Présentation du didacticiel

Dans ce didacticiel, nous présentons les caractéristiques de base de l'interface de Tanagra, en analysant le fichier d'exemple « Breast.txt ».

Ce fichier, bien connu, est issu du domaine médical, et contient les caractéristiques physiologiques de cellules ponctionnées sur des patientes atteintes (ou non) du cancer du sein.

Vous apprendrez à utiliser les opérateurs suivants :

Onglet	Opérateur	Fonction
Data visualization	View dataset	Visualisation du fichier de données
Feature selection	Define status	Précision des variables à utiliser
Descriptive stats	Univariate continuous stat	Statistiques descriptives basiques pour variables de type continu
Descriptive stats	Univariate discrete stat	Statistiques descriptives basiques pour variables de type discret
Descriptive stats	Group caracterization	Statistiques par sous-population

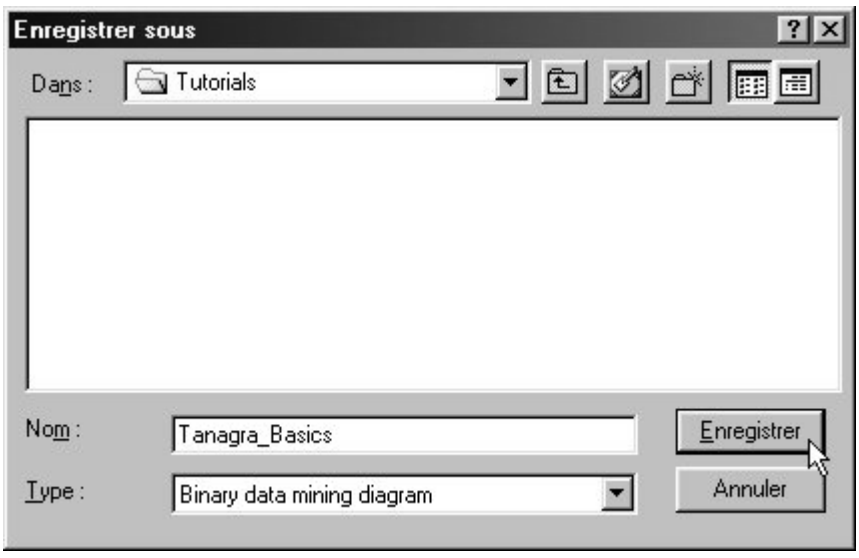
Importer et visualiser des données dans Tanagra

➢ Créer un nouveau diagramme

1 – Choisissez *File/New…* dans le menu de Tanagra.

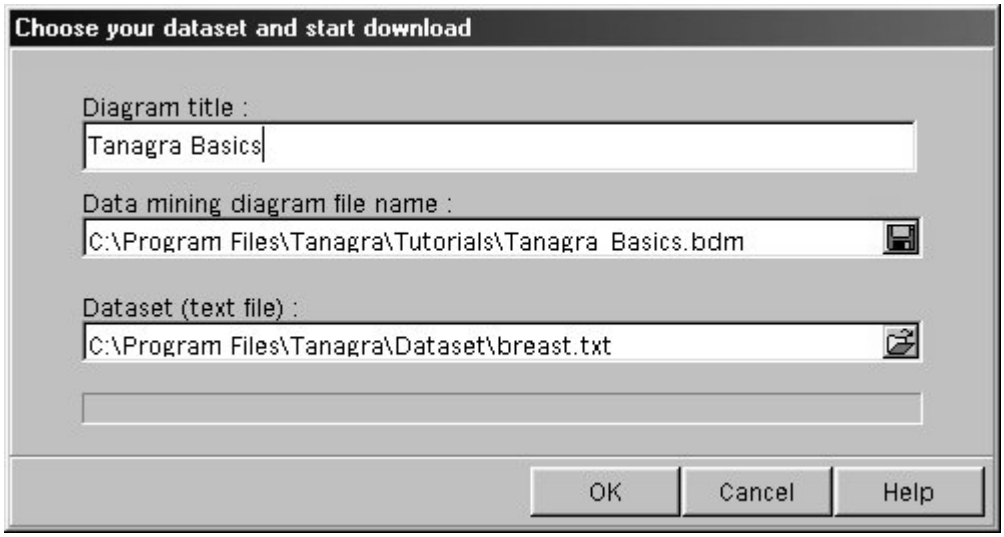
2 – Entrez un titre pour le diagramme : « Tanagra Basics ».

3 – Entrez le nom du fichier associé dans lequel votre travail sera sauvegardé (« Tanagra_Basics.bdm »). Cliquez avant sur le bouton pour parcourir le disque dur et vous placer dans le répertoire « …\Tanagra\Tutorials ».



