1 Objectif

Régression – Déploiement de modèles.

Le déploiement est une des principales finalités du Data Mining. Il s'agit d'appliquer les modèles sur de nouveaux individus de la population. En apprentissage supervisé, il s'agit de leur attribuer leur classe d'appartenance (ex. http://tutoriels-data-mining.blogspot.com/2008/03/dploiement-de-modles-avec-tanagra.html) ; en apprentissage non supervisé, l'objectif est de les associer à un groupe qui leur serait le plus similaire (ex. http://tutoriels-data-mining.blogspot.com/2008/10/classification-automatique-dploiement.html). Concernant la régression, appliquer le modèle sur des nouveaux individus consiste à prédire la valeur de la variable dépendante quantitative (variable endogène, variable cible) à partir de leur description c.-à-d. les valeurs prises par les variables indépendantes (variables exogènes).

L'opération est simple lorsqu'il s'agit d'implémenter une régression linéaire multiple ou une régression PLS. Nous récupérons les coefficients du modèle, nous les appliquons sur la description des nouveaux individus à étiqueter. L'affaire devient compliquée lorsque nous souhaitons manipuler des modèles plus complexes, soit parce qu'issus d'enchaînements d'opérations (ex. analyse factorielle + régression sur axes), soit parce que nous ne disposons pas d'une expression explicite simple du modèle (Support Vector Regression avec un noyau non linéaire). Il est donc primordial que le logiciel qui a servi à la construction des modèles puisse se charger lui-même du déploiement.

Avec Tanagra, il est possible de déployer facilement les modèles dans le cadre de la régression, même lorsqu'ils sont le fruit d'une succession d'opérations. Il faut simplement préparer le fichier de données d'une manière particulière.

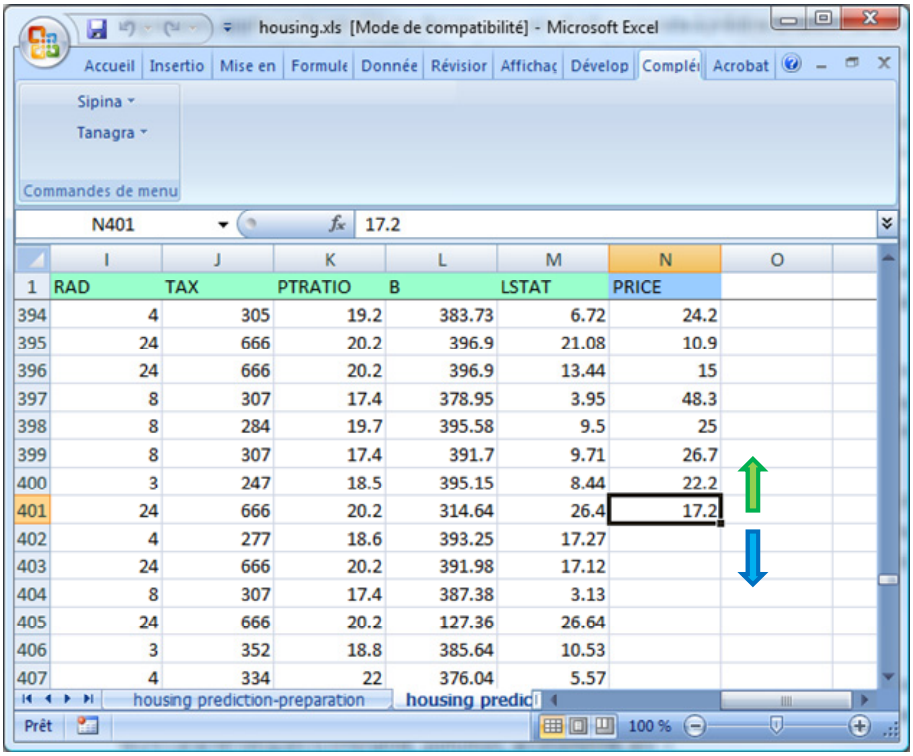
Dans ce didacticiel, nous montrons comment organiser efficacement le fichier pour faciliter le déploiement. Par la suite, nous apprenons plusieurs modèles prédictifs (régression linéaire multiple, régression PLS, support vector régression avec un noyau RBF, arbre de régression, régression sur axes factoriels), que nous appliquons sur les nouvelles observations à étiqueter. Nous exportons les prédictions dans un fichier au format Excel. Enfin, nous vérifions leur cohérence. L'idée est d'identifier les techniques qui produisent des prédictions similaires.

2 Données – Préparation du fichier

Nous utilisons le fichier HOUSING (http://archive.ics.uci.edu/ml/datasets/Housing). On souhaite prédire la valeur médiane des logements (PRICE) dans différentes zones géographiques à partir de leurs caractéristiques (criminalité, pollution, accessibilité, etc.).

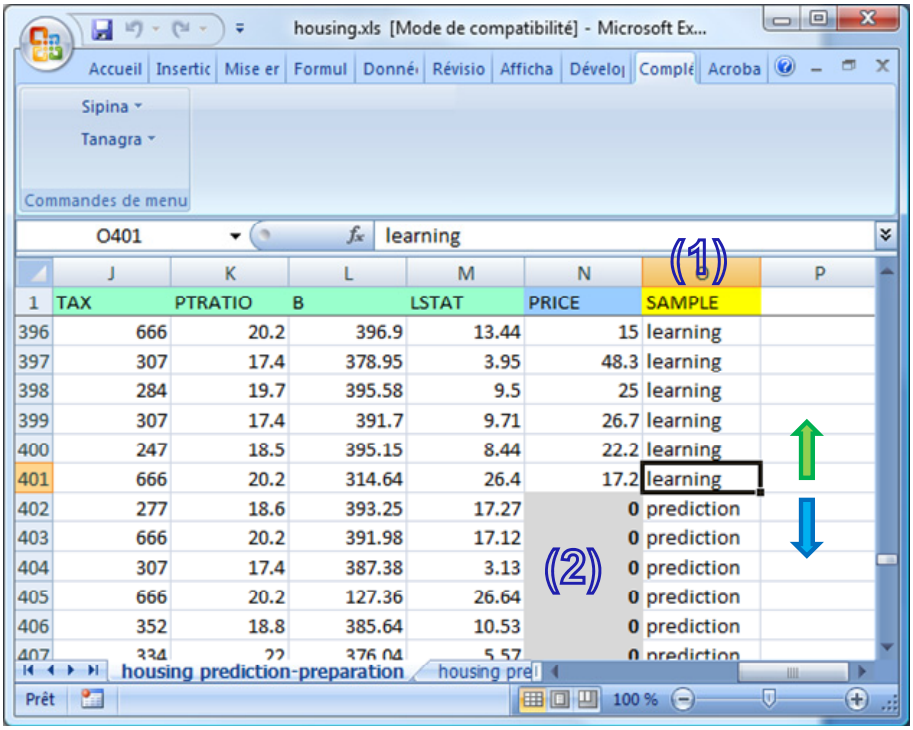
Nous disposons de 400 observations pour la construction des modèles, que nous souhaitons déployer sur 106 observations non étiquetées (c.-à-d. pour lesquelles nous disposons de la description mais pas des valeurs de la variable cible). Concrètement, il y a 506 lignes dans notre fichier, mais seules les 400 premières lignes de PRICE (la variable dépendante) sont renseignées.

Cette organisation pose problème. Tanagra comprend les cellules vides de la colonne PRICE comme des données manquantes… qu'il ne sait pas gérer. De fait, à l'importation, le fichier sera tronqué aux 400 premières lignes, les autres observations sur lesquelles nous souhaitons appliquer les modèles ne seront pas disponibles. Ce qui est fâcheux.



	I	J	K	L	M	N	O
1	RAD	TAX	PTRATIO	B	LSTAT	PRICE	
394	4	305	19.2	383.73	6.72	24.2	
395	24	666	20.2	396.9	21.08	10.9	
396	24	666	20.2	396.9	13.44	15	
397	8	307	17.4	378.95	3.95	48.3	
398	8	284	19.7	395.58	9.5	25	
399	8	307	17.4	391.7	9.71	26.7	
400	3	247	18.5	395.15	8.44	22.2	
401	24	666	20.2	314.64	26.4	17.2	
402	4	277	18.6	393.25	17.27		
403	24	666	20.2	391.98	17.12		
404	8	307	17.4	387.38	3.13		
405	24	666	20.2	127.36	26.64		
406	3	352	18.8	385.64	10.53		
407	4	334	22	376.04	5.57		

Pour que Tanagra puisse appréhender correctement les données, nous devons modifier l'organisation du fichier. Tout d'abord, nous rajoutons la colonne SAMPLE pour indiquer le rôle des observations (1) : « learning » pour celles utilisées lors de la construction des modèles ; « prediction » pour celles que nous souhaitons étiqueter. Ensuite, nous insérons des valeurs par défaut dans la colonne de la variable cible non renseignée (2). La valeur **0** convient parfaitement. Cette opération peut paraître étrange. Elle sert uniquement à contourner le problème des données manquantes.



