

Objectif

Comparer TANAGRA, ORANGE et WEKA lors de la construction d'une courbe ROC à partir de la régression logistique.

TANAGRA, ORANGE et WEKA sont trois logiciels de data mining gratuits. S'ils poursuivent le même objectif, permettre aux utilisateurs de définir une succession de traitements sur les données, ils présentent néanmoins des différences. C'est tout à fait normal. Leurs auteurs n'ont pas la même culture informatique, cela se traduit par des choix technologiques différents ; ils n'ont pas la même culture de la fouille de données, ce qui se traduit par un vocabulaire et par un mode de présentation des résultats parfois différents.

Au-delà de leurs propres spécificités, ces outils permettent de définir les mêmes analyses et par conséquent produisent les mêmes résultats. La comparaison sera d'autant plus aisée qu'ils adoptent le même mode de représentation graphique des séquences d'opérations, à l'aide d'un graphe pour ORANGE et WEKA, à l'aide d'un arbre pour TANAGRA : chaque sommet représente un traitement, le lien entre chaque sommet représente le flux de données.

Dans ce didacticiel, nous montrons la construction de la courbe ROC à partir d'une régression logistique. Quel que soit le logiciel utilisé, nous devons impérativement passer par les étapes suivantes :

- Importer les données dans le logiciel ;
- Calculer les statistiques descriptives pour se donner une idée de la nature des données ;
- Définir le problème à résoudre, c.-à-d. choisir la variable à prédire (l'attribut « classe ») et les descripteurs ;
- Définir la modalité « positive » de la variable à prédire ;
- Subdiviser le fichier en données d'apprentissage (70% par exemple), servant à construire le modèle de prédiction, et en données test (30 %), pour construire la courbe ROC ;
- Choisir l'algorithme d'apprentissage, nous voulons mettre en œuvre la régression logistique, selon la méthode réellement implémentée dans le logiciel, c'est le principal point de différenciation, nous pouvons obtenir des résultats légèrement différents ;
- Lancer l'apprentissage et visualiser les résultats ;
- Elaborer enfin la courbe ROC sur les données en test afin d'évaluer les performances en classement.

La progression peut ne pas être la même pour chaque logiciel. Néanmoins, à un moment ou un autre, **il faudra passer par chacune des étapes ci-dessus, de manière explicite ou non, pour arriver à nos fins.**

Fichier

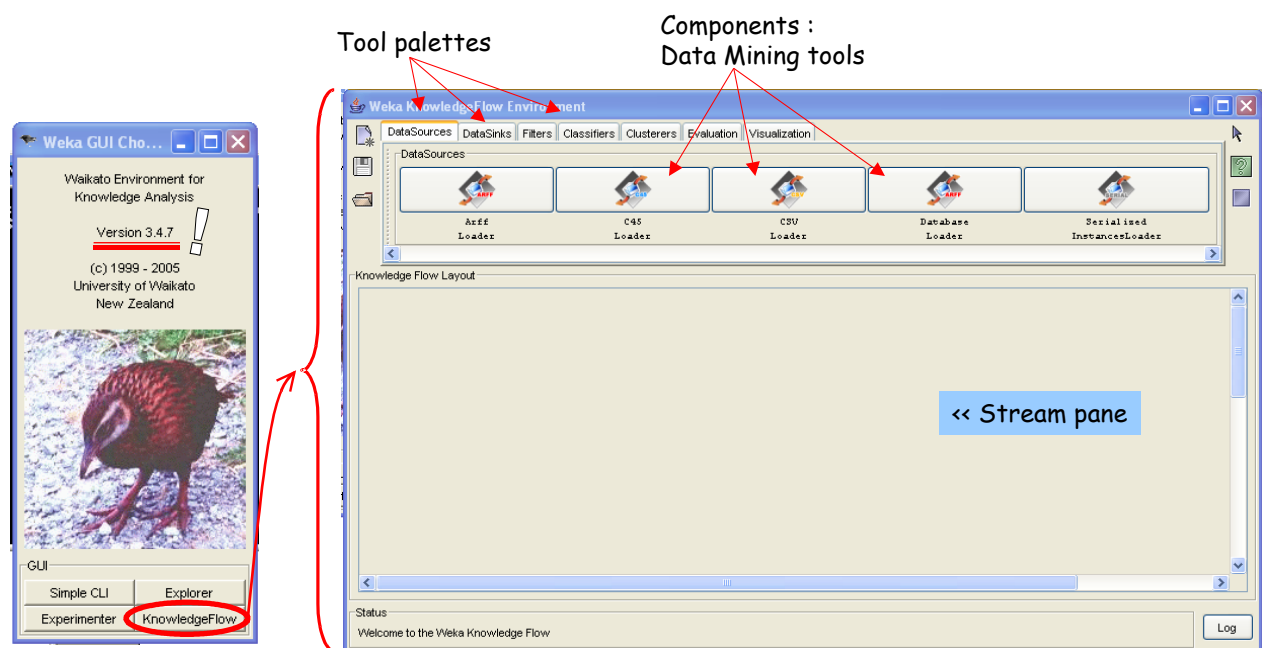
Les données utilisées proviennent du site de la bibliothèque LR-TRIRLS que nous avons reprise dans TANAGRA (Régression Logistique -- <http://komarix.org/ac/lr>). Le fichier DS1-10 comporte 26733 observations, 10 descripteurs continus ; la variable à prédire est binaire, la modalité positive représentant 5% des observations. Les données sont disponibles en deux formats : ARFF pour WEKA et TANAGRA, TXT pour ORANGE (TANAGRA peut traiter également les fichiers TXT).

