

## Tutorial overview

---

This tutorial shows basic characteristics of TANAGRA user interface, through the analysis of the « Breast.txt » dataset.

This well-known dataset come from the medical domain, consists of the characteristics of cells sampled on women presenting (or not) a malignant tumor.

In this tutorial, you'll learn to use the following components:

Tab	Operator (Component)	Function
Data visualization	View dataset	View the contents of the data file in a grid
Feature selection	Define status	Specify the attributes to use
Descriptive stats	Univariate continuous stat	Descriptive statistics for continuous attributes
Descriptive stats	Univariate discrete stat	Descriptive statistics for discrete attributes
Descriptive stats	Group characterization	Statistics for sub-populations

## Importing and viewing data in TANAGRA

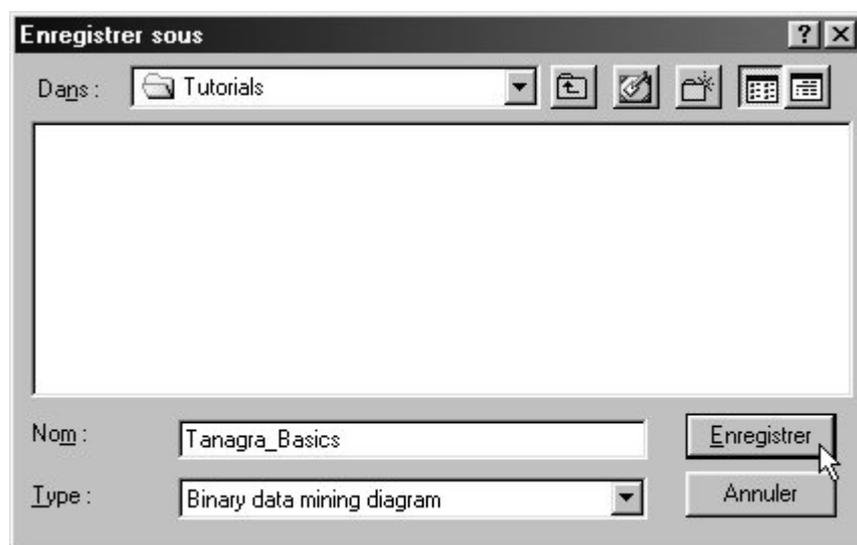
---

### ➤ Creating a new data mining diagram

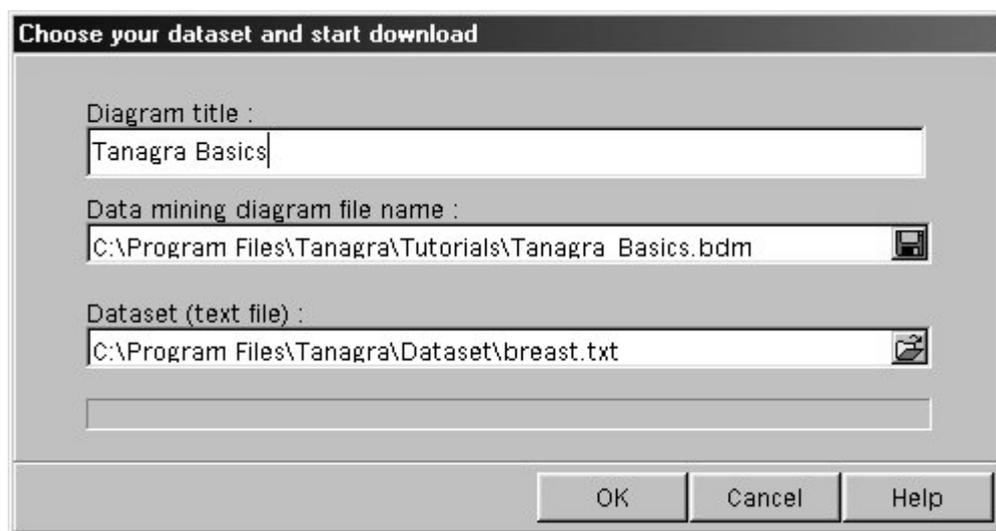
1 – Choose *File/New...* in the main menu of TANAGRA.

2 – Enter a title for the diagram : « TANAGRA Basics ».

3 – Enter the name of the associated file in (« TANAGRA\_Basics.bdm »). Before click on  button: to run through the hard disk and place yourself in the directory « ...\TANAGRA\Tutorials ».