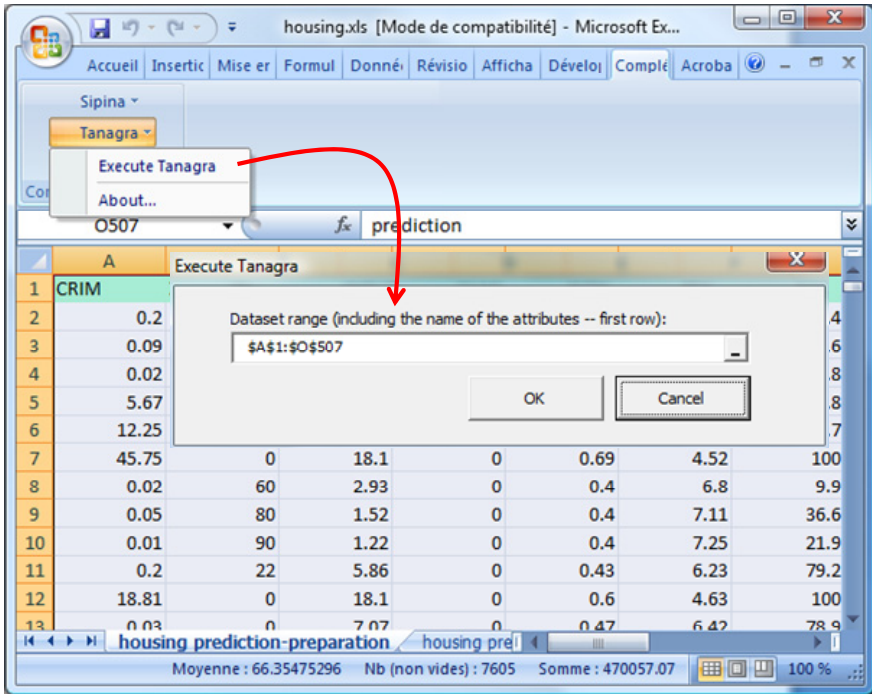
	J	K	L	M	N	O	P
1	TAX	PTRATIO	B	LSTAT	PRICE	SAMPLE	
396	666	20.2	396.9	13.44	15	learning	
397	307	17.4	378.95	3.95	48.3	learning	
398	284	19.7	395.58	9.5	25	learning	
399	307	17.4	391.7	9.71	26.7	learning	
400	247	18.5	395.15	8.44	22.2	learning	
401	666	20.2	314.64	26.4	17.2	learning	
402	277	18.6	393.25	17.27	0	prediction	
403	666	20.2	391.98	17.12	0	prediction	
404	307	17.4	387.38	3.13	0	prediction	
405	666	20.2	127.36	26.64	0	prediction	
406	352	18.8	385.64	10.53	0	prediction	
407	334	22	376.04	5.57	0	prediction	

Nous pouvons maintenant importer les données dans Tanagra.

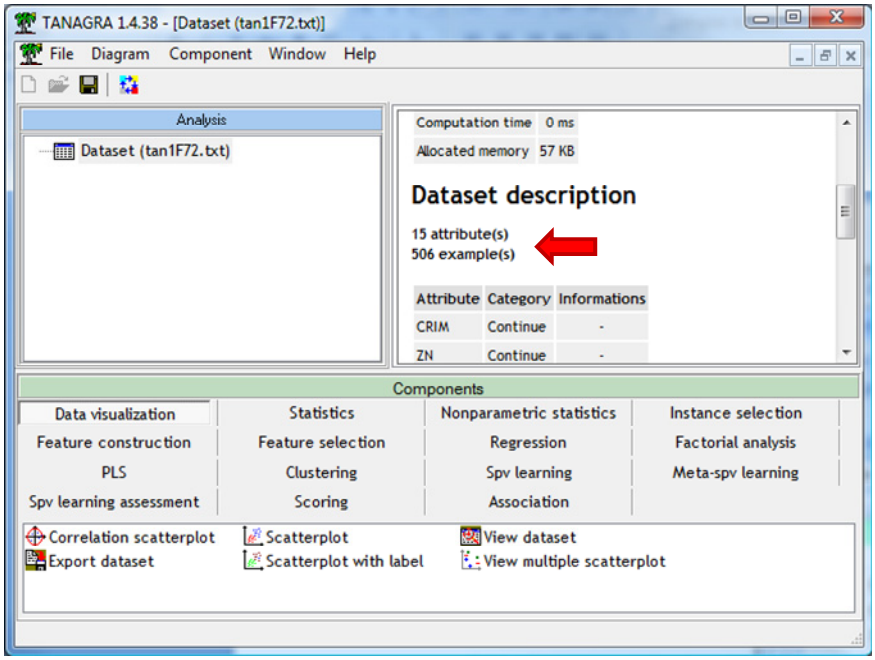
3 Déploiement de modèles en régression

3.1 Importation des données

Pour importer les données dans Tanagra, le plus simple est de les charger dans le tableur Excel (Open Office fait très bien l'affaire aussi) et de les envoyer vers Tanagra via la macro complémentaire « Tanagra.xla » (*voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html pour l'intégration de l'add-in dans Excel 2003 et antérieures ; http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour Excel 2007 et 2010*).

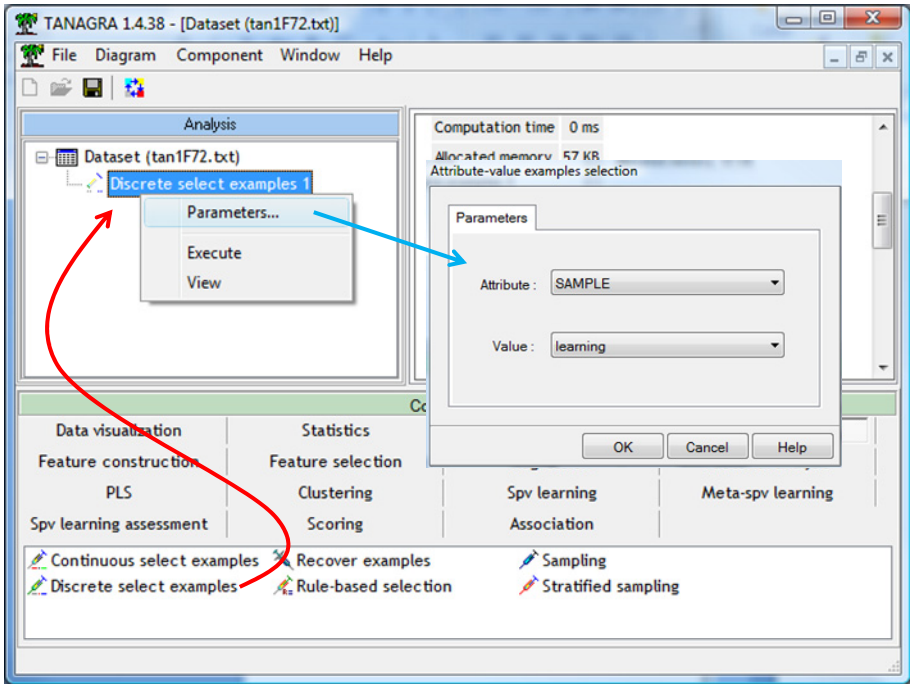


La totalité des observations, soit 506 lignes, ont été importées correctement.

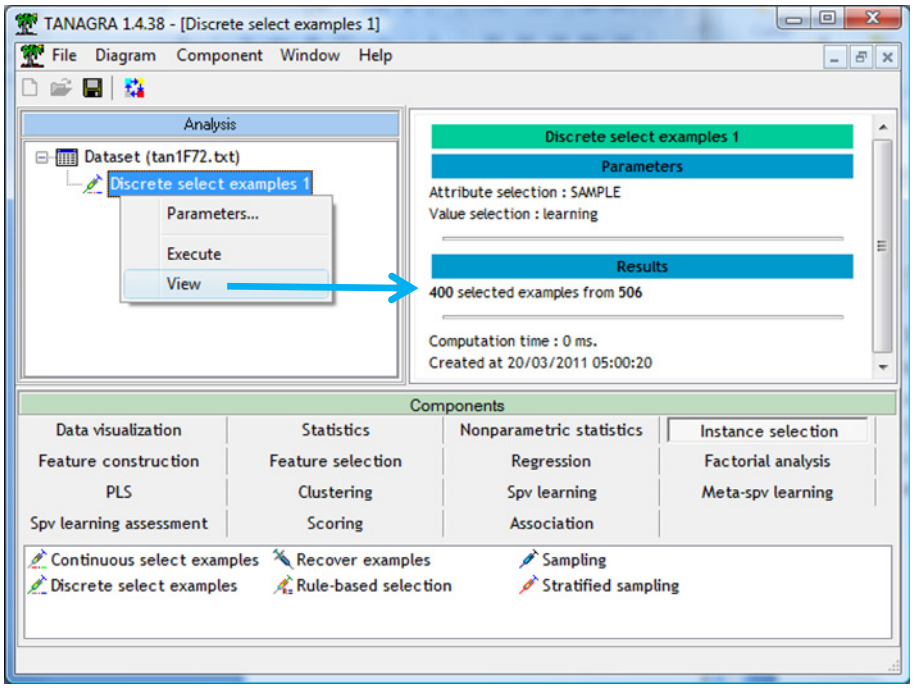


3.2 Subdivision « données d'apprentissage » - « données à étiqueter »

Nous insérons le composant DISCRETE SELECT EXAMPLES (onglet INSTANCE SELECTION) dans le diagramme. Il permet de scinder les observations en se basant sur un des champs de la base. Nous le paramétrons : l'attribut de sélection est SAMPLE, les observations actives correspondent à la modalité LEARNING.



Nous validons : 400 observations sont maintenant disponibles pour les calculs.