Construire la courbe ROC avec WEKA

WEKA impressionne par la richesse de sa bibliothèque de méthodes, mais paradoxalement c'est également son principal défaut : les possibilités sont immenses mais la présentation est très touffue, il est difficile de repérer les bonnes fonctionnalités et les composants à mettre en œuvre. L'intérêt de ce didacticiel est justement de discerner les composants adéquats dans le traitement que nous voulons mettre en œuvre. Nous verrons que la succession des opérations est loin d'être évidente pour le néophyte.

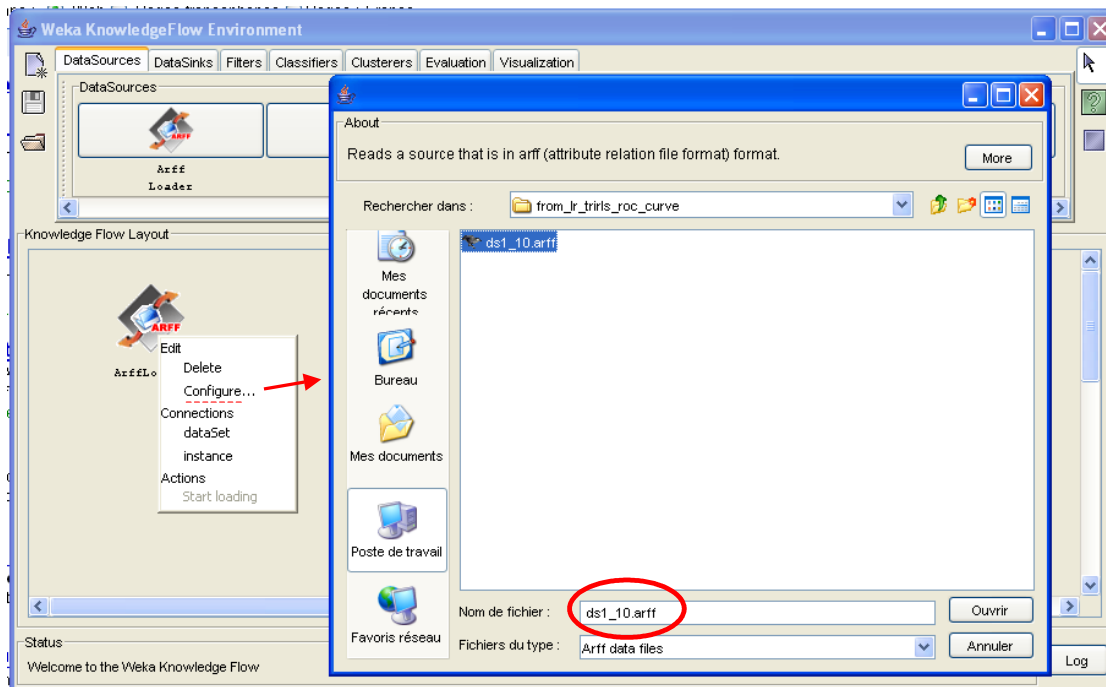
Lancement de WEKA

Au lancement de WEKA, un panneau permet de choisir le mode d'exécution du logiciel. Nous choisissons le mode KNOWLEDGE FLOW. Nous avons utilisé la version 3.4.7 dans ce didacticiel. Nous parvenons alors dans l'espace de travail dans lequel nous allons définir nos traitements. Dans la partie haute, nous trouvons les icônes organisées dans des palettes, ils représentent les opérateurs de traitement.