4 – Select the text file containing the data you want to explore by clicking on this icon:   
For this tutorial, choose the file “breast.txt”, located in TANAGRA subdirectory « Dataset ».



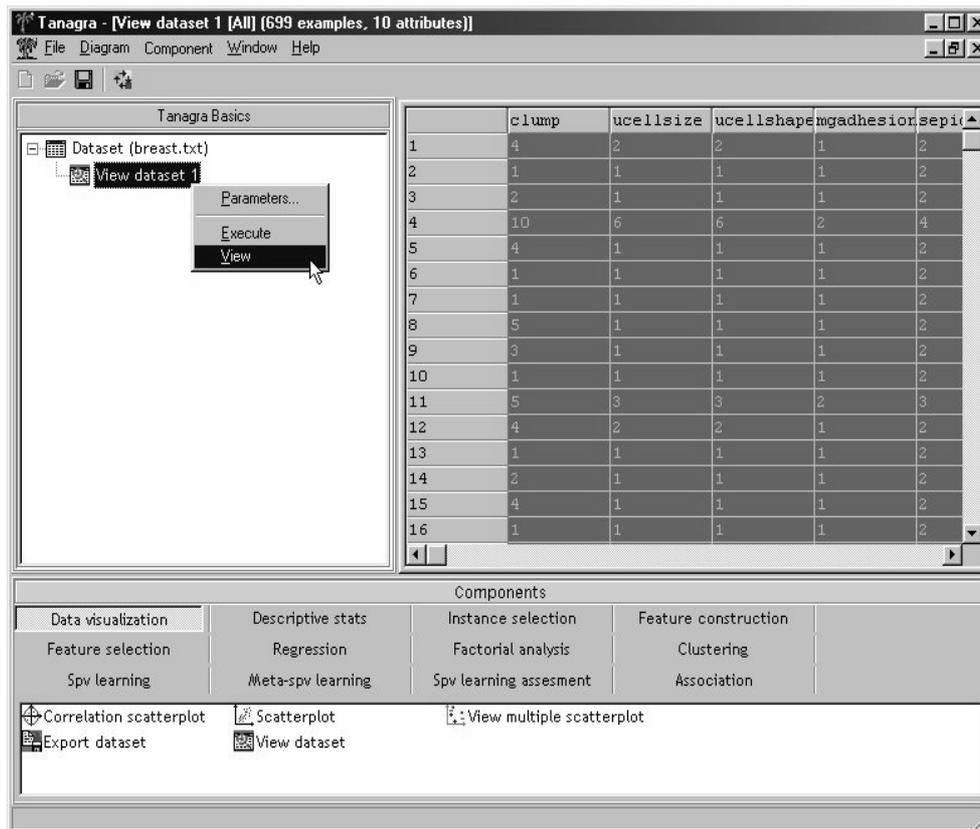
5 – Validate by clicking the OK button to start loading.

➤ Adding an operator to the diagram in order to visualize data

1 – Add a **View dataset** component to the diagram. To do this, click on the DATA VISUALIZATION tab of the components palette.

Drag and drop **View Dataset** from components palette to stream diagram, under the “Dataset” node (the node must appear to be selected).

2 – Then click on the “View dataset” node to select it (if not yet), and right-click on it to activate the popup menu: choose the *View* command. Data are displayed in the right frame.



## Getting some descriptive statistics

- Use and importance of the Define status operator

In TANAGRA you can build sequences of operations.

However, almost all operators require, before executing it, that you have defined the attributes to use, and how to use them (View dataset, that we've just used before, is an exception to this rule).

To avoid repetition of status definition for each operator, TANAGRA centralizes this declaration in the **Define status** component.