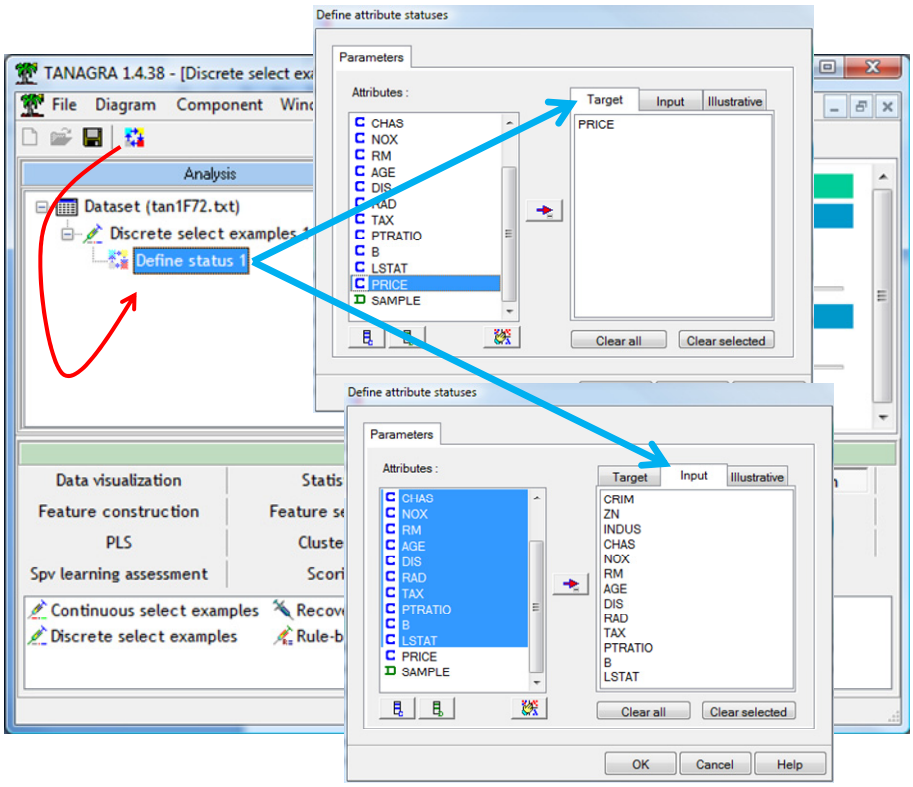


Les lignes restantes ne sont pas utilisées lors de l'apprentissage. **En revanche, et c'est là le secret de la manipulation, lorsque le composant de modélisation réapplique le modèle sur les données, il le fait sur la totalité du fichier, y compris donc sur les observations non sélectionnées. Nous tirons profit de cette propriété pour obtenir les prédictions sur les individus non étiquetés**.

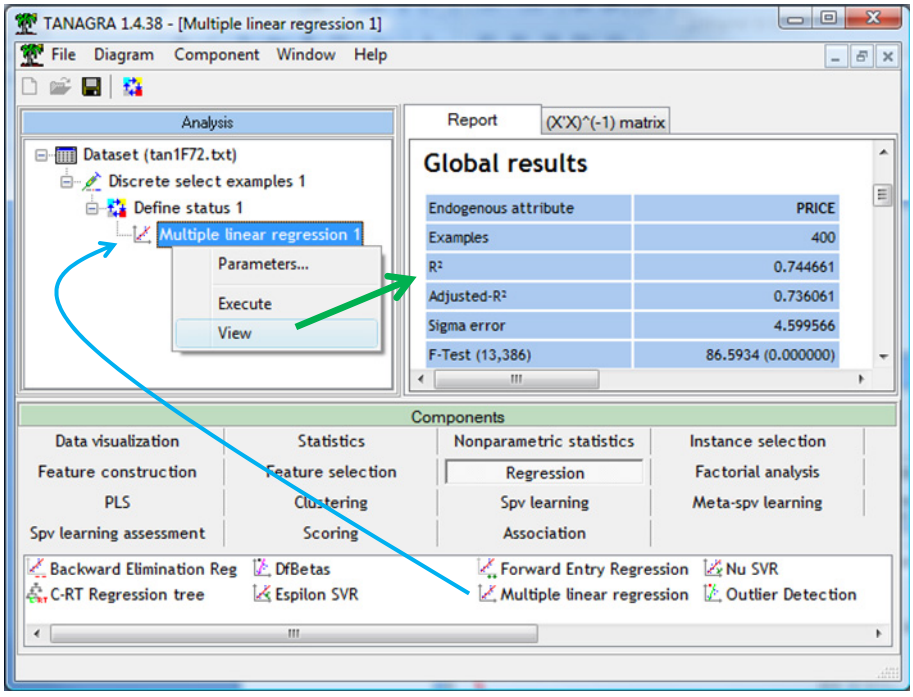
3.3 Construction des modèles

3.3.1 Régression linéaire multiple

Nous utilisons le composant DEFINE STATUS pour définir le rôle des variables : PRICE est la variable à prédire (TARGET), les autres correspondent aux variables explicatives (INPUT). La colonne SAMPLE ne sert plus à ce stade.

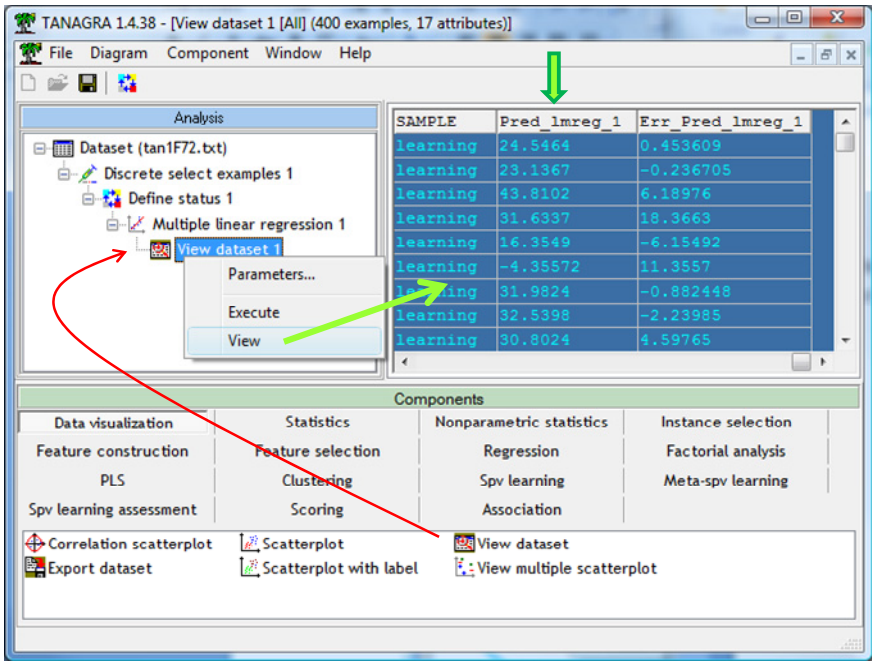


Nous plaçons ensuite le composant MULTIPLE LINEAR REGRESSION (onglet REGRESSION). Nous actionnons le menu VIEW pour obtenir les résultats.



Le coefficient de détermination $R^2 = 0.744$. Certaines variables ne semblent pas significatives. Nous ne nous attarderons pas trop sur les résultats dans ce didacticiel, notre objectif étant d'étudier le déploiement de modèles sur de nouveaux individus.

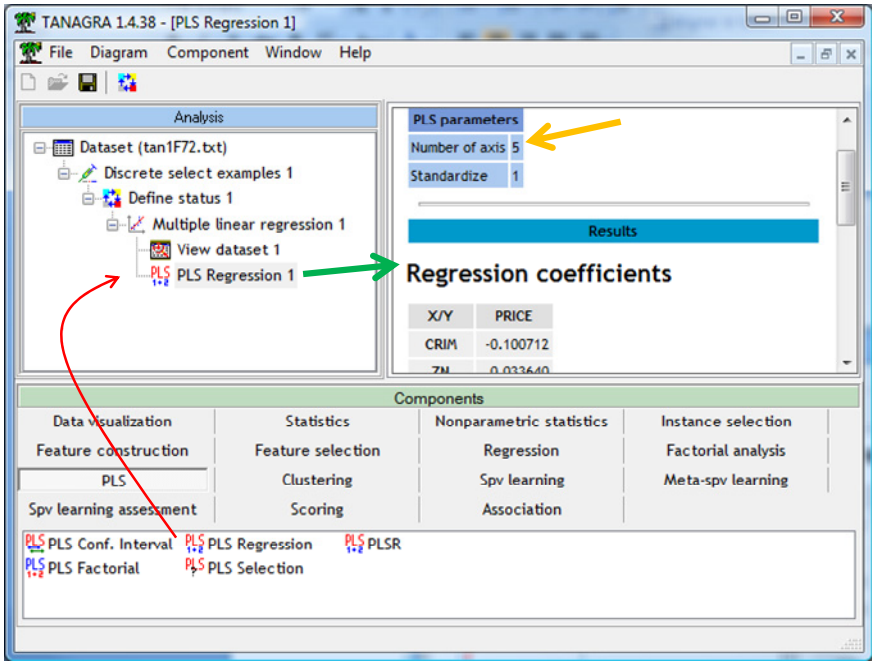
Nous pouvons visualiser les données à l'aide de VIEW DATASET (onglet DATA VISUALIZATION).