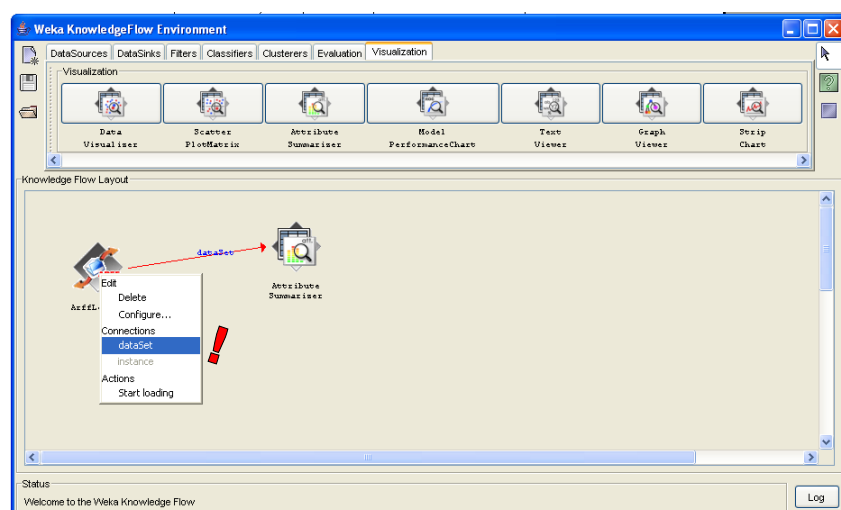
Charger les données

Le composant ARFF LOADER permet de charger les données. Nous ajoutons donc ce composant dans notre espace de travail, nous pouvons le paramétrer à l'aide de l'option CONFIGURE de son menu contextuel.