4 – Sélectionnez le fichier texte contenant les données en cliquant sur l'icône suivante : 

Pour ce didacticiel, choisissez le fichier "breast.txt", situé dans le sous-répertoire de Tanagra « Dataset ».



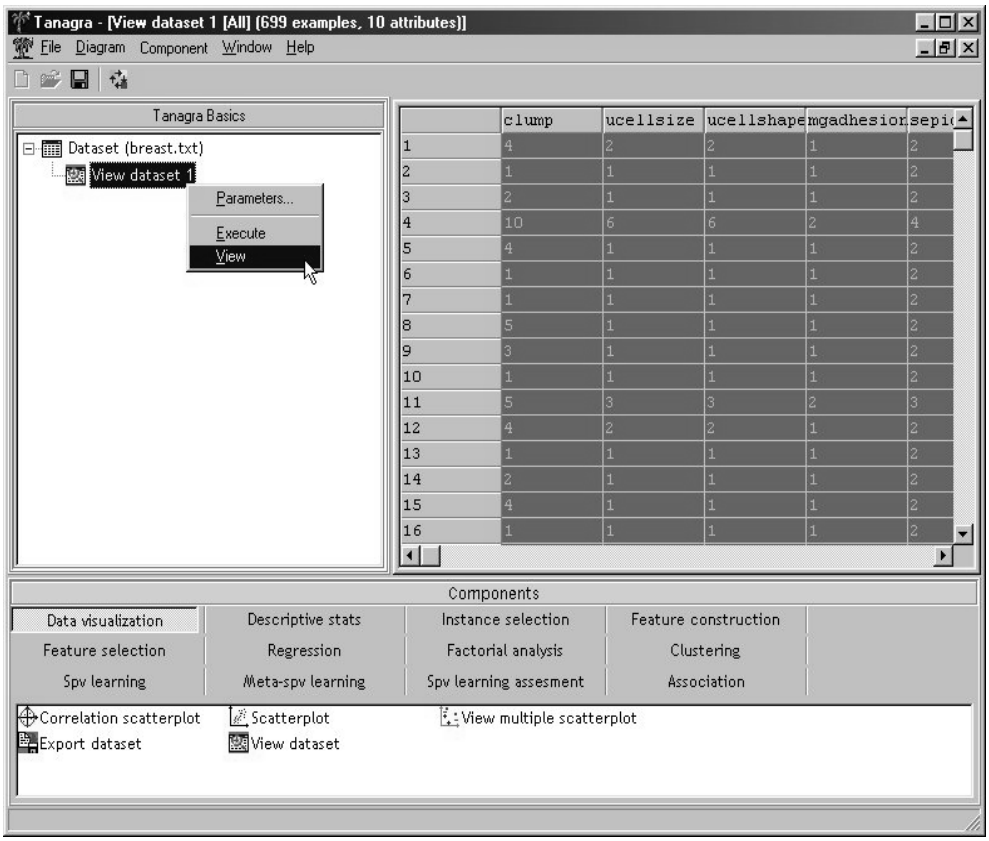
5 – Cliquez sur OK pour débuter l'importation.

➢ Ajouter un opérateur au diagramme pour visualiser les données

1 – Ajoutez un opérateur **View dataset** au diagramme. Pour cela, cliquez sur l'onglet DATA VISUALIZATION de la palette des opérateurs.

Positionnez la souris sur l'opérateur **View dataset** et, en maintenant le bouton gauche de la souris enfoncé, amenez-le sur le diagramme. Relâchez quand vous êtes au-dessus du nœud "Dataset" (il doit apparaître sélectionné comme ci-dessous).

2 – Cliquez ensuite sur le nœud "View dataset" pour le sélectionner (s'il ne l'est pas déjà), et faites apparaître son menu contextuel par clic droit : choisissez la commande *View*. Les données apparaissent dans le cadre de droite.



Obtenir des informations sur les données (statistiques descriptives)

➢ Note sur l'opérateur Define status

Tanagra permet de constituer des enchaînements d'opérateurs.

Or presque tous les opérateurs requièrent, avant de les exécuter, qu'on ait défini les variables à utiliser et leur rôle (View dataset, que nous venons d'utiliser, est justement une exception).

Pour ne pas avoir à spécifier le statut des variables au niveau de chaque opérateur, Tanagra centralise cette déclaration dans l'opérateur **Define status**.

