Objectif

Montrer comment importer directement dans TANAGRA un fichier au format EXCEL (version 97 & 2000).

Le principal avantage de cette procédure est qu'il est possible de modifier à la volée le fichier source *sans avoir à reconstruire le diagramme de traitements (!)*.

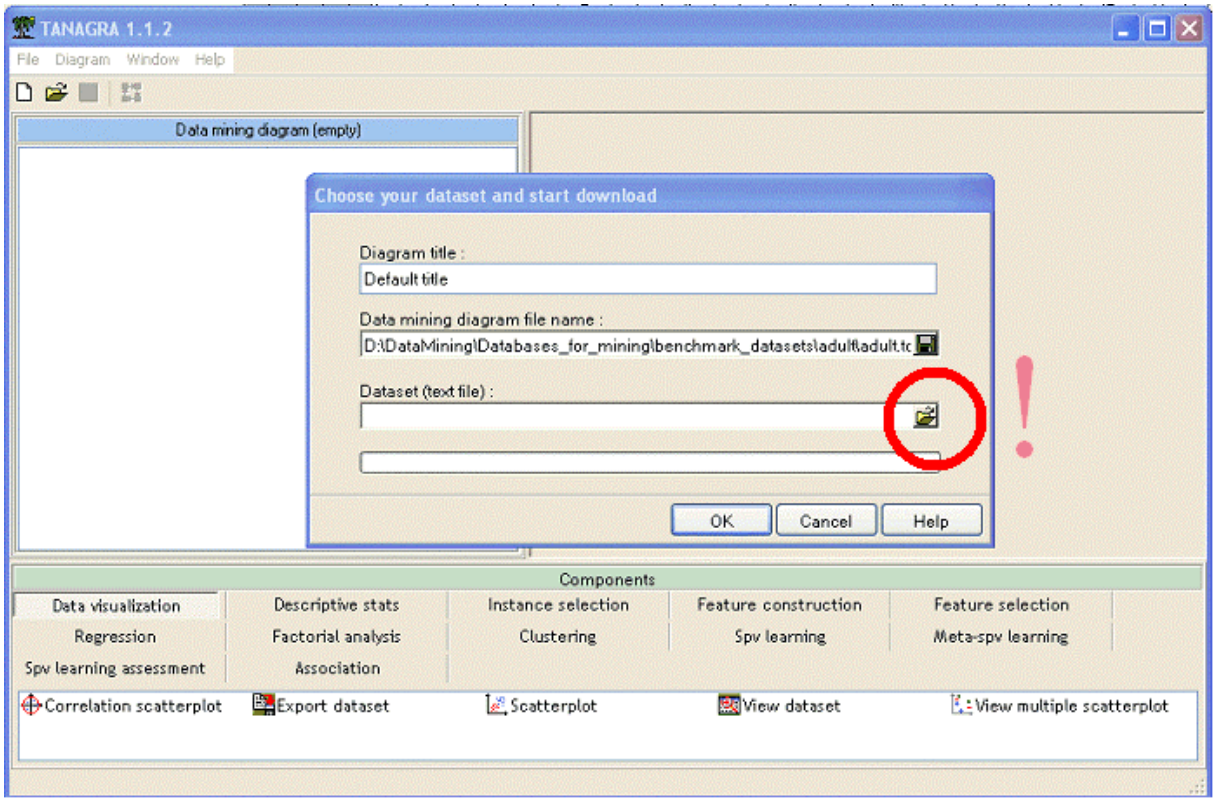
Fichier

Nous allons utiliser le fichier ADULT, il contient 48842 individus et 15 variables. L'objectif est de caractériser les individus ayant un revenu élevé (variable « class » : supérieur à 50K$ ou non)