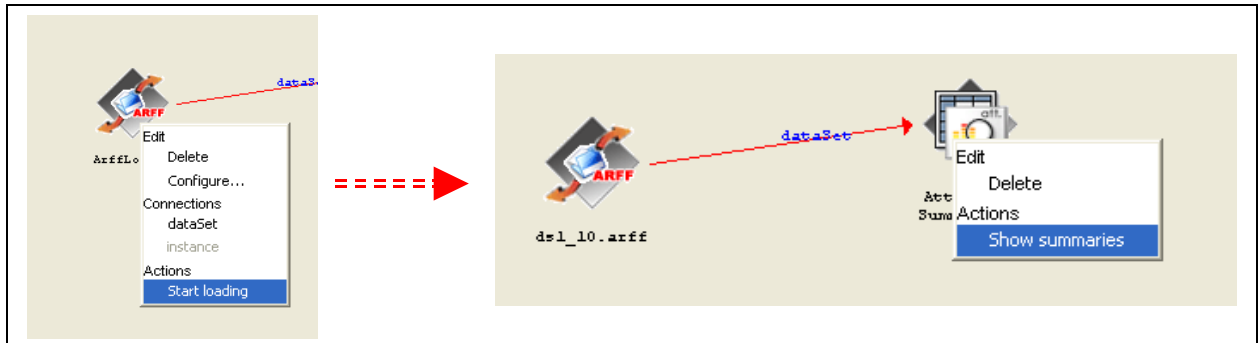


Statistiques descriptives

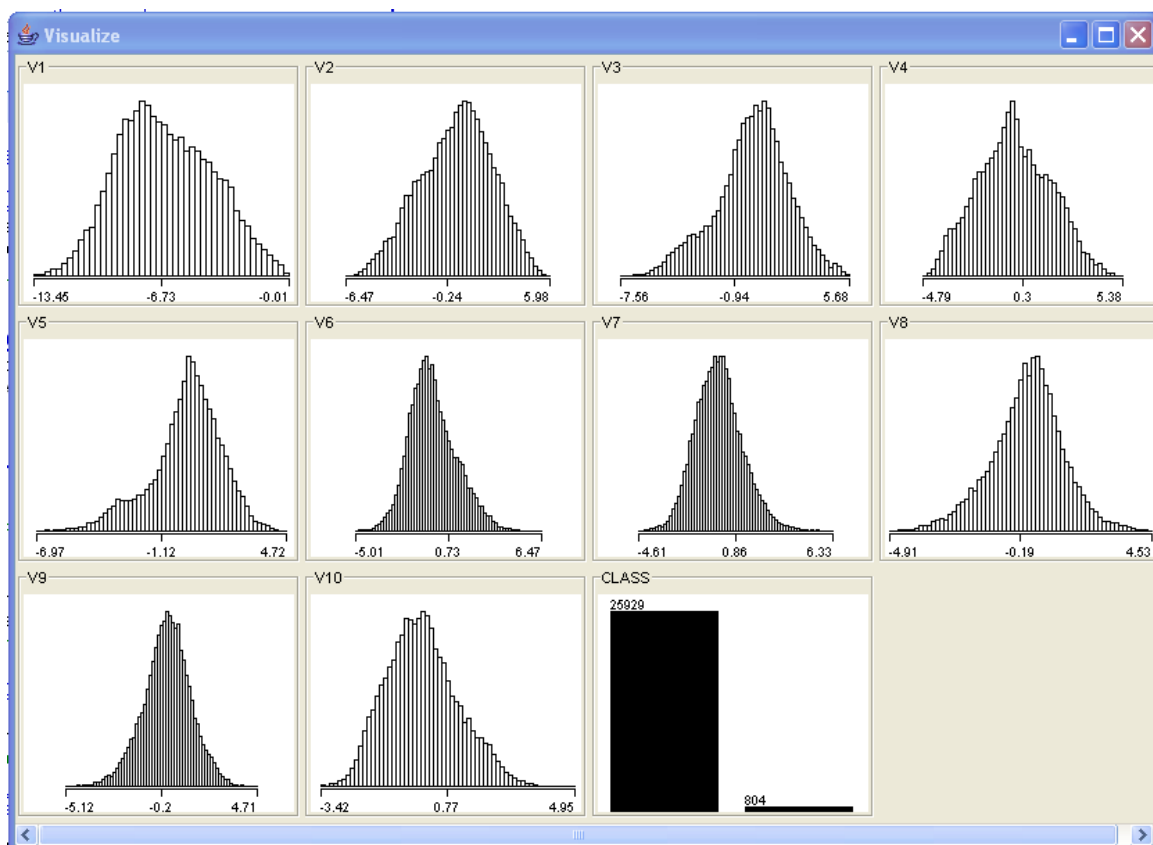
Dans un premier temps, il est conseillé de visualiser rapidement la nature des données. Le composant ATTRIBUT SUMMARIZER (palette VISUALIZATION) permet de voir en un coup d'œil la distribution des descripteurs. Nous l'ajoutons donc dans le diagramme, puis nous lui connectons le composant ARFF LOADER en sélectionnant l'option DATASET du menu contextuel.



Le traitement sera effectivement exécuté lorsque nous aurons activé le menu START LOADING du composant ARFF LOADER. Pour accéder aux résultats, il faut activer le menu SHOW SUMMARIES du second composant.



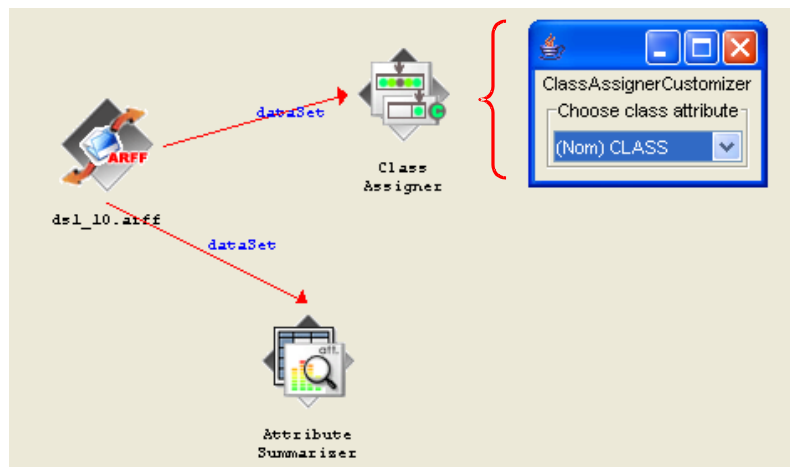
Nous obtenons l'histogramme de distribution de toutes les variables du fichier. Un rapide coup d'œil permet de voir qu'il n'y a pas de problème d'asymétrie ni de multi-modalité chez les descripteurs ; la variable à prédire CLASS est en revanche très déséquilibrée (804 « positifs » contre 25929 « négatifs »), ce qui laisse prévoir quelques difficultés lors de l'apprentissage.