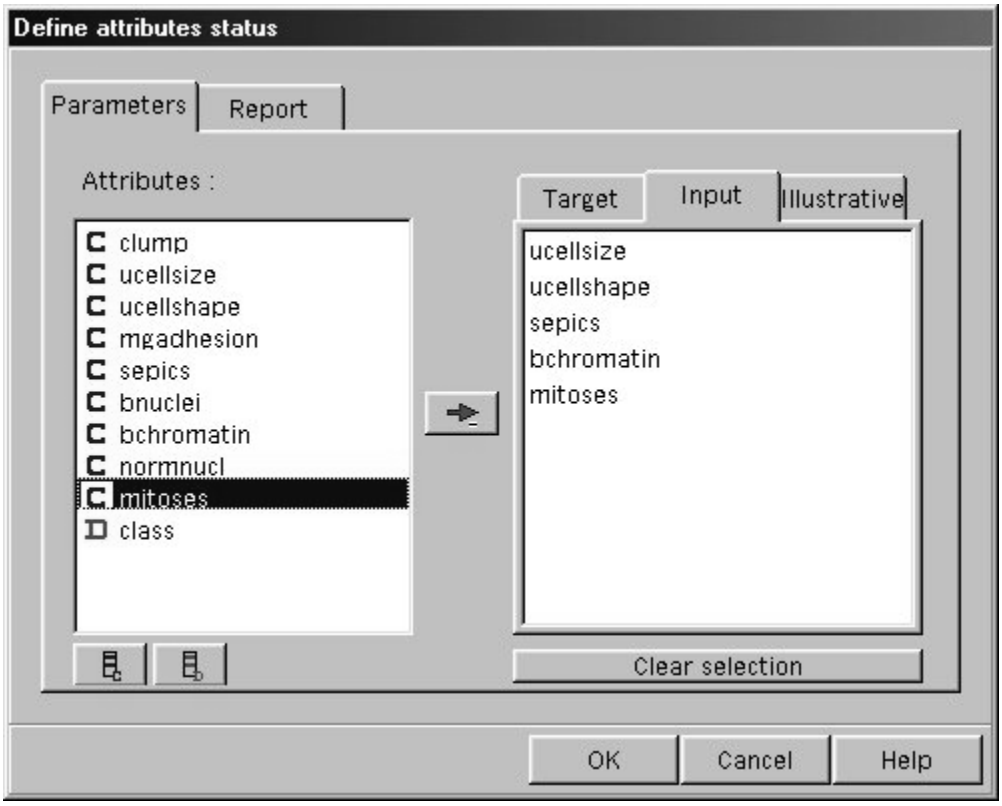
➢ Statistiques de base sur chaque variable (min, max, moyenne, écart-type)

1 – Ajoutez un opérateur **Define Status** (onglet FEATURE SELECTION) au diagramme, en l'accrochant au nœud « Dataset ». (si vous l'accrochez par erreur au nœud « View Dataset », supprimez-le via le menu *Diagram / Delete component*)

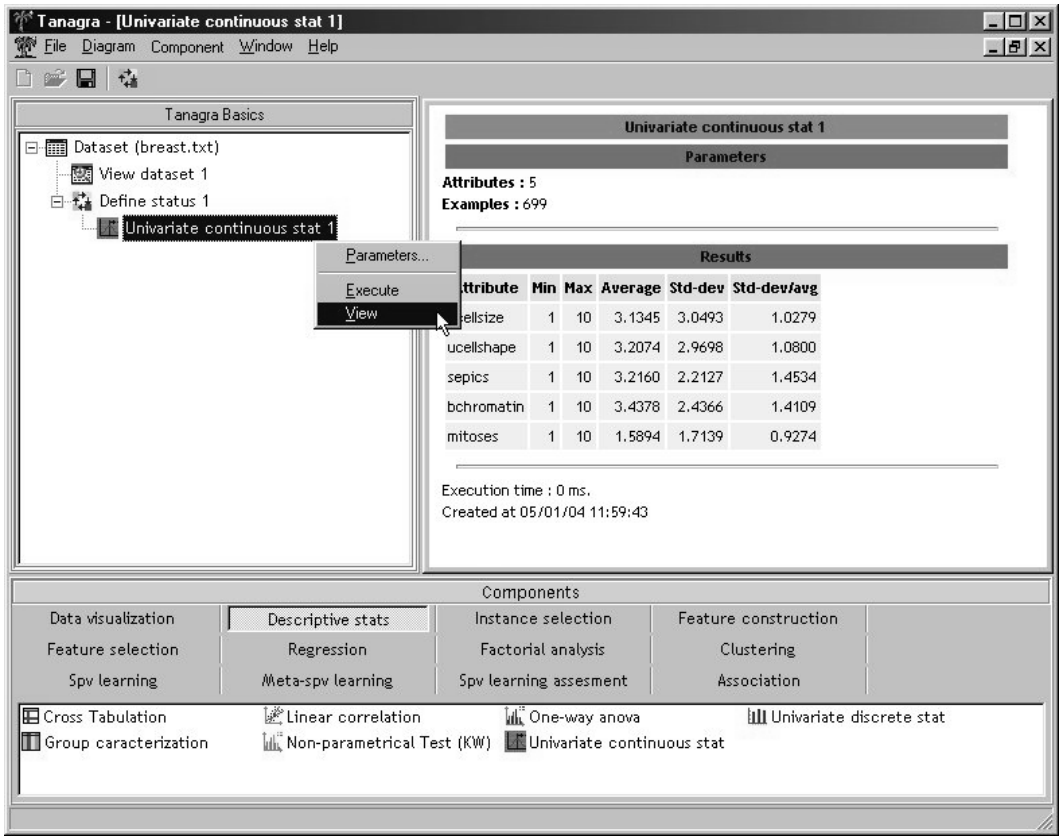
2 – Cliquez ensuite sur le nœud "Define status" pour le sélectionner, et faites apparaître son menu contextuel par clic droit : choisissez la commande *Parameters…*.