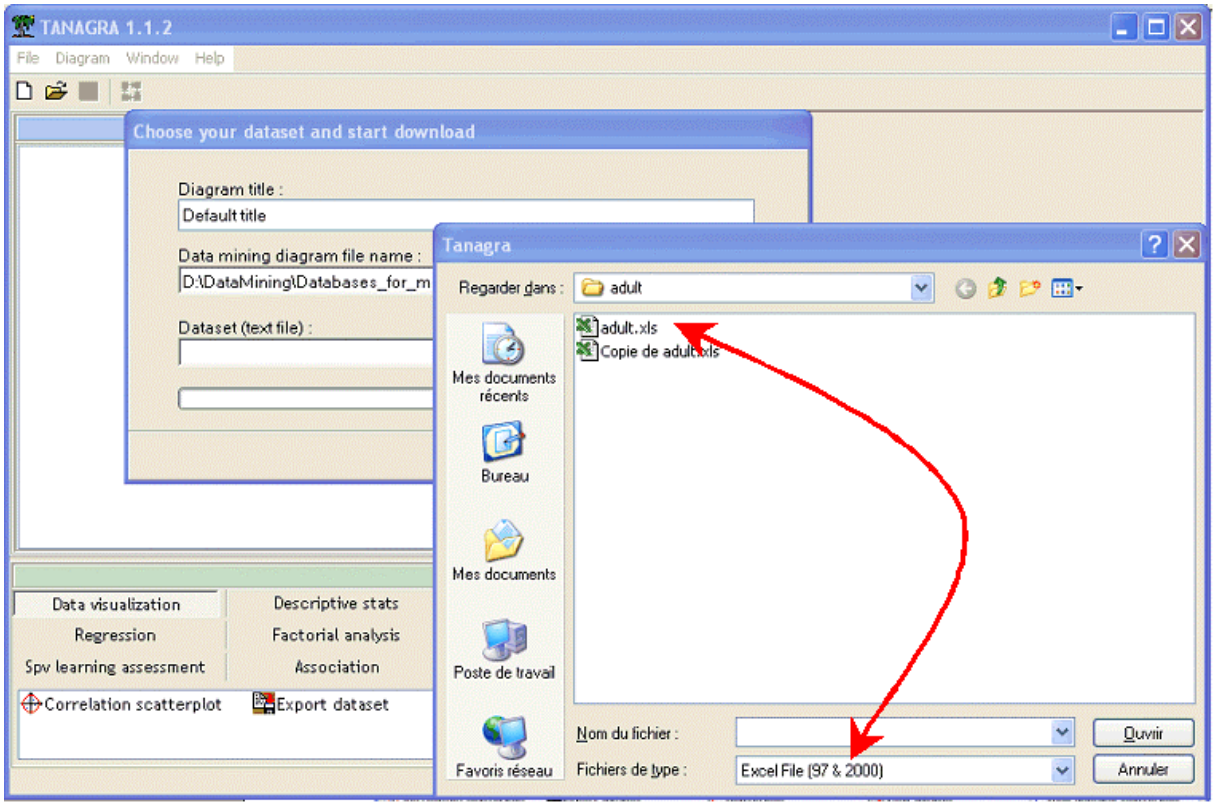
Manipuler un fichier EXCEL

Importer le fichier EXCEL

La première étape consiste à importer les données. Il faut donc construire un nouveau diagramme de traitements selon la procédure habituelle (File / New), puis choisir le fichier de données.