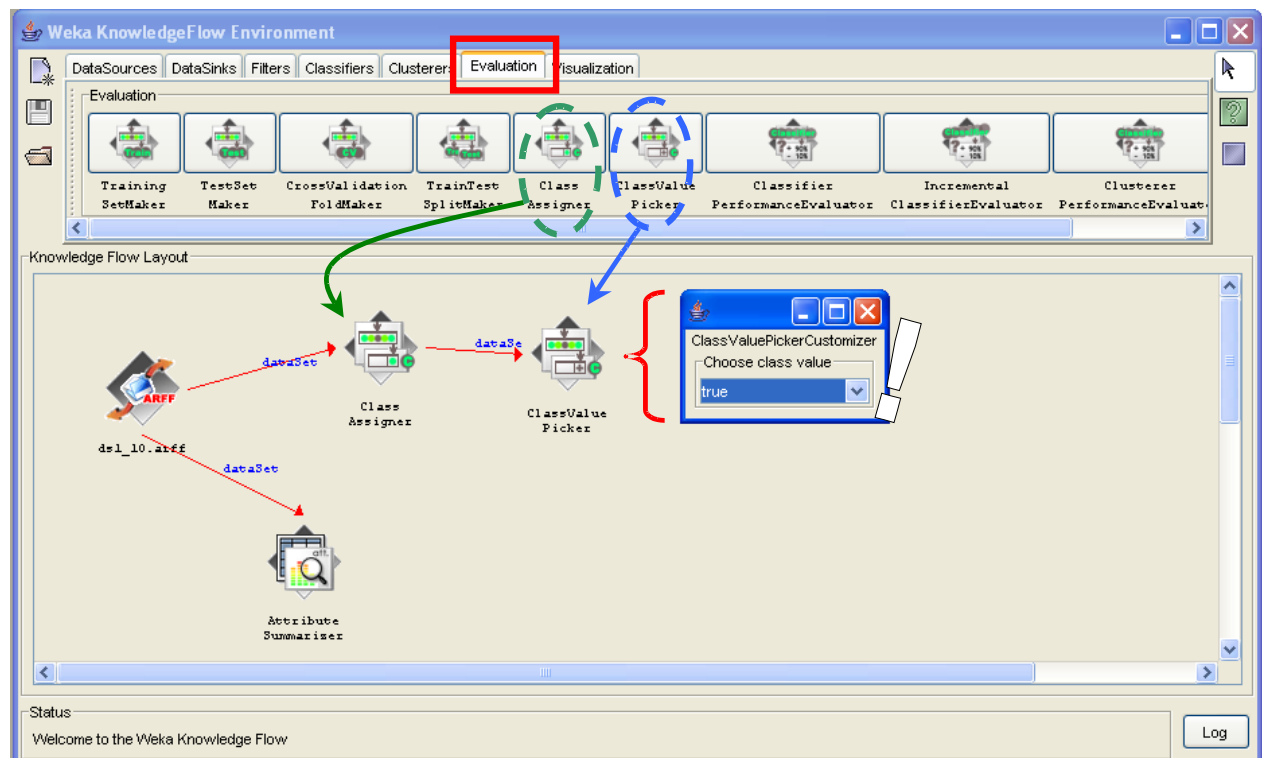
Choisir la variable à prédire et la modalité positive

L'étape suivante consiste à définir la variable à prédire. Par défaut, c'est la dernière variable du fichier, ce qui est notre cas ici. Néanmoins nous allons quand même le faire explicitement

afin que nous sachions comment procéder lorsqu'elle est située à un autre endroit dans le fichier. Nous sélectionnons le composant CLASS ASSIGNER (palette EVALUATION), auquel nous connectons le composant ARFF LOADER (connexion DATASET), nous le paramétrons à l'aide du menu CONFIGURE.



