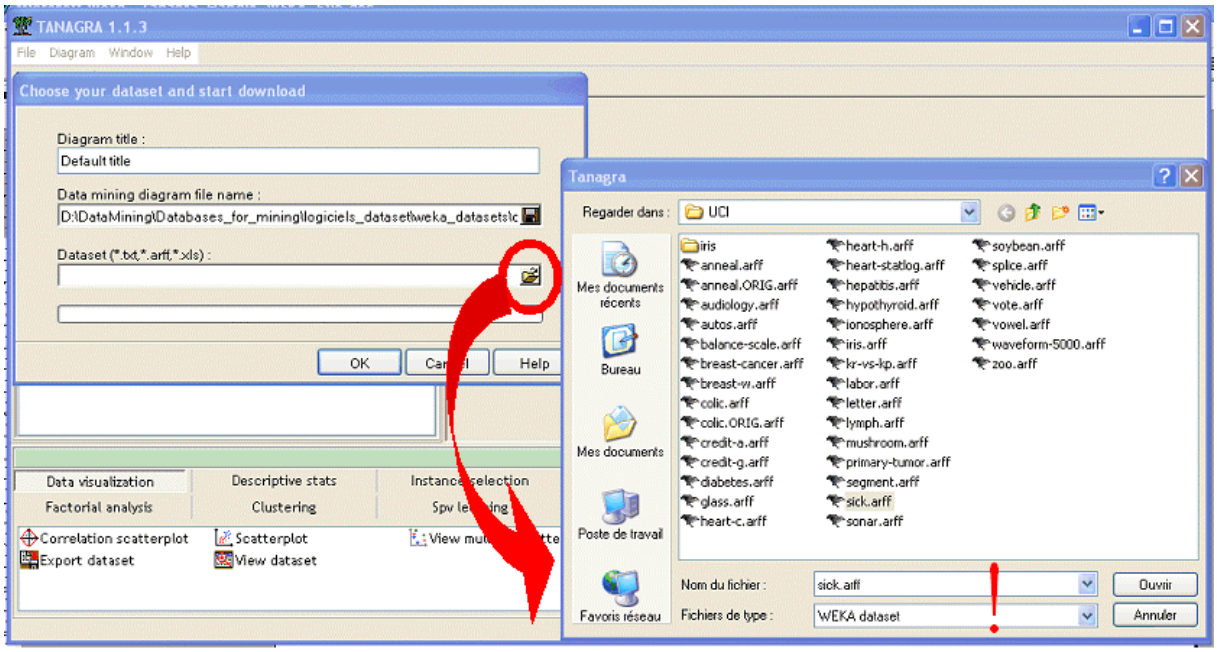
SICK.ARFF.

# Handling WEKA file format (.ARFF)

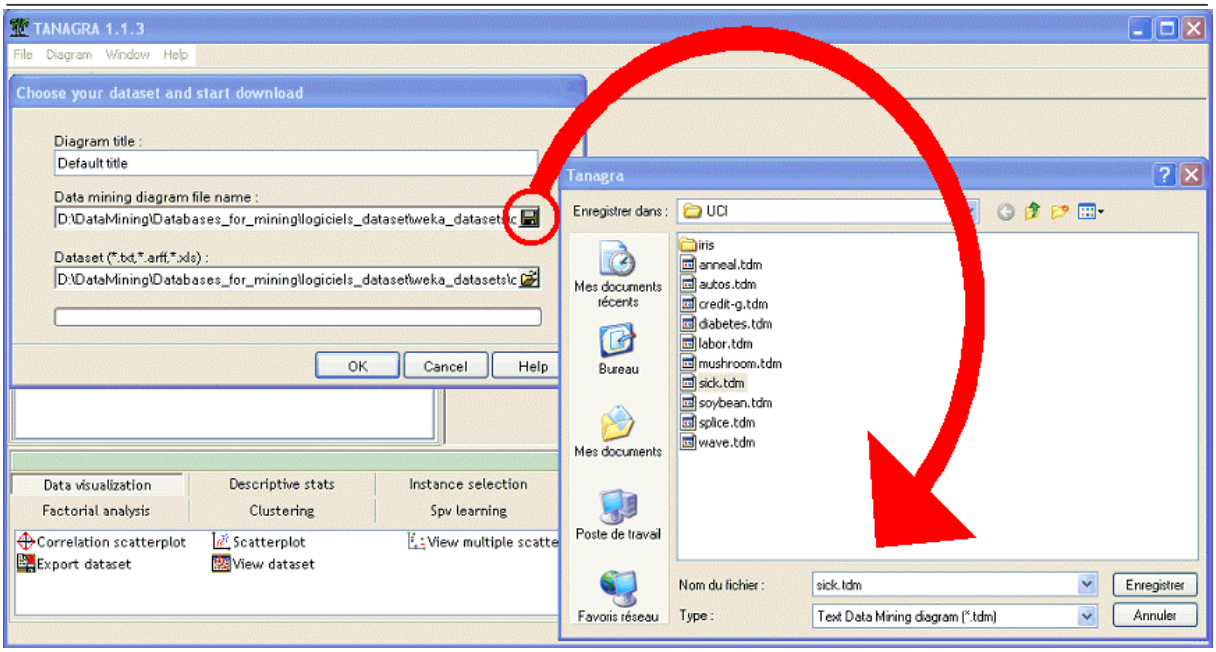
## Download .ARFF file

The first step is to import the dataset and create a diagram. Click on the menu “FILE/NEW”.