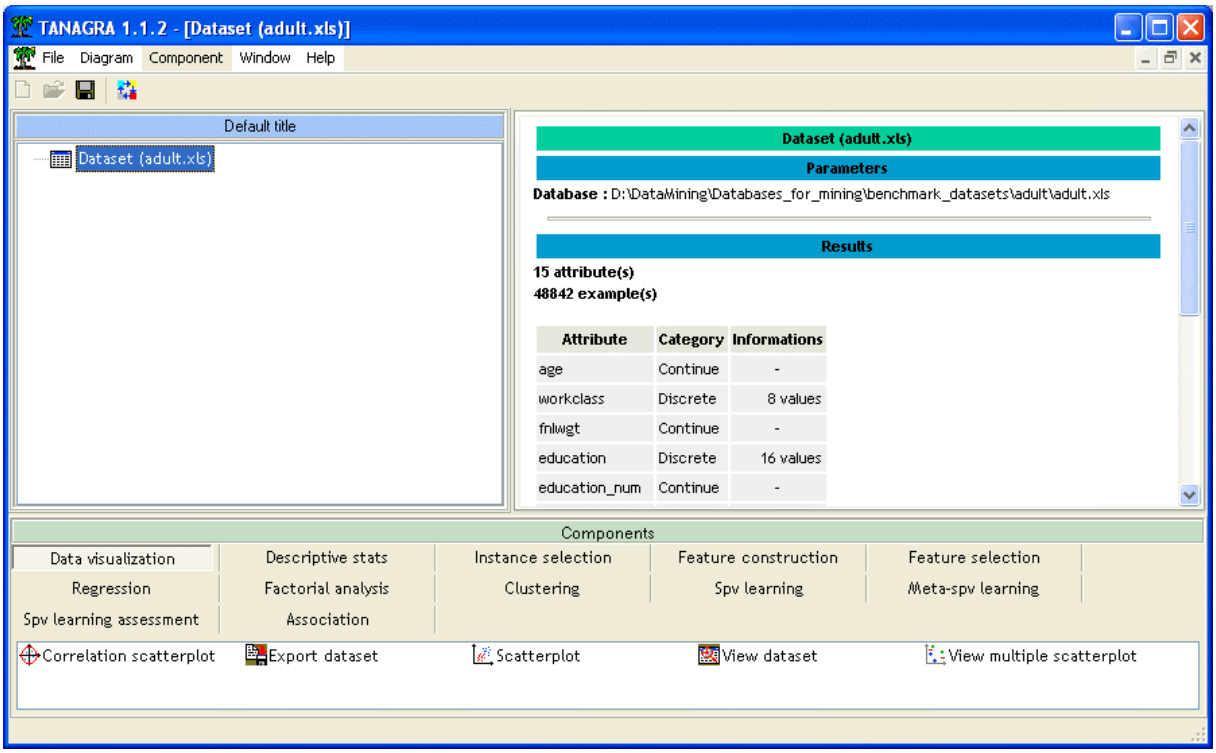
Dans la boîte de dialogue permettant de sélectionner le fichier, plusieurs formats sont maintenant disponibles, notamment le format EXCEL. Sélectionnez-le, puis chargez le fichier ADULT.XLS.



Plusieurs précautions sont nécessaires pour importer correctement un fichier au format EXCEL :
(1) les seuls formats acceptés correspondent aux versions 97 & 2000 de EXCEL ;
(2) si le classeur contient plusieurs feuilles, les données doivent être situées dans la première feuille ;
(3) les données doivent être alignées en haut et à gauche dans la feuille de calcul càd elles doivent commencer aux coordonnées A1 ;
(4) la première ligne des données correspond au nom des variables ;
(5) il n'y a pas d'identifiant des lignes ;
(6) la plage de données ne doit pas contenir des cellules / lignes / colonnes vides.

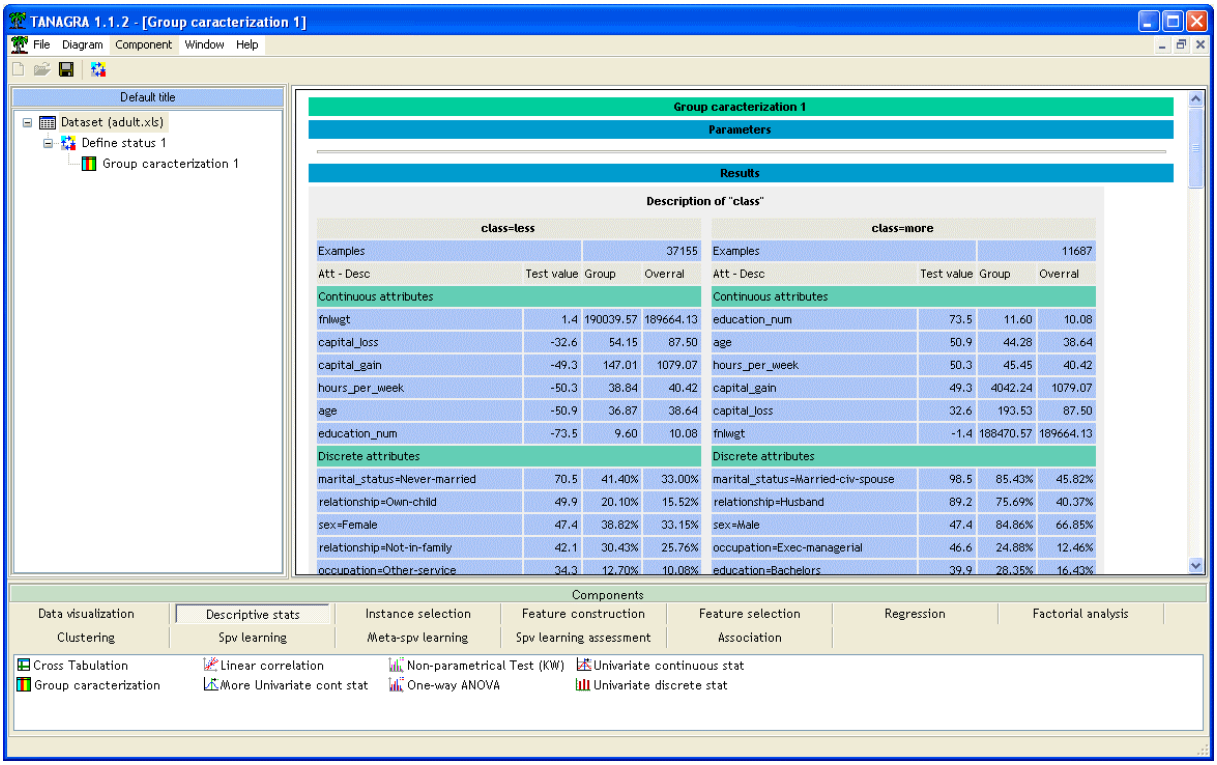
Il est à noter que le nombre maximum de variables et d'individus que l'on peut traiter est limité par les capacités d'EXCEL et non de TANAGRA : 65534 observations et 256 variables.