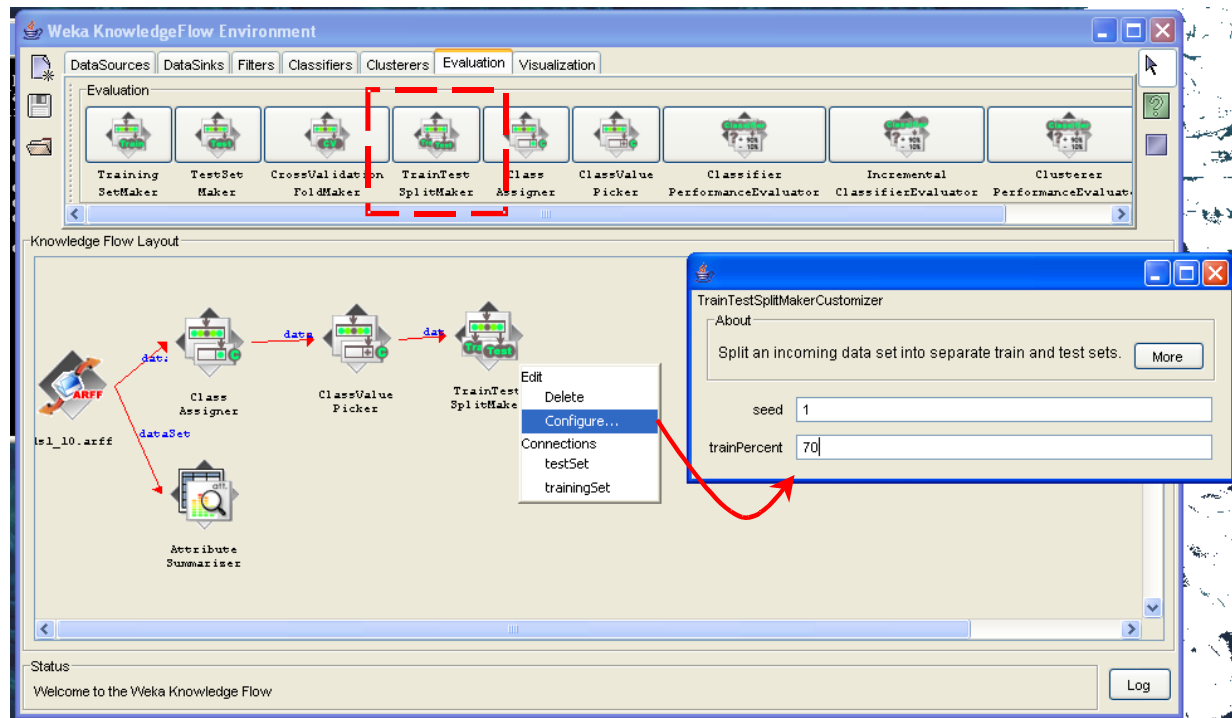
Etape suivante, il nous faut préciser quelle est la modalité « positive » de la variable à prédire. Toujours dans la palette EVALUATION, nous ajoutons le composant CLASSVALUE PICKER dans le diagramme, auquel nous connectons le composant précédent. A l'aide du menu CONFIGURE, nous sélectionons la modalité TRUE dans la liste.



Subdivision des données en « apprentissage » et « test »