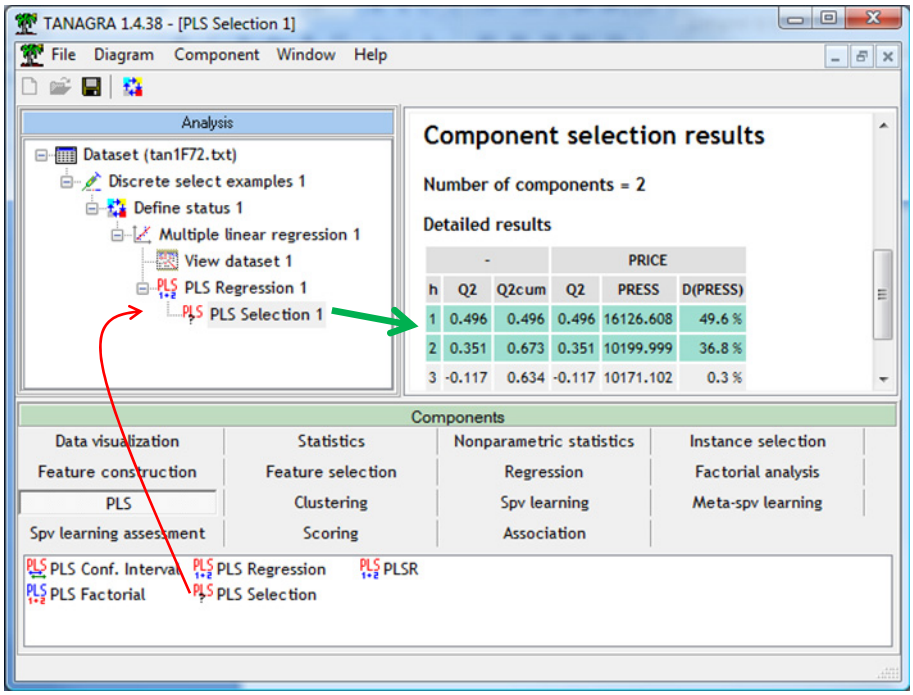
Nous remarquons deux nouvelles colonnes : **PRED_LMREG_1** correspond à la prédiction du modèle, ***calculée sur toutes les observations***, y compris celles qui ne sont pas sélectionnées ; ERR_PRED_LMREG_1 est le résidu, la différence entre les valeurs observées et prédites.

3.3.2 Régression PLS

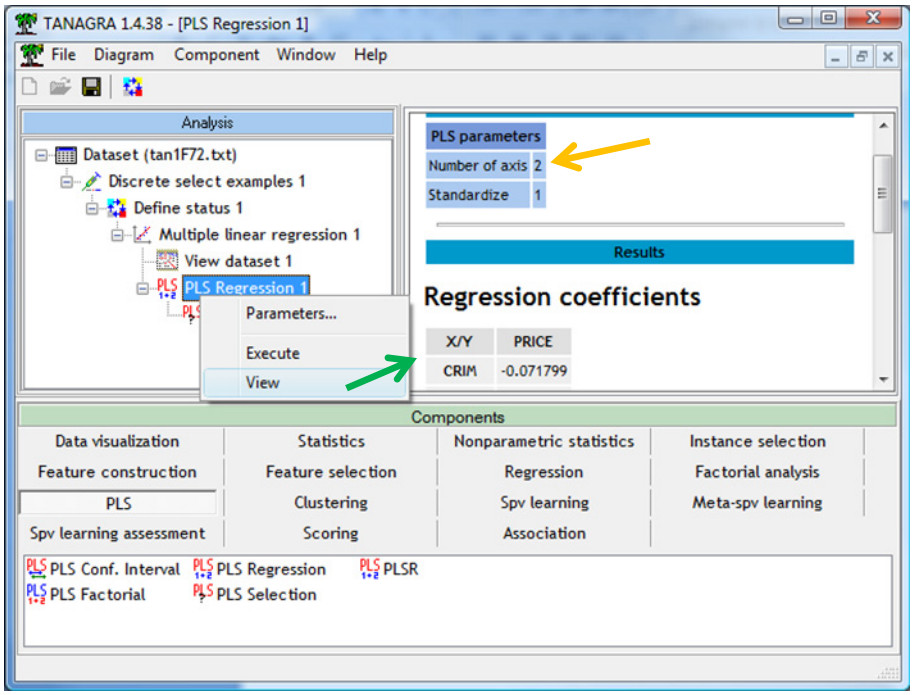
Nous insérons le composant PLS REGRESSION (onglet PLS) à la suite de la régression linéaire. Nous le lançons directement sans nous préoccuper du paramétrage (menu VIEW).



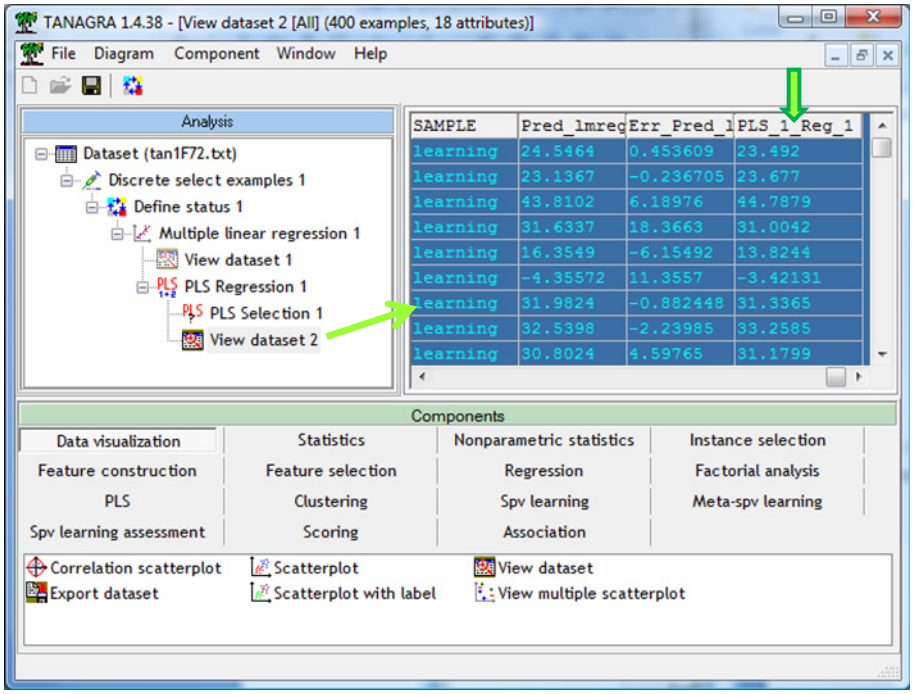
Par défaut, 5 axes factoriels sont produits. Ce choix n'est certainement pas « optimal » (si tant est qu'il y en ait une valeur optimale d'ailleurs). Nous pouvons demander à Tanagra à sélectionner le « bon » nombre d'axes par validation croisée avec le composant PLS SELECTION.



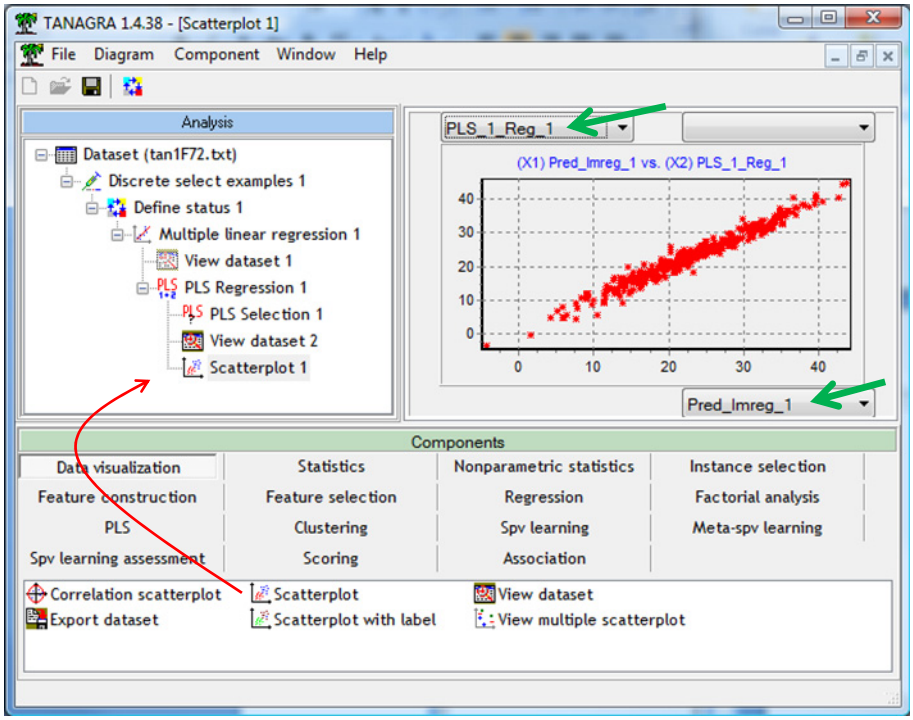
2 axes semblent être un bon compromis pour la prédiction. En re-cliquant sur le menu VIEW du composant PLS REGRESSION, nous obtenons les coefficients du modèle définitif.



Ici également, en visualisant l'ensemble des données, nous notons qu'une nouvelle colonne (Prédiction – **PLS_1_REG_1**) a été rajoutée.



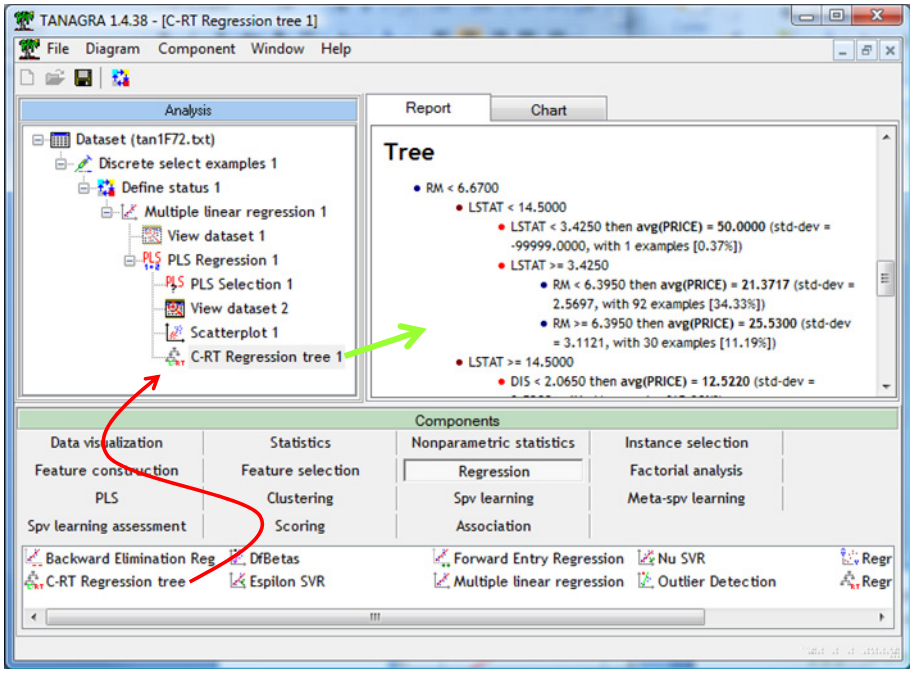
Nous pouvons confronter les prédictions des deux modèles (Régression Linéaire et Régression PLS) sur les données d'apprentissage en introduisant le composant SCATTERPLOT (onglet DATA VISUALIZATION).