Voici les résultats de l'importation, il est relativement rapide (7 secondes sur un P4 à 3 Ghz).



Caractériser la variable « CLASS »

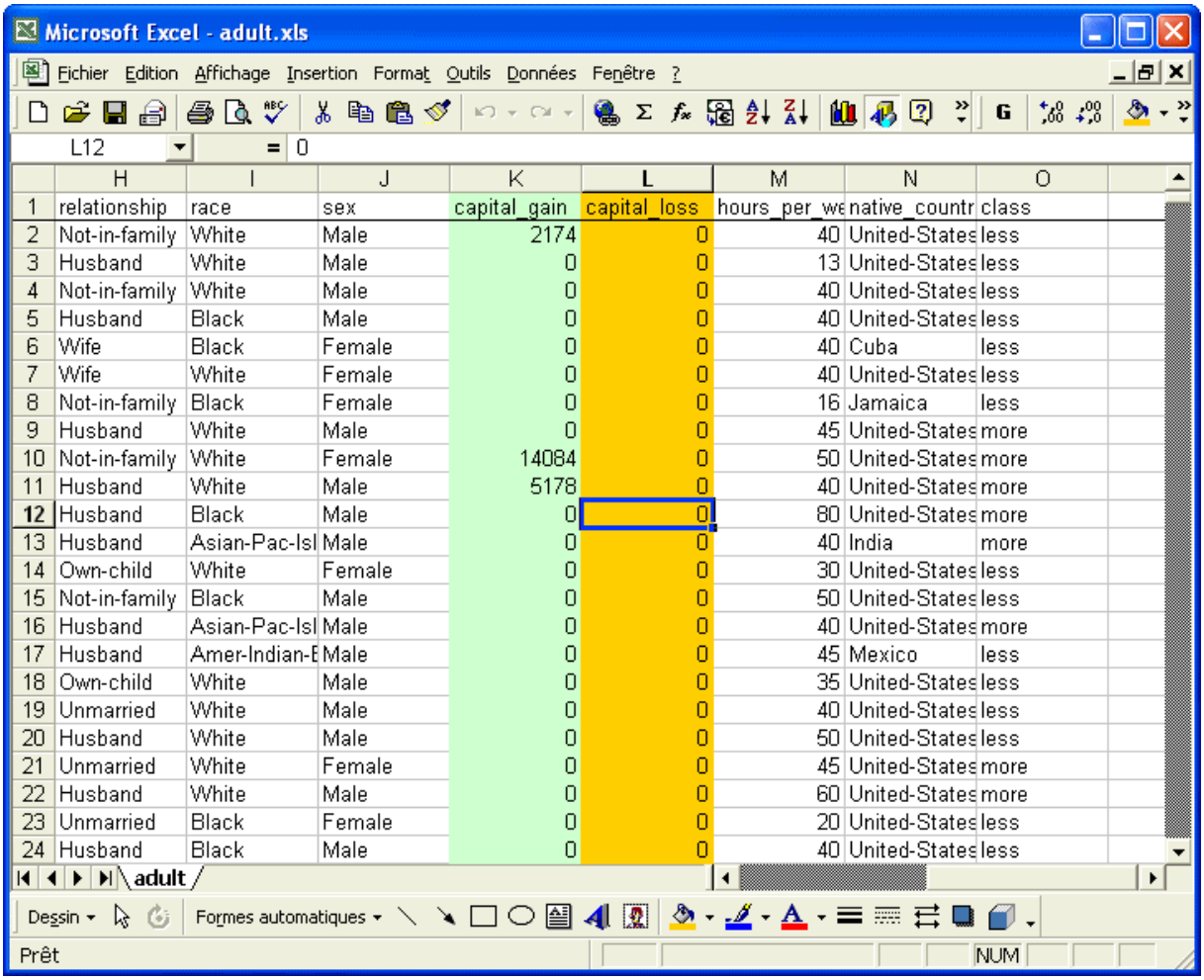
Nous allons essayer de prédire la variable classe à partir des autres variables. Nous utilisons pour cela le composant GROUP CHARACTERIZATION après avoir placé CLASS en TARGET et les autres attributs en INPUT. **Enregistrer le diagramme de traitements correspondant.**