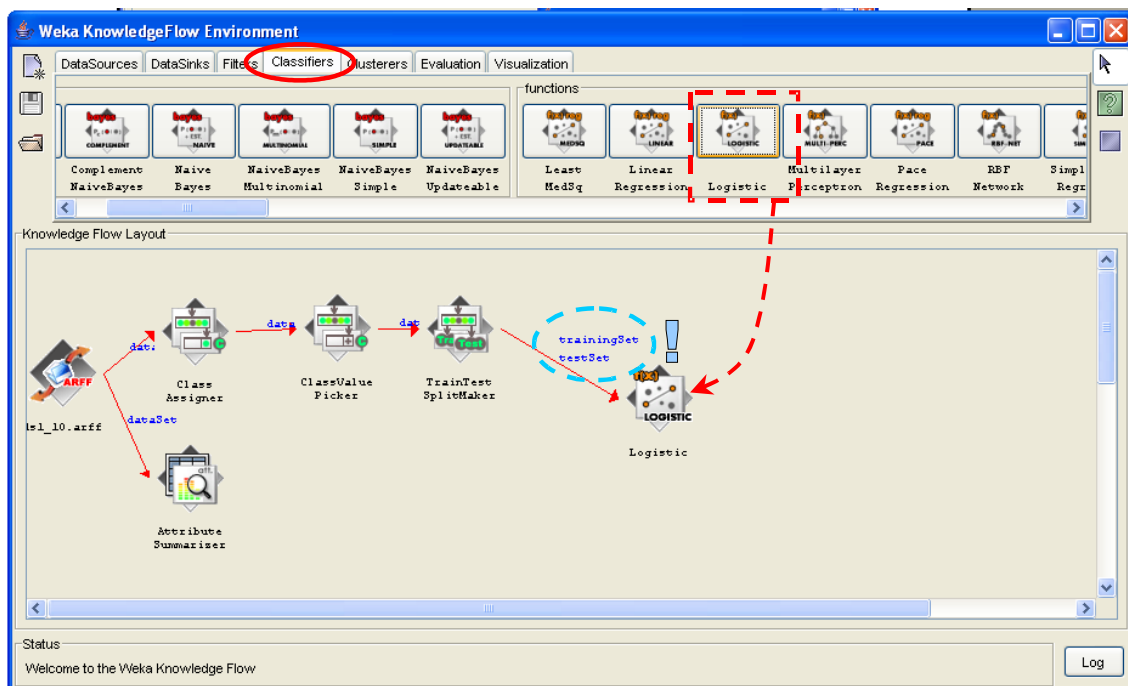
Nous voulons construire notre modèle de prédiction sur 70% de notre base, puis l'évaluer sur la courbe ROC calculée sur les 30% restants. Il est donc nécessaire de subdiviser en deux parties nos données avant de brancher la méthode d'apprentissage, le composant

TRAINTEST SPLIT MAKER (palette EVALUATION) permet de le faire. Nous l'ajoutons dans notre diagramme, puis nous le configurons.



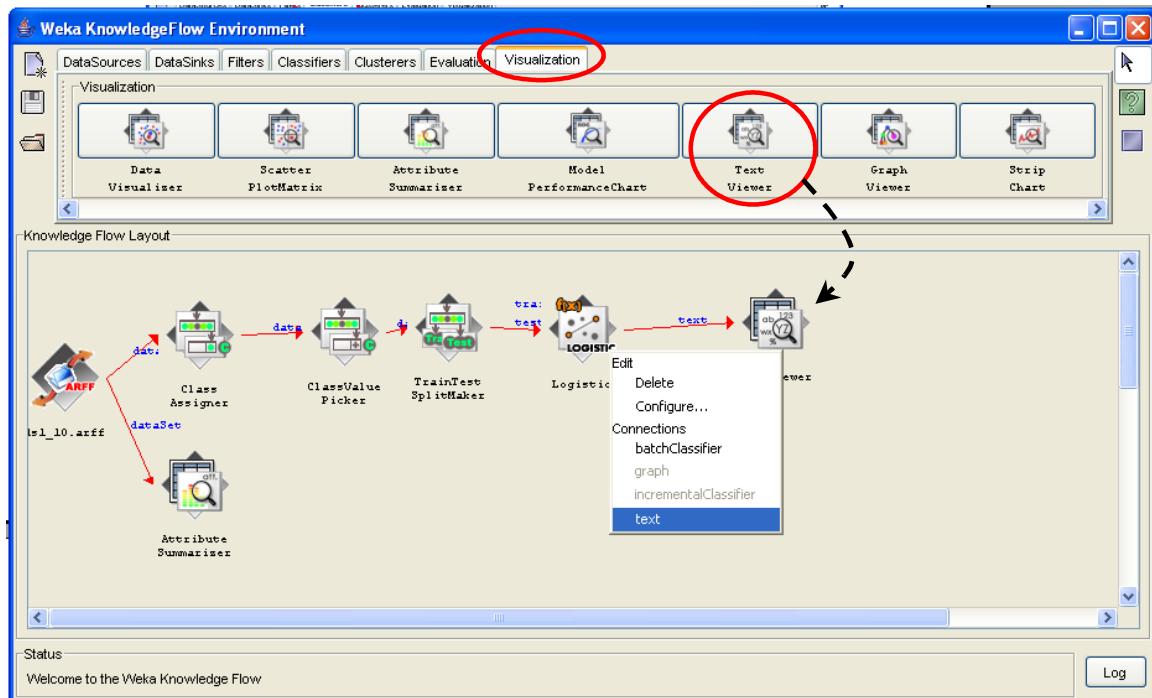
Régression logistique

La méthode Régression Logistique est située dans la palette CLASSIFIERS. Nous la plaçons dans notre diagramme, puis nous lui connectons deux fois le composant TRAIN TEST SPLIT MAKER : deux fois car nous devons transmettre les données d'apprentissage et de test. Cette particularité, assez étrange de prime abord, s'explique par le fait que c'est le même composant qui produit les ensembles apprentissage et test.