Les prédictions sont très cohérentes ! C'est plutôt rassurant d'ailleurs. Les deux modèles étant linéaires, de fortes disparités auraient été inquiétantes.

3.3.3 Arbre de régression

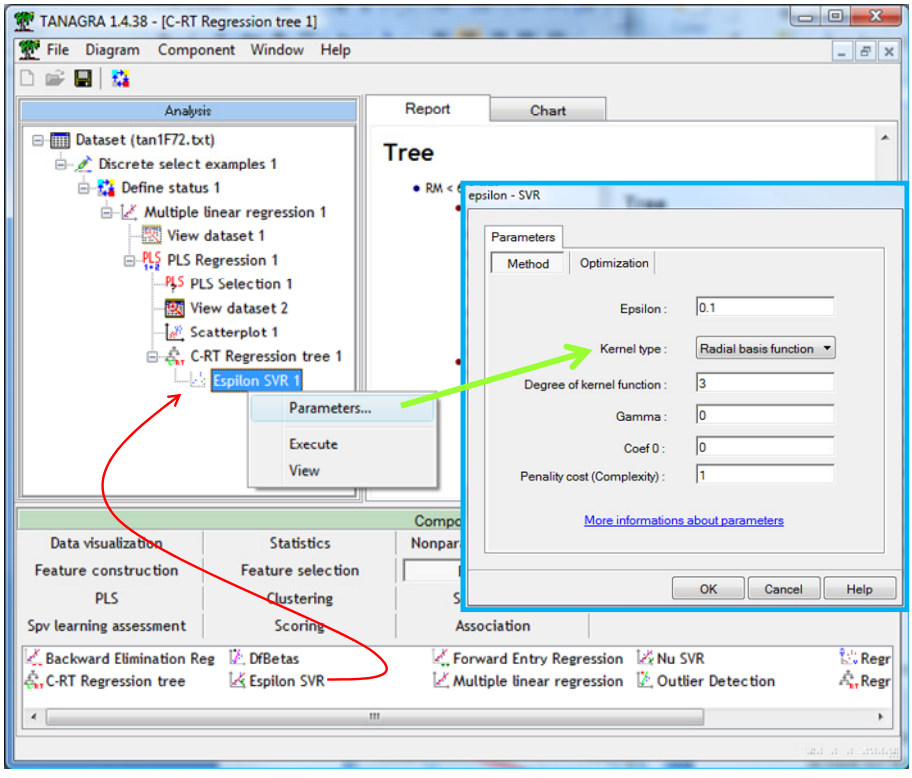
Nous introduisons maintenant un arbre de régression (méthode CART) dans le diagramme : nous insérons le composant C-RT Regression Tree (onglet Regression).



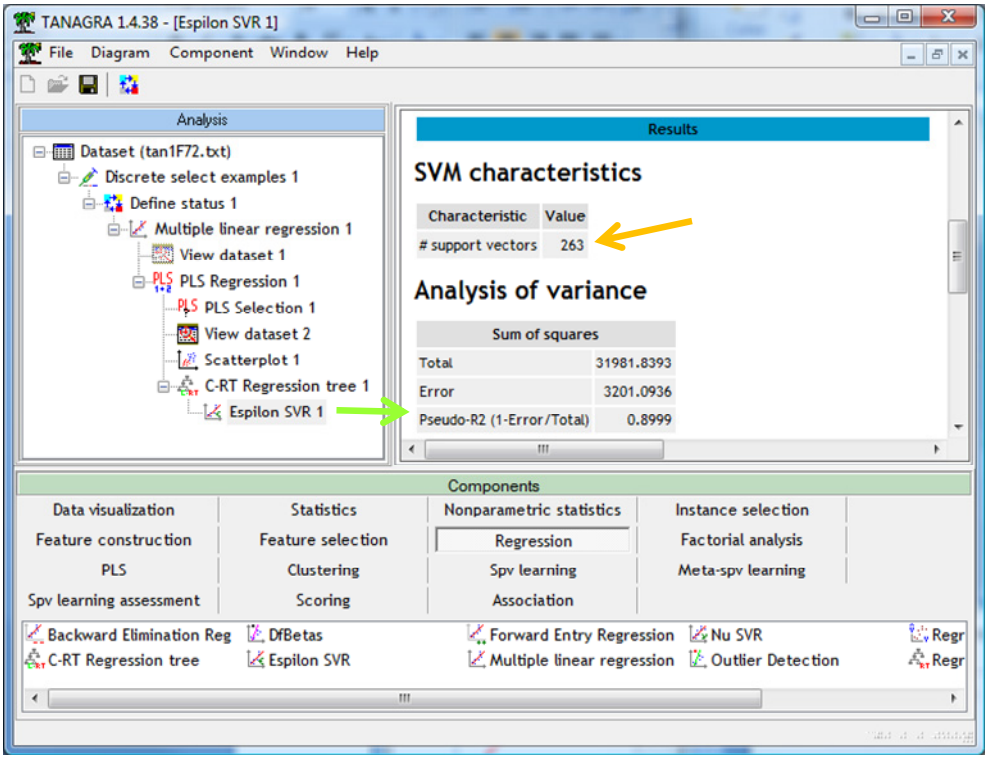
L'arbre obtenu comporte 12 feuilles. Comme pour les modèles linéaires, le déploiement dans les systèmes d'informations est aisé parce que le modèle est très facile à déchiffrer.

3.3.4 Support vector regression

Prochaine technique de régression : un support vector regression avec le composant EPSILON SVR (onglet REGRESSION). Nous avons présenté cet outil par ailleurs (http://tutoriels-data-mining.blogspot.com/2009/04/support-vector-regression.html). Nous le paramétrons ici de manière à utiliser un noyau RBF (Radial Basis Function).

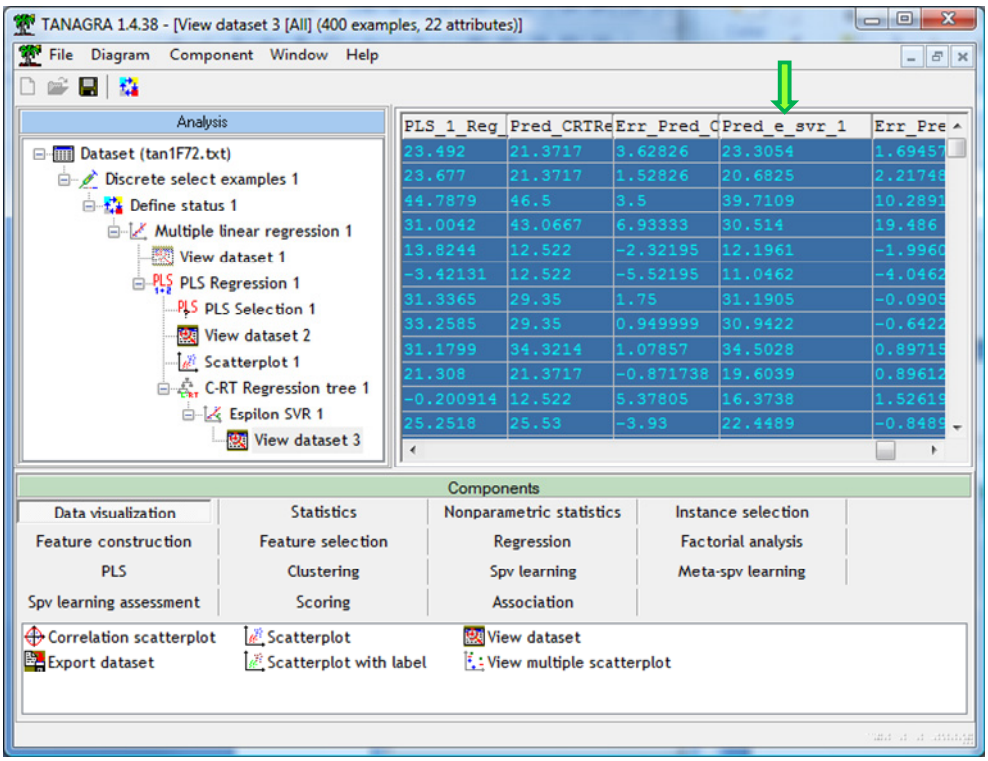


Nous actionnons le menu VIEW pour obtenir les résultats.



Ici commencent réellement les problèmes. En effet, nous ne disposons pas de modèle explicite. On se base sur les points supports (« Support Vectors » , ***263*** en ce qui nous concerne) proposés par SVR pour étiqueter les nouveaux individus. Réaliser le calcul manuellement est très compliqué. Il serait heureux que le logiciel ayant servi à la construction du modèle puisse s'en charger directement. C'est ce que fait Tanagra, et la grande majorité des logiciels de Data Mining d'ailleurs.