Nous constatons que les personnes à revenus élevés sont surtout des hommes mariés assez âgés exerçant une activité managériale, ils ont un haut niveau d'éducation et travaillent beaucoup (sic !).

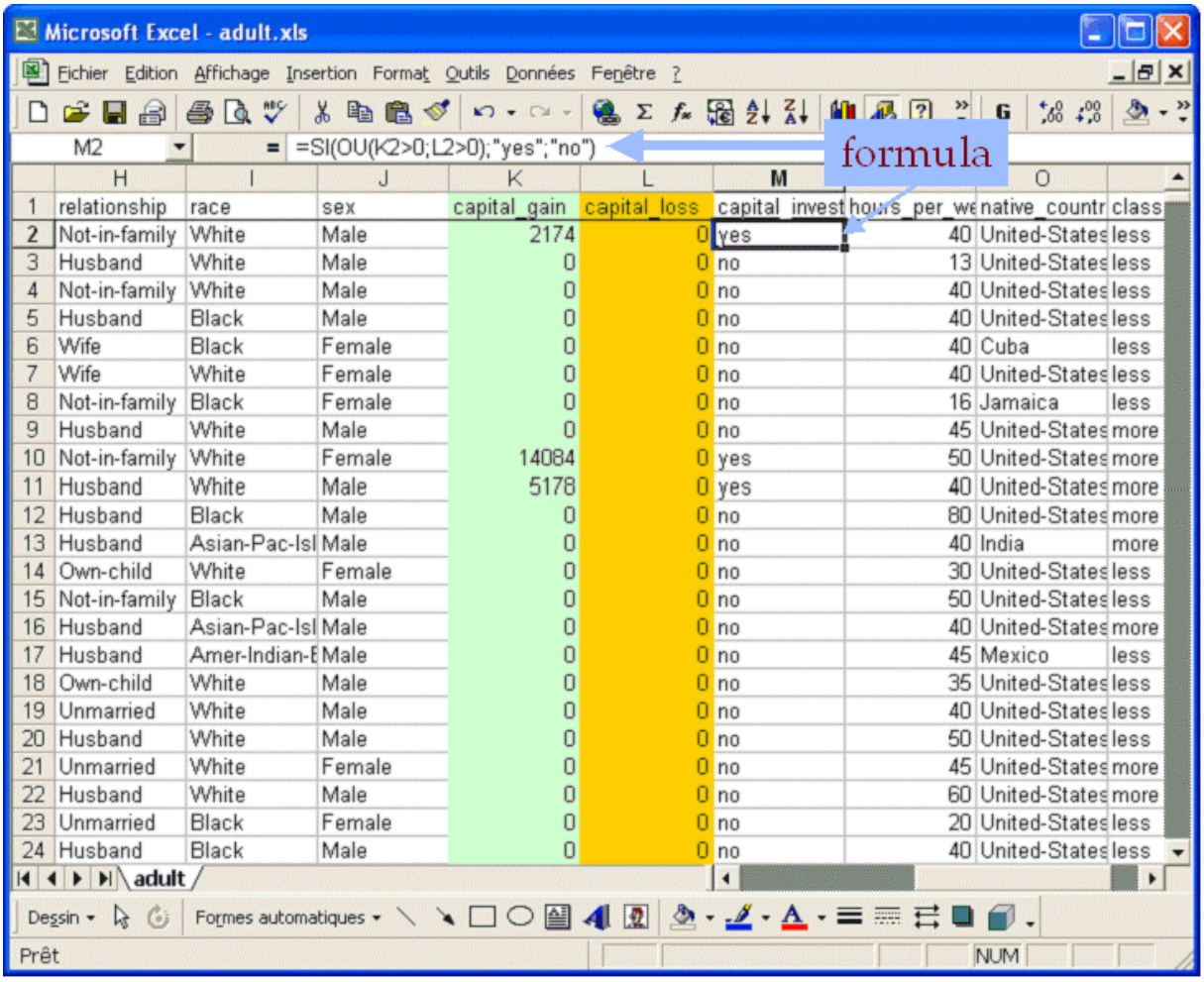
Ce sont également des personnes qui ont eu à la fois des gains et des pertes en capital plus élevés que la moyenne. La question est de savoir comment interpréter cette dernière information ?