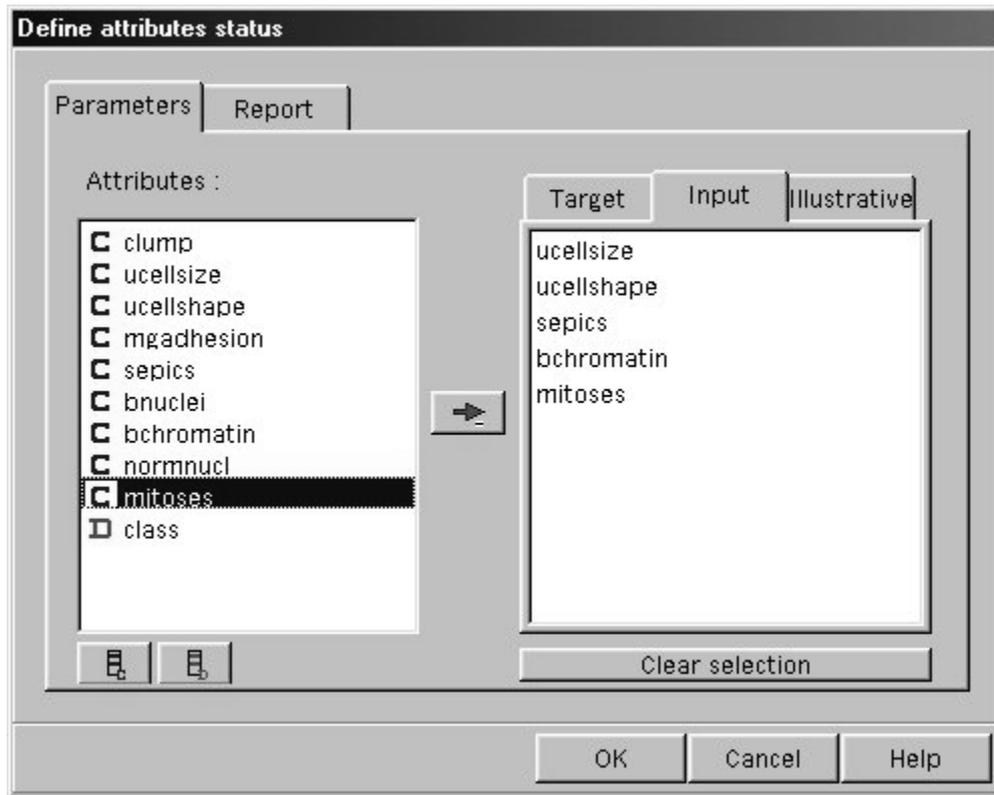
- Basic statistics on each attribute (min, max, average, standard deviation)

1 – Add a **Define Status** component (FEATURE SELECTION tab) to the diagram, under the « Dataset » node. (if, by error, you put it under the « View Dataset » node, you can delete it via the *Diagram / Delete component* menu)

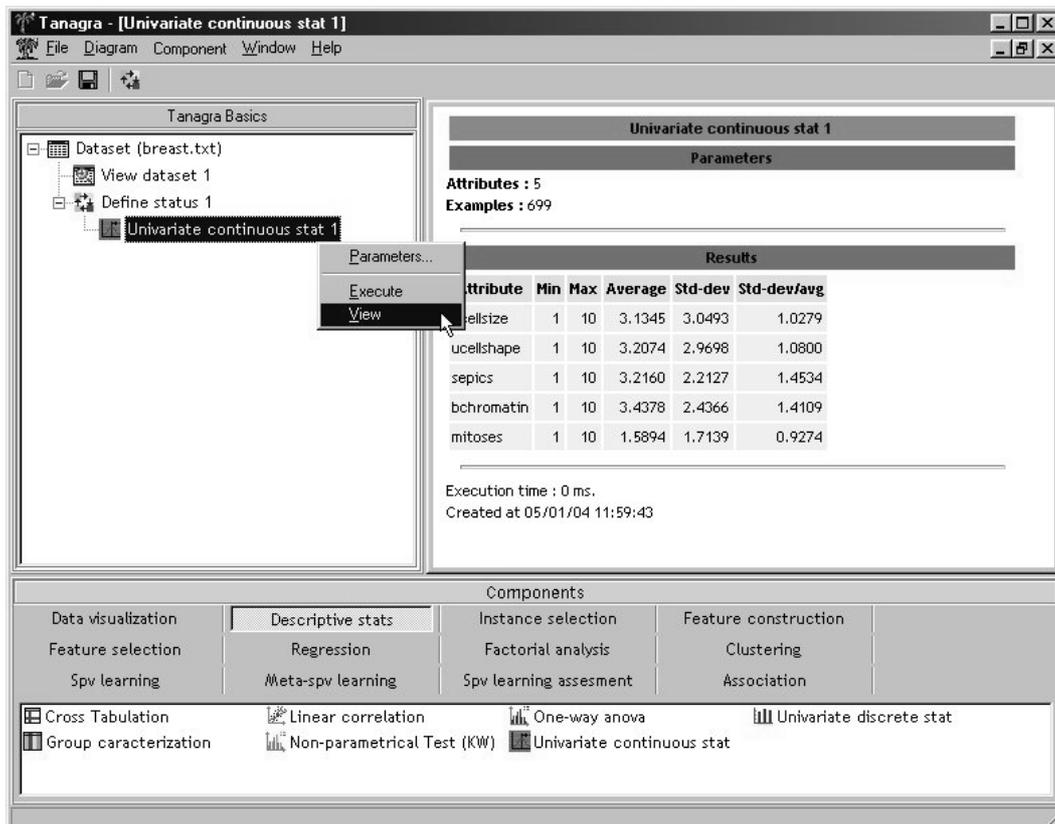
2 – Then click on the "Define status" node to select it, and activate its popup menu by clicking the mouse right button: choose the *Parameters...* Command.