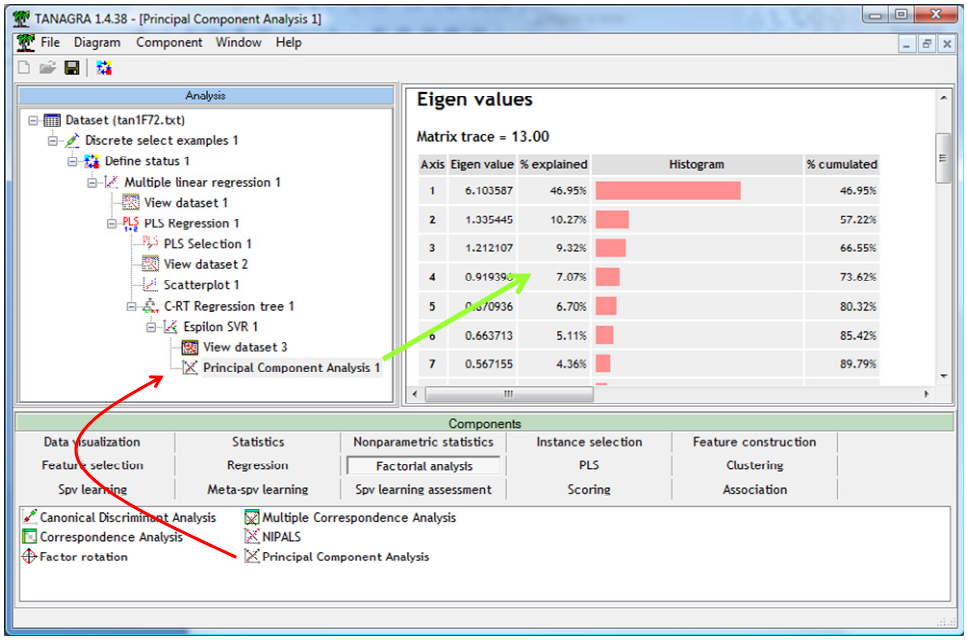
Avec le composant VIEW DATASET, nous constatons que EPSILON SVR a ajouté une colonne prédiction dans l'ensemble de données.



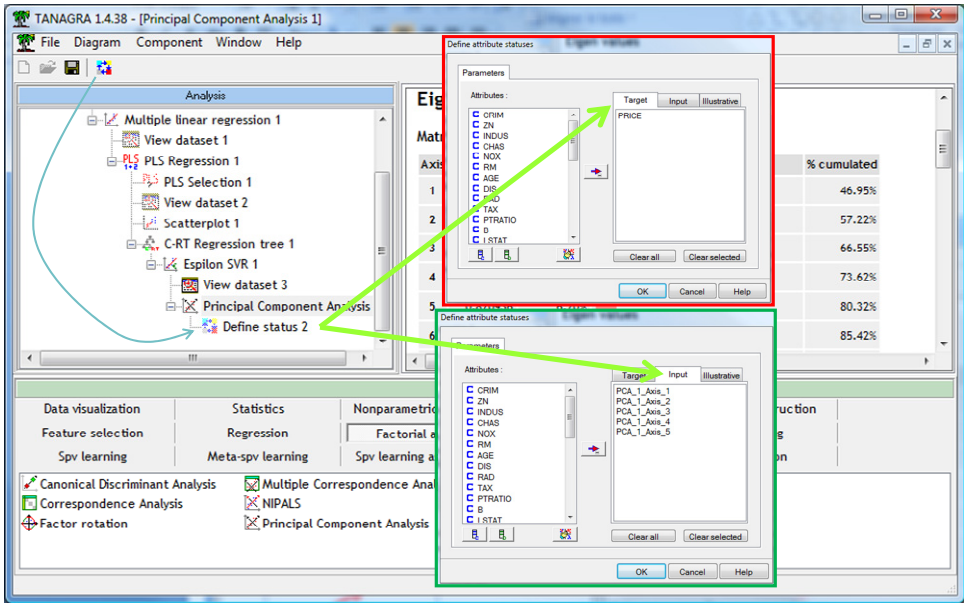
3.3.5 Régression sur axes

Jusqu'à maintenant, nous utilisions les techniques individuellement, le déploiement ne fait intervenir qu'un seul modèle. Dans cette section, nous souhaitons faire coopérer plusieurs techniques. Dans une première phase, nous réalisons une analyse factorielle des correspondances sur les variables explicatives ; puis, dans une seconde phase, nous utilisons les 5 premiers axes factoriels dans une régression linéaire multiple. Lors de la prédiction sur de nouvelles observations, il faudrait déployer les deux « modèles », la manipulation est mal aisée (***Remarque*** : *certes, on peut retrouver une équation unique puisque nous avons une combinaison linéaire de combinaisons linéaires, mais ce type d'opération n'est pas facile à appréhender*).

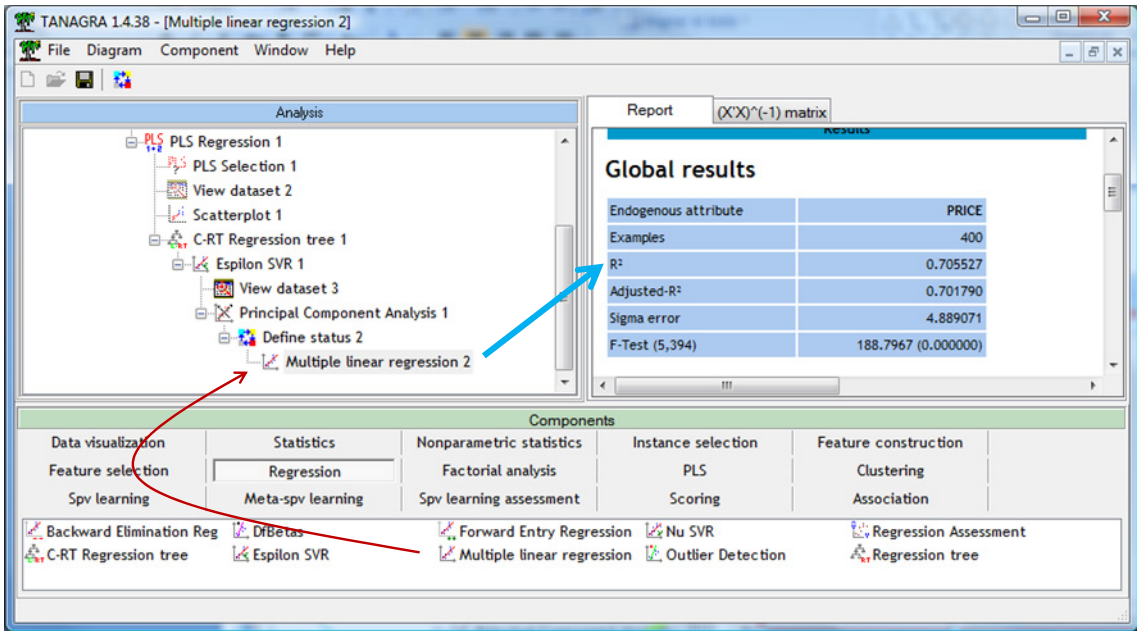
Nous insérons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme. Nous obtenons les résultats suivants.



Avec le composant DEFINE STATUS, nous plaçons les 5 axes factoriels comme variables explicatives (INPUT), PRICE comme variable cible (TARGET).



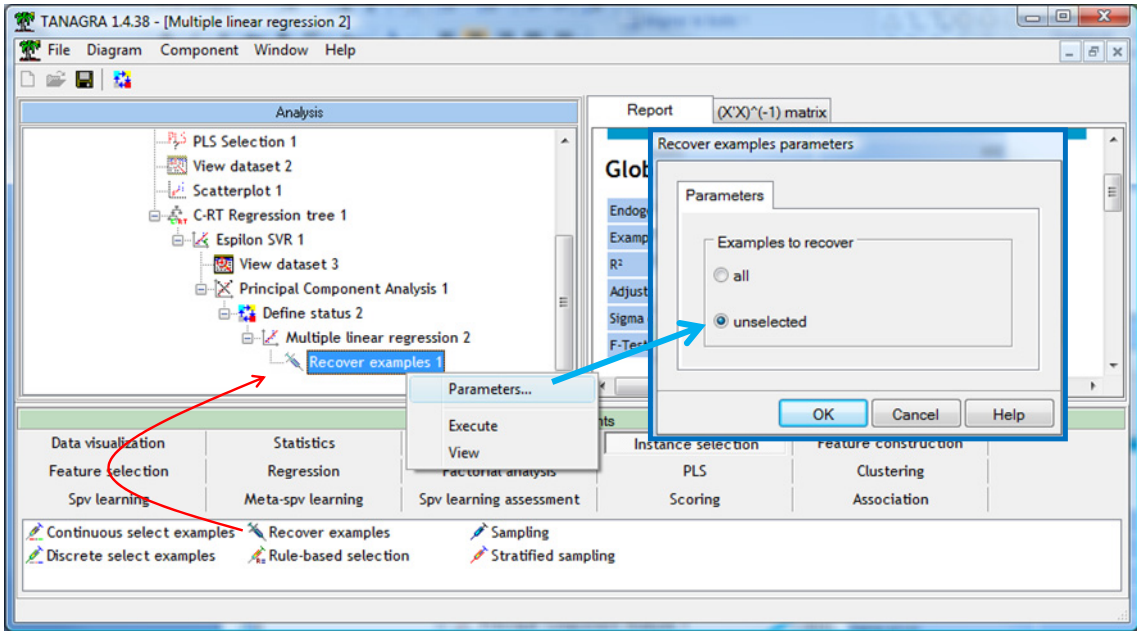
Nous plaçons la régression linéaire multiple, puis nous cliquons sur VIEW.