3 – Dans la fenêtre de dialogue qui apparaît, sélectionnez quelques variables continues (marquées de la lettre bleue C) : cliquez dessus dans la liste à gauche, puis appuyez sur le bouton flèche. Par défaut, elles auront un rôle d' « input », puisque c'est l'onglet actif.

Appuyez sur OK pour valider et fermer cette fenêtre.

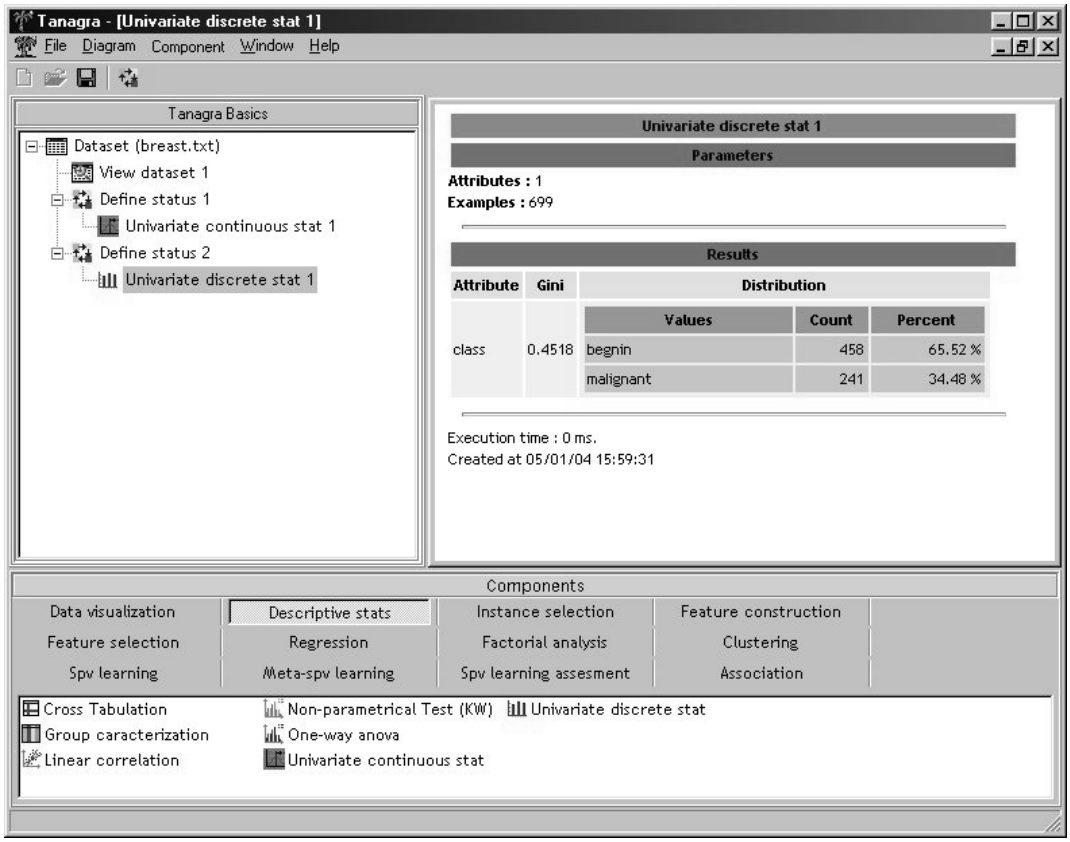


4 – Ajoutez un opérateur **Univariate continuous stats** (onglet DESCRIPTIVE STATS) au diagramme, sous le nœud «Define status 1». Dans le menu contextuel, choisissez la commande *View*. Les statistiques descriptives de ces variables apparaissent dans le cadre de droite.