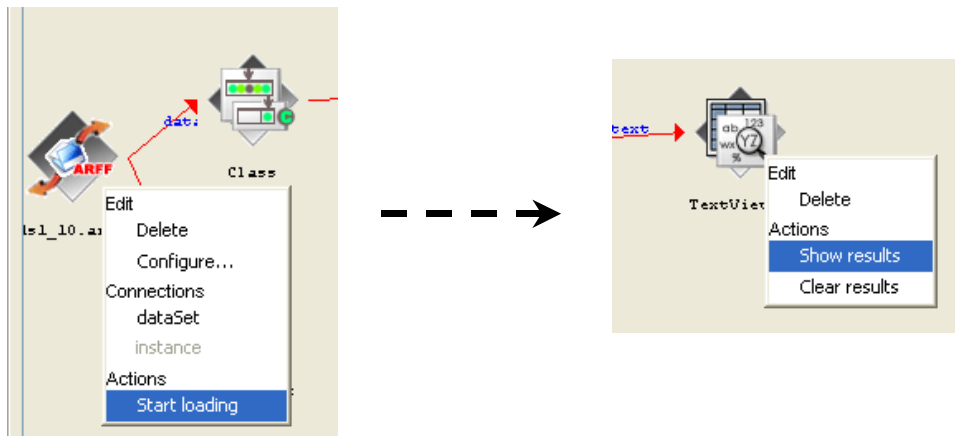


Voir les résultats

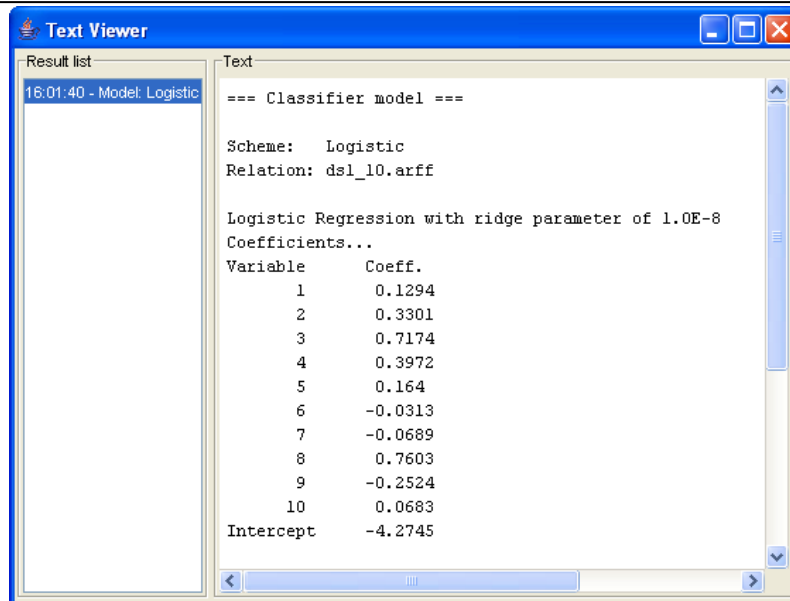
Pour voir les résultats de la régression, il faut connecter (connexion TEXT) le composant LOGISTIC au composant TEXT VIEWER (palette VISUALIZATION) que nous ajoutons dans le diagramme.



L'exécution de la chaîne de traitement s'effectue alors en activant le menu START LOADING du premier composant (ARFF LOADER). Le menu SHOW RESULTS du composant TEXTVIEWER permet d'afficher les résultats.



La fenêtre indique les paramètres du modèle estimé, en l'occurrence les coefficients de l'équation de régression. Ce composant de WEKA est peu loquace, nous ne retrouvons pas ici les indicateurs usuels de la régression logistique tel que l'évaluation globale de la régression, l'évaluation individuelle des coefficients, etc.



```

Result list
16:01:40 - Model: Logistic

=== Classifier model ===