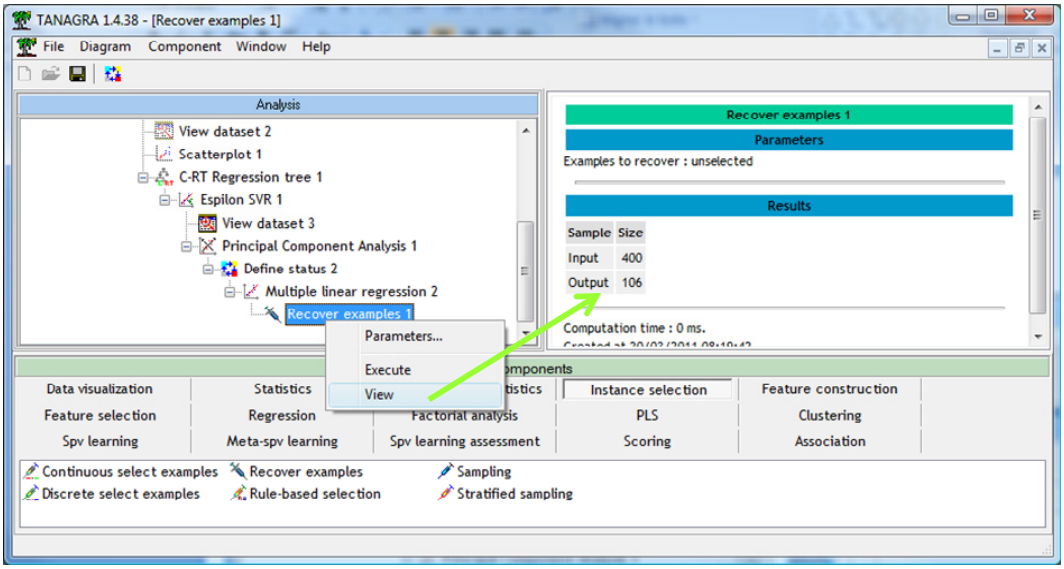
3.4 Récupération des valeurs prédites

En sous main, Tanagra réalise les prédictions sur les individus non sélectionnés pour chaque modèle élaboré. Il est temps maintenant de les visualiser. Pour ce faire, nous allons inverser la sélection des observations c.-à-d. rendre actives les individus auparavant masquées, et inversement.

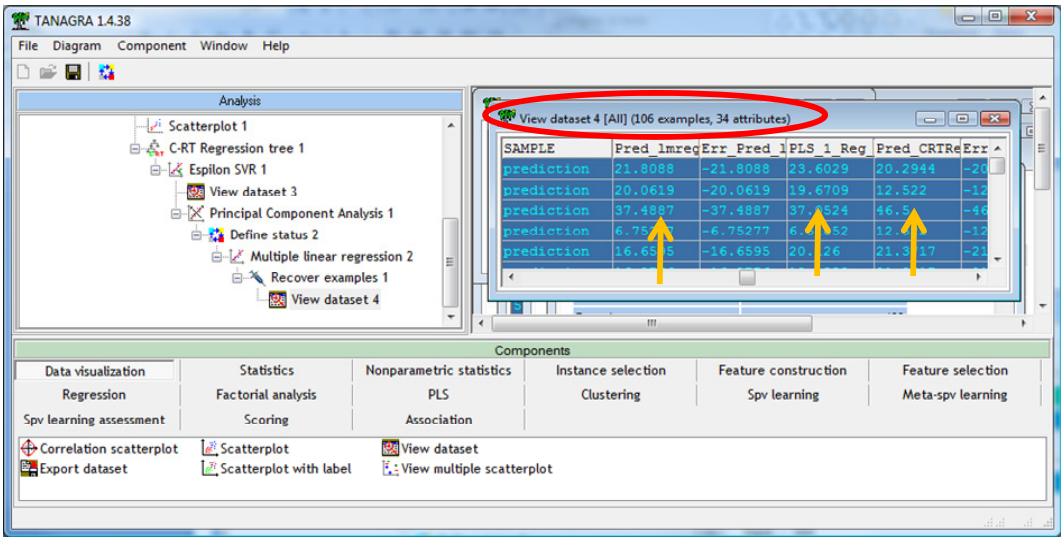
Nous insérons le composant RECOVER EXAMPLES dans le diagramme, nous le paramétrons de manière à recouvrir les UNSELECTED.



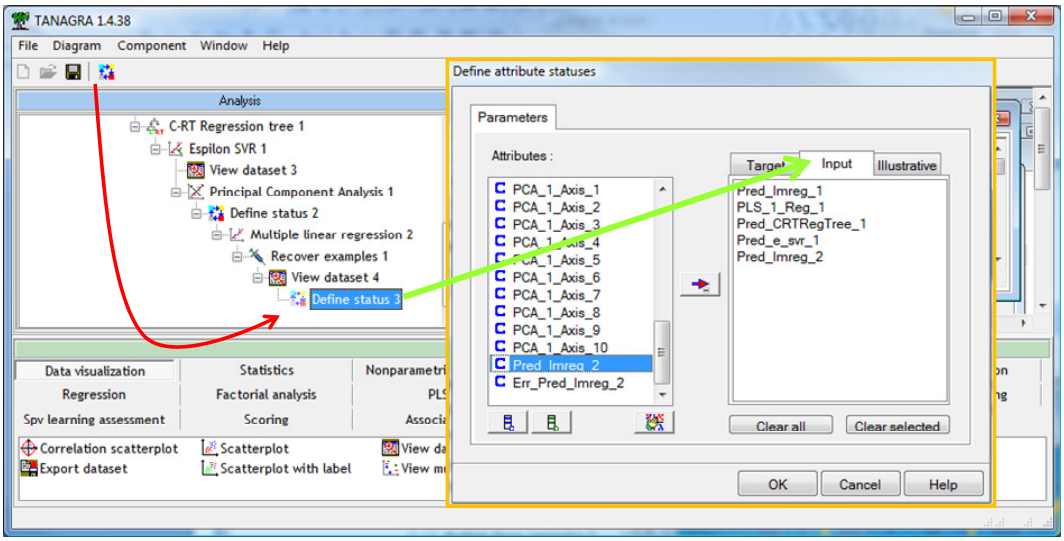
Nous actionnons le menu VIEW : **106** observations sont maintenant activées.



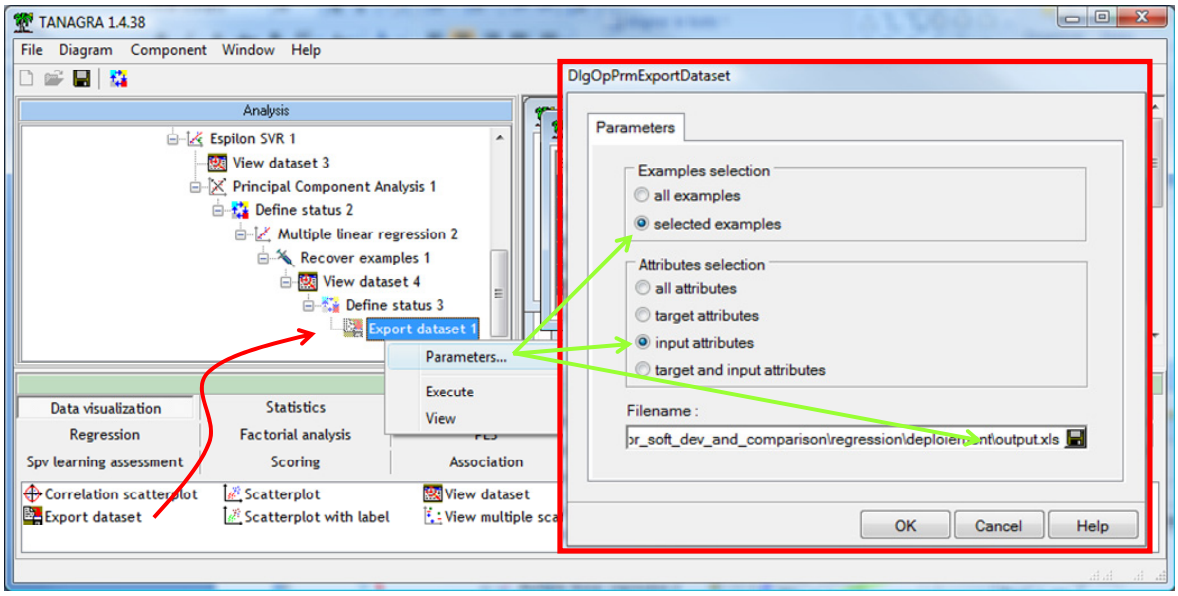
Nous pouvons les visualiser avec VIEW DATASET. Dans les dernières colonnes, nous distinguons les prédictions de la régression linéaire multiple, de la régression PLS, de l'arbre de décision, etc.