En revenant sur le fichier initial (ADULT.XLS), nous constatons que ces deux variables regroupent deux types d'informations : si la personne ne place pas sont argent, le gain (la perte) est égal à zéro ; et s'ils placent leur argent, la variable indique le montant du gain (de la perte). De plus, un individu ayant réalisé un gain ne peut pas réaliser de perte, et inversement.



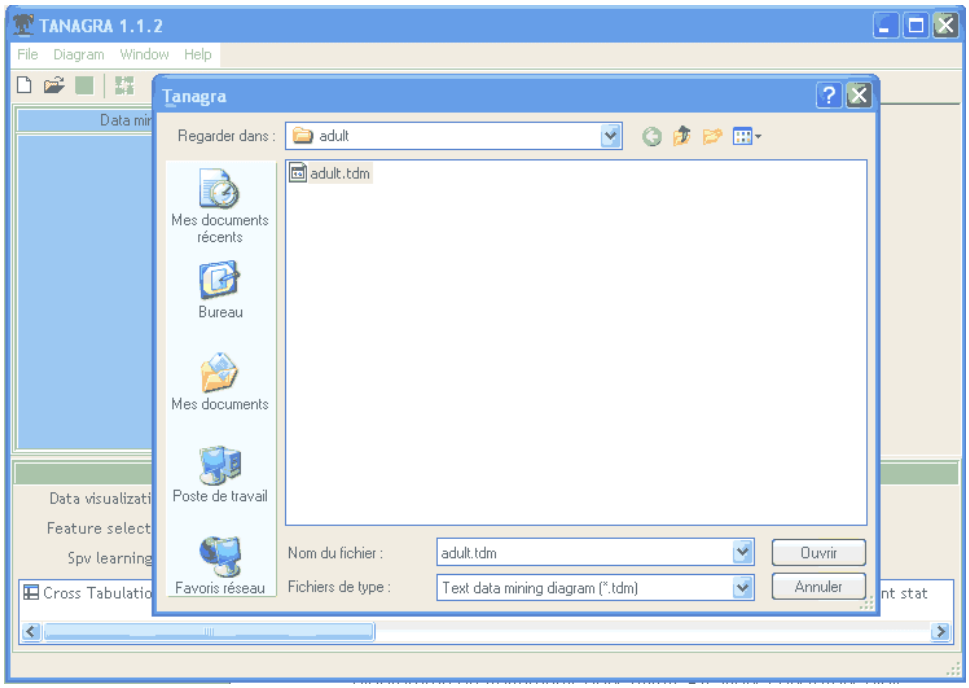
	H	I	J	K	L	M	N	O
1	relationship	race	sex	capital_gain	capital_loss	hours_per_we	native_countr	class
2	Not-in-family	White	Male	2174	0	40	United-States	less
3	Husband	White	Male	0	0	13	United-States	less
4	Not-in-family	White	Male	0	0	40	United-States	less
5	Husband	Black	Male	0	0	40	United-States	less
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

Nous allons donc créer une variable indicatrice permettant de savoir si l'individu place son argent ou pas.

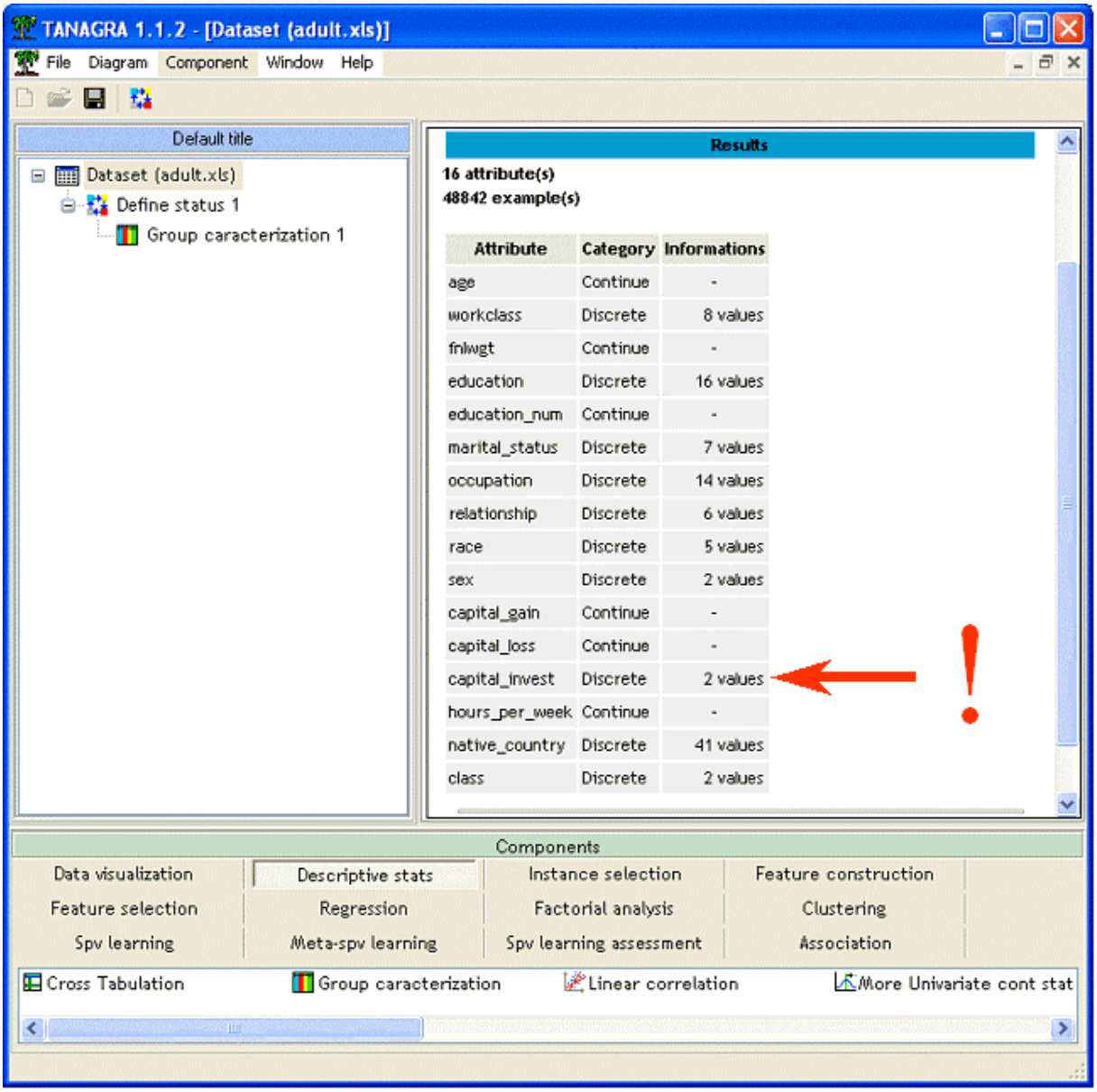


	H	I	J	K	L	M	O	
1	relationship	race	sex	capital_gain	capital_loss	capital_invest	native_countr	class
2	Not-in-family	White	Male	2174	0	yes	United-States	less
3	Husband	White	Male	0	0	no	United-States	less
4	Not-in-family	White	Male	0	0	no	United-States	less
5	Husband	Black	Male	0	0	no	United-States	less
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