In the importation dialog box, select the data source, WEKA file format is now available.




Define the diagram file name (use SICK.TDM) and click on OK.

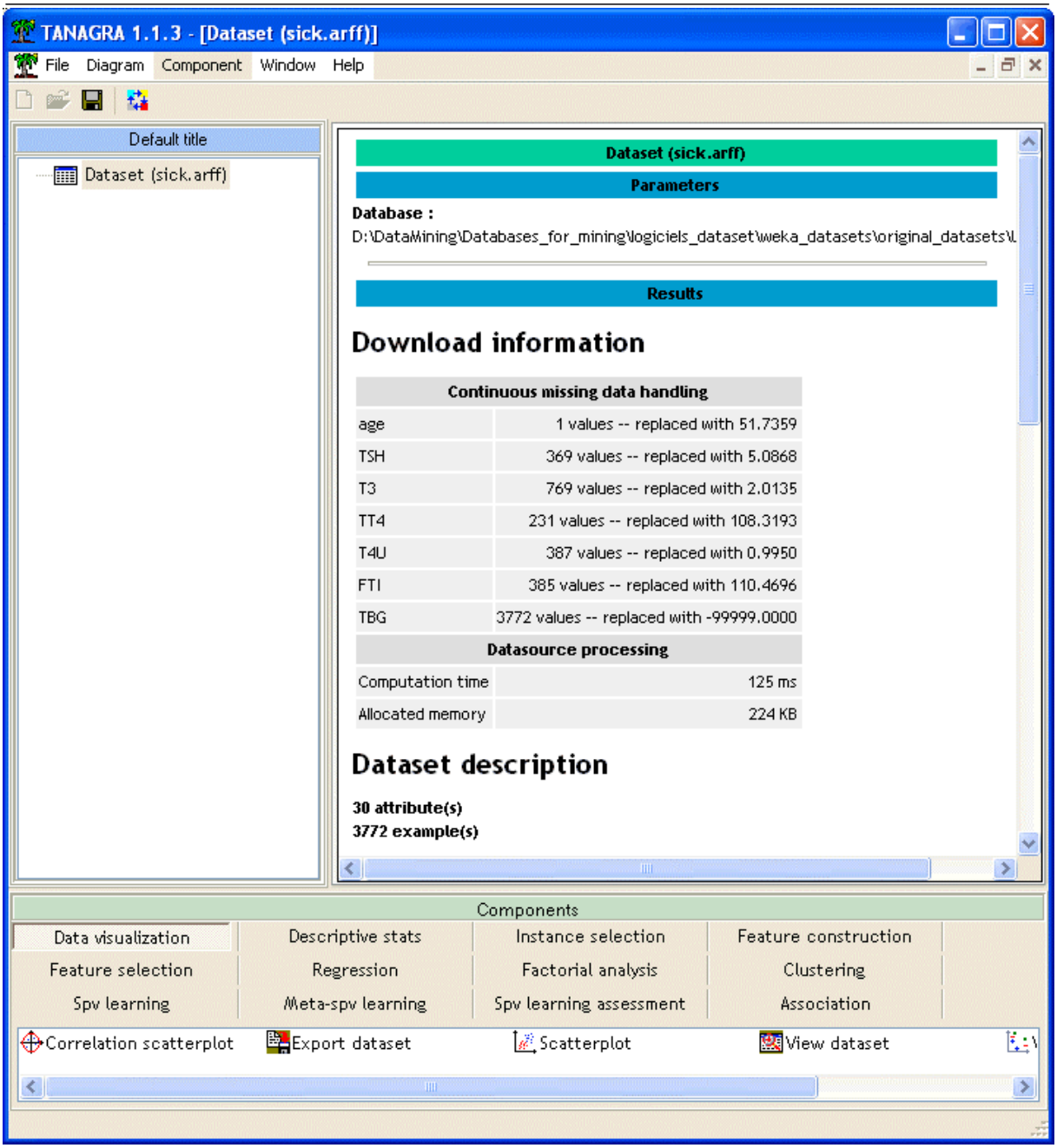



## Report

Additional informations are available in the report, they describe pre-treatment applied on the ARFF dataset. This relates essentially missing data handling. There are two kind of operations :

- For a discrete attribute, missing value (? or '?' in ARFF file) becomes a new value. It is added to the data dictionary. Because there are no replacements of values, this operation is not reported, the number of values of the variable is only incremented.

- For a continuous attribute, missing values are replaced with computed average on available examples. In this case, there is really a treatment, the modifications are reported with 2 kinds of informations : the number of missing values in the column, the replacement value.




There are 30 attributes and 3772 examples in the dataset. Among continuous attributes, 7 have missing values.

AGE for instance has 1 missing value, it was replaced with the computed average on available examples (51.7359).

TBG is particular. There is no valid values in the column (3772 missing values), in this case, we use the default value (-99999).

## Open a diagram

TDM format keeps a reference on the data source (.ARFF). So, you can apply the same treatments diagram on eventually updated dataset.

The next time, when you open the diagram (FILE/OPEN on SICK.TDM), the source file SICK.ARFF will be parsed again, with missing values handling.