5 – Ajoutez un autre opérateur **Define status** au nœud « Dataset », et sélectionnez cette fois-ci la variable discrète « class ».

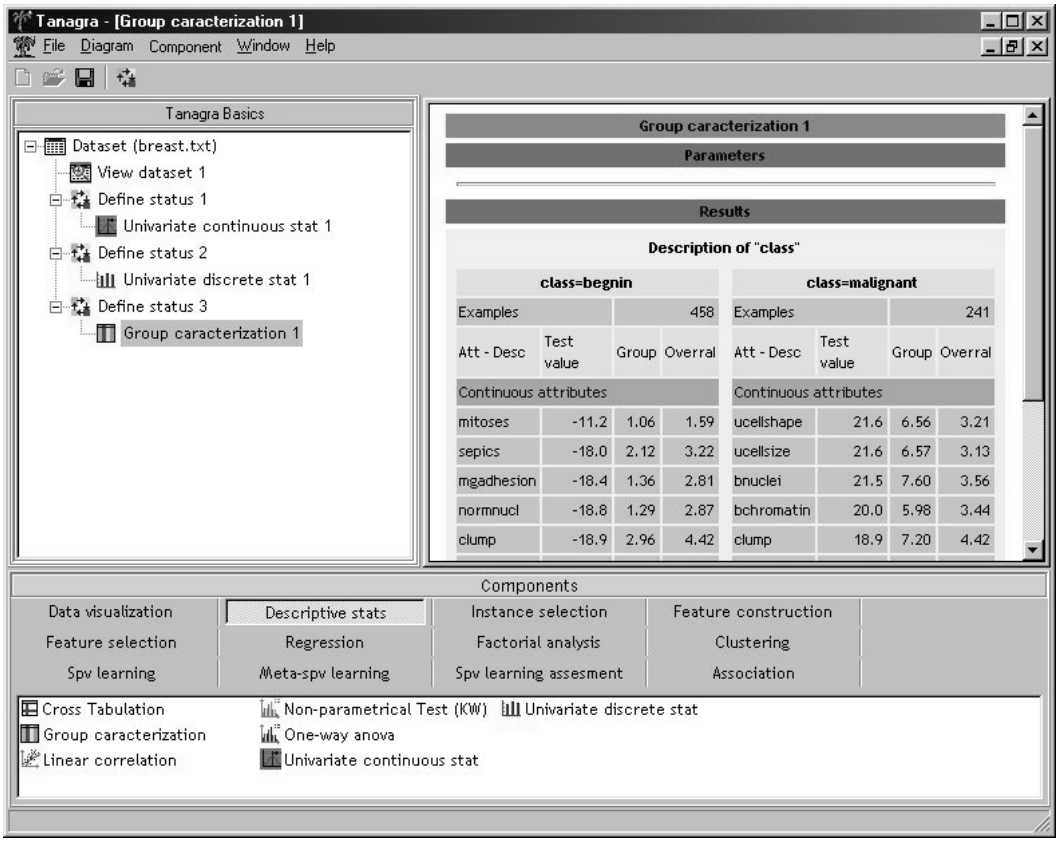
Ajoutez un opérateur **Univariate discrete stats** en-dessous du nœud « Define status », et choisissez la commande *View* comme précédemment. Vous devez obtenir le résultat ci-dessous.



- Statistiques par sous-population (comparaison des caractéristiques des patientes atteintes du cancer du sein et de celles des patientes non atteintes)

1 –- Ajoutez un autre opérateur **Define status** au nœud « Dataset ». Choisissez la commande *Parameters…* dans son menu contextuel. Dans la fenêtre de dialogue, choisissez quelques variables continues en Input, et la variable discrète en Target.

2 – Sous ce nœud ajoutez un opérateur **Group caracterization**, et dans son menu contextuel choisissez *View*.



Au vu des résultats, on constate qu'en moyenne les patientes non atteintes du cancer du sein présentent des valeurs de mitoses plus petites (1.06, contre 1.59 sur l'ensemble de la population). A l'inverse, pour les patientes atteintes du cancer, les valeurs de la variable ucellshape sont en moyenne plus élevées (6.56 contre 3.21).