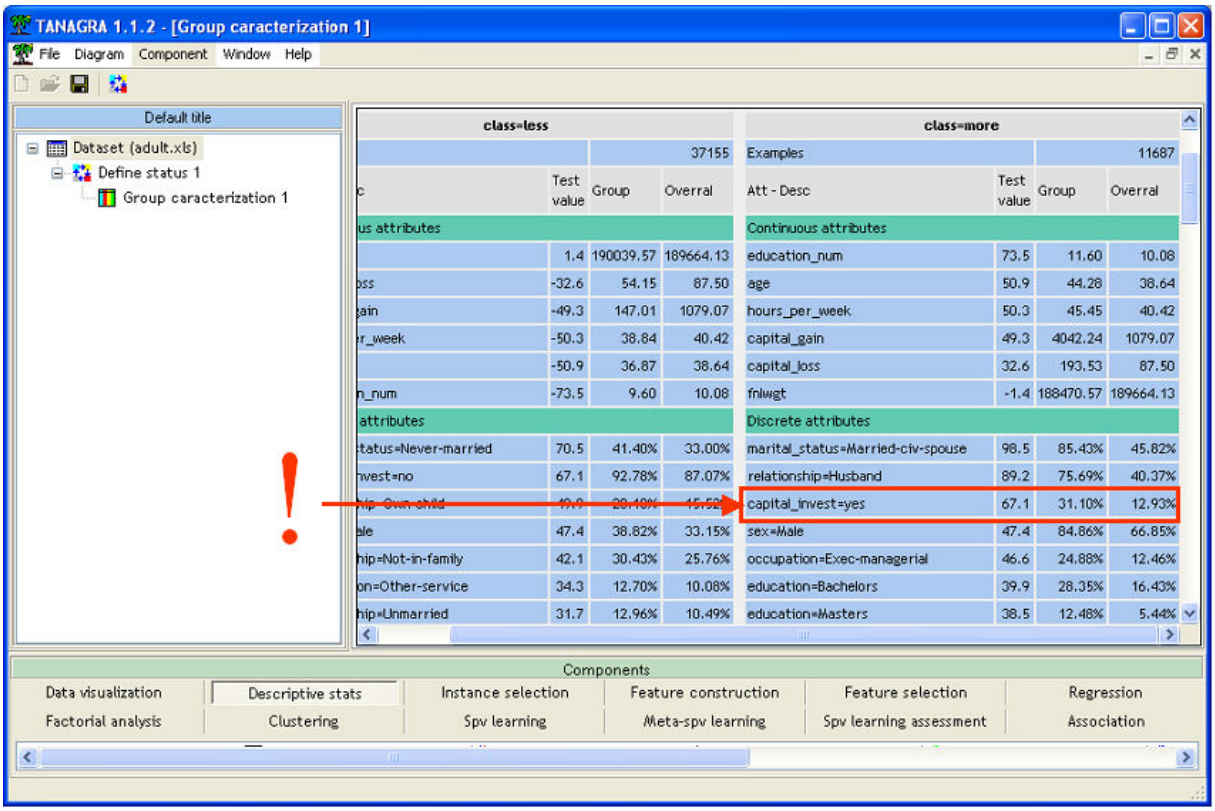
Après avoir enregistré le fichier modifié puis fermé EXCEL, il suffit de fermer puis de ré-ouvrir le diagramme de traitements dans TANAGRA.



Nous constatons alors :
(1) que le diagramme est bien conservé ;
(2) que les données maintenant sont enrichies de la nouvelle colonne que nous venons de créer.



Dans DEFINE STATUS, ajouter la nouvelle variable parmi les INPUT. Nous constatons dès lors que la variable CAPITAL indique surtout un comportement vis à vis de son épargne : les personnes ayant un revenu élevé ont tendance à placer leur capital (12.9% des personnes placent leur argent, ils sont 31% à le faire parmi les personnes à revenus élevés).


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
37155
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%
class=more
Examples 11687
1.4 190039.57 189664.13
-32.6 54.15 87.50
-49.3 147.01 1079.07
-50.3 38.84 40.42
-50.9 36.87 38.64
-73.5 9.60 10.08
70.5 41.40% 33.00%
67.1 92.78% 87.07%
47.4 38.82% 33.15%
42.1 30.43% 25.76%
34.3 12.70% 10.08%
31.7 12.96% 10.49%
Components
Data visualization
Descriptive stats
Instance selection
Feature construction
Feature selection
Regression
Factorial analysis
Clustering
Spv learning
Meta-spv learning
Spv learning assessment
Association