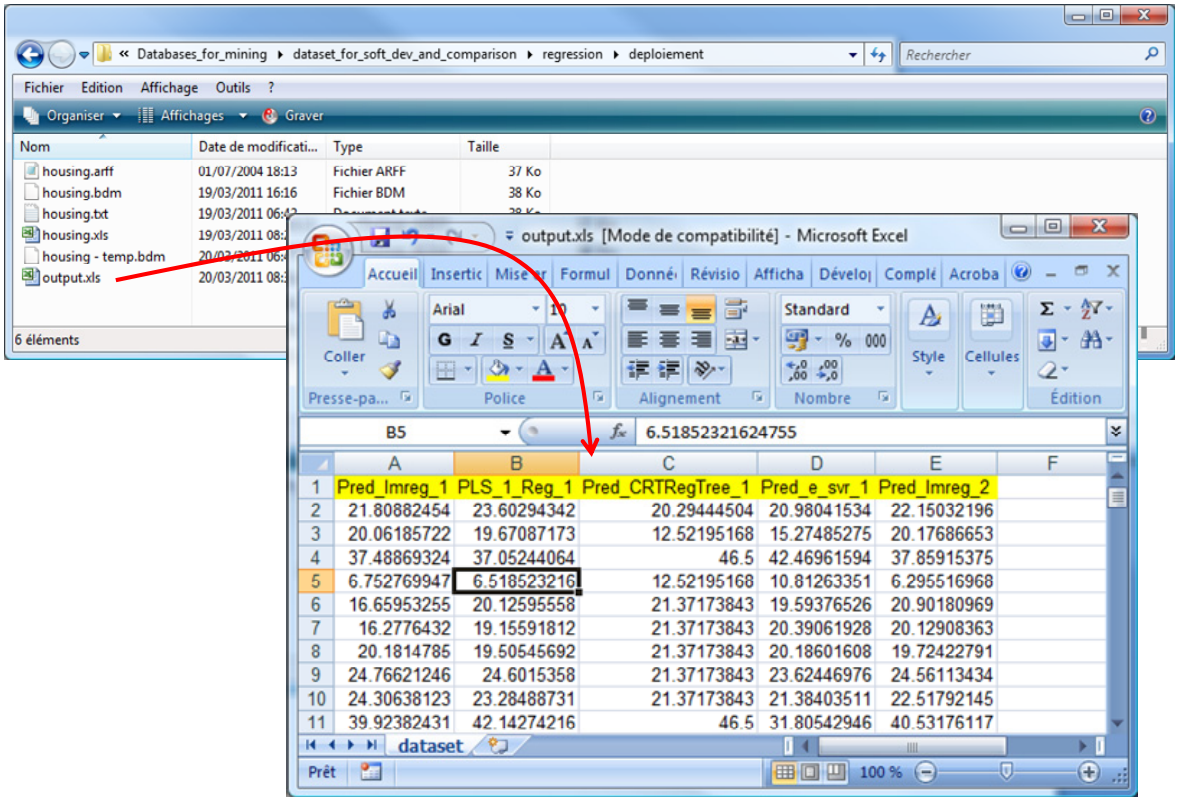
Enfin, nous pouvons récupérer ces valeurs en les exportant dans un fichier Excel. Nous plaçons un DEFINE STATUS pour ne sélectionner que les prédictions des modèles en INPUT.



Nous plaçons le composant EXPORT DATASET. Nous choisissons de n'exporter que les observations sélectionnées (Examples selection = selected examples) et les variables auparavant placées en INPUT (Attributes selection = input attributes). Le nom du fichier de sortie est « output.xls ».

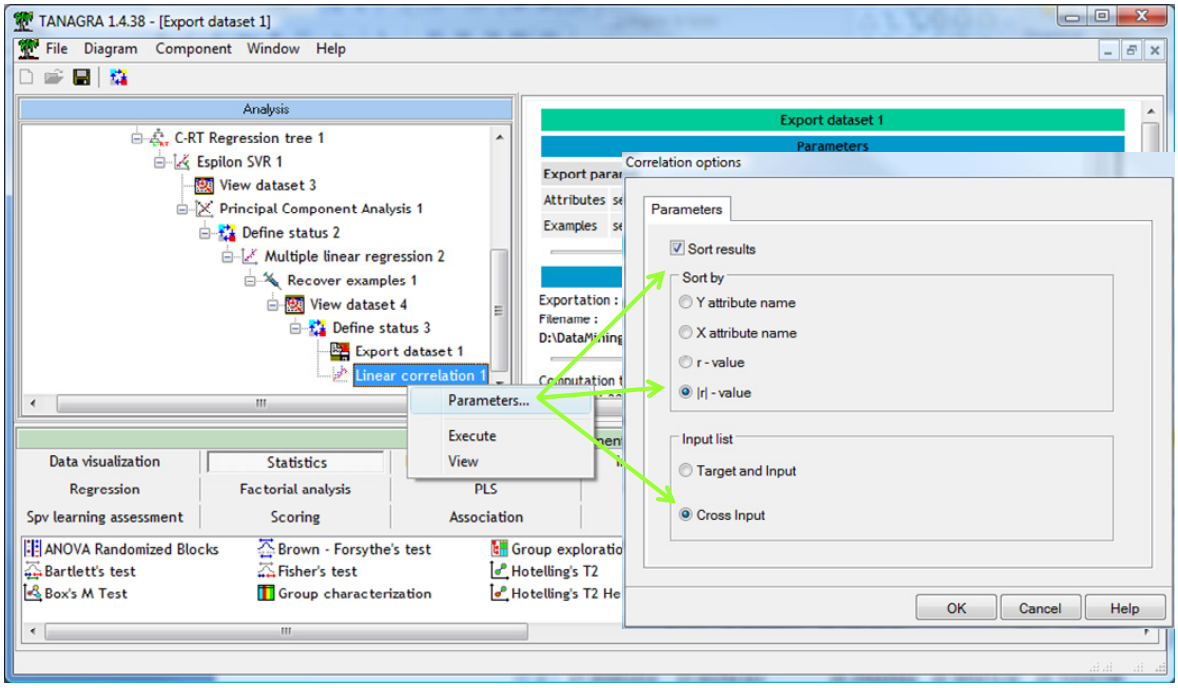


Nous validons puis nous cliquons sur VIEW. Le fichier est bien généré sur le disque, nous pouvons le visualiser via le tableur Excel.

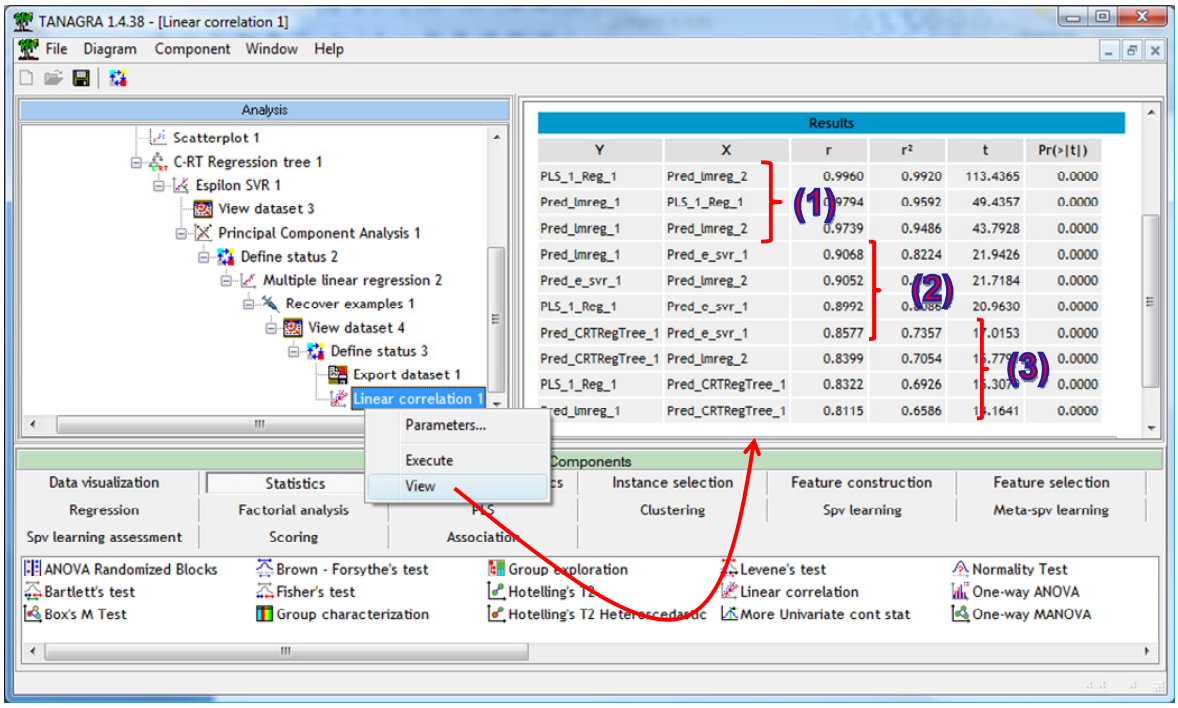


3.5 Comparaison des prédictions

Dernière étape enfin, nous souhaitons analyser la cohérence des prédictions. Le plus simple est calculer les corrélations croisées. Nous ajoutons le composant LINEAR CORRELATION (onglet STATISTICS) dans le diagramme. Nous le paramétrons de la manière suivante.



Nous constatons que les prédictions issues des modèles linéaires sont très proches (régression linéaire multiple, régression PLS, régression linéaire sur axes factoriels) (1). Le SVR avec un noyau RBF, non linéaire donc, se démarque un peu (2). L'arbre de décision, un modèle non linéaire aussi mais avec des caractéristiques très différentes (il morcelle l'espace de représentation pour réaliser une prédiction locale avec une moyenne conditionnelle dans chaque zone), est vraiment à part (3).



4 Conclusion

Bien évidemment, d'autres pistes existent pour le déploiement de modèles, surtout dans un cadre professionnel où l'on souhaite les industrialiser dans les systèmes d'information. Une solution simple consiste à exporter les modèles dans un format standard, le format PMML par exemple

(http://www.dmg.org/v4-0-1/Regression.html pour la régression ; http://www.dmg.org/v4-0-1/TreeModel.html, pour les arbres ; la chose n'est pas facile quand il s'agit de SVR, http://www.dmg.org/v4-0-1/SupportVectorMachine.html ; etc.), puis à utiliser des outils capables de les appliquer sur les observations à étiqueter (http://www.pentaho.com/products/data_integration/ -- Pentaho PDI par exemple). Nous avions décrit cette solution dans le cadre du déploiement des classifieurs en apprentissage supervisé (http://tutoriels-data-mining.blogspot.com/2010/09/le-format-pmml-pour-le-deploiement-de.html).