3 – In the dialog box that appears, choose some continuous variables (these marked with a C blue letter): click on them (in the left list), then press the arrow button. They will act as input, since it is the active tab.

Validate with OK, the window closes.



4 – Add a **Univariate continuous stats** component (DESCRIPTIVE STATS tab) to the diagram, under the «Define status 1» node. In the popup menu, choose the *View* command. Descriptive statistics for the selected attributes are displayed in the right frame.



5 – Add another **Define status** component to the « Dataset » node, and select the discrete attribute « class ».

Add a **Univariate discrete stats** operator under the node « Define status 2», and choose the *View* command as previously. Below is the result you should obtain.

The screenshot shows the Tanagra software interface. The main window is titled "Tanagra - [Univariate discrete stat 1]". The project tree on the left shows a hierarchy: Dataset (breast.txt) > Define status 2 > Univariate discrete stat 1. The main results pane displays the following information:

**Univariate discrete stat 1**

**Parameters**

Attributes : 1  
Examples : 699

**Results**

Attribute	Gini	Distribution		
		Values	Count	Percent
class	0,4518	begin	458	65,52 %
		malignant	241	34,48 %

Execution time : 0 ms.  
Created at 05/01/04 15:59:31

The components palette at the bottom shows various statistical operators, with "Univariate discrete stat" highlighted.

- Statistics for each sub-population (comparing the characteristics of the women presenting -- or not -- a malignant tumor)

1 – Add another **Define status** operator to the « Dataset » node. Choose the *Parameters...* command in its popup menu. In the dialog window, select some continuous attributes as Input, and the discrete attribute as Target.

2 – Under this node add a **Group characterization** operator, and choose *View* in its popup menu.

**Tanagra - [Group characterization 1]**

File Diagram Component Window Help

Tanagra Basics

- Dataset (breast.txt)
  - View dataset 1
  - Define status 1
    - Univariate continuous stat 1
  - Define status 2
    - Univariate discrete stat 1
  - Define status 3
    - Group characterization 1

**Group characterization 1**

Parameters

Results

**Description of "class"**

class=begin				class=malignant			
Examples		458		Examples		241	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes				Continuous attributes			
mitoses	-11.2	1.06	1.59	ucellshape	21.6	6.56	3.21
sepics	-18.0	2.12	3.22	ucellsize	21.6	6.57	3.13
mgadhesion	-18.4	1.36	2.81	bnuclei	21.5	7.60	3.56
normnucl	-18.8	1.29	2.87	bchromatin	20.0	5.98	3.44
clump	-18.9	2.96	4.42	clump	18.9	7.20	4.42

Components

- Data visualization
- Descriptive stats
- Instance selection
- Feature construction
- Feature selection
- Regression
- Factorial analysis
- Clustering
- Spv learning
- Meta-spv learning
- Spv learning assesment
- Association

Cross Tabulation     Non-parametrical Test (KW)     Univariate discrete stat  
 Group characterization     One-way anova  
 Linear correlation     Univariate continuous stat

Inspecting these results, we notice that, on average, women with benign tumor present smaller values of "mitoses" (1.06, versus 1.59 for the complete population). On the other hand, the values of "ucellshape" attribute are, still on average, higher for women with malignant tumor (6.56 versus 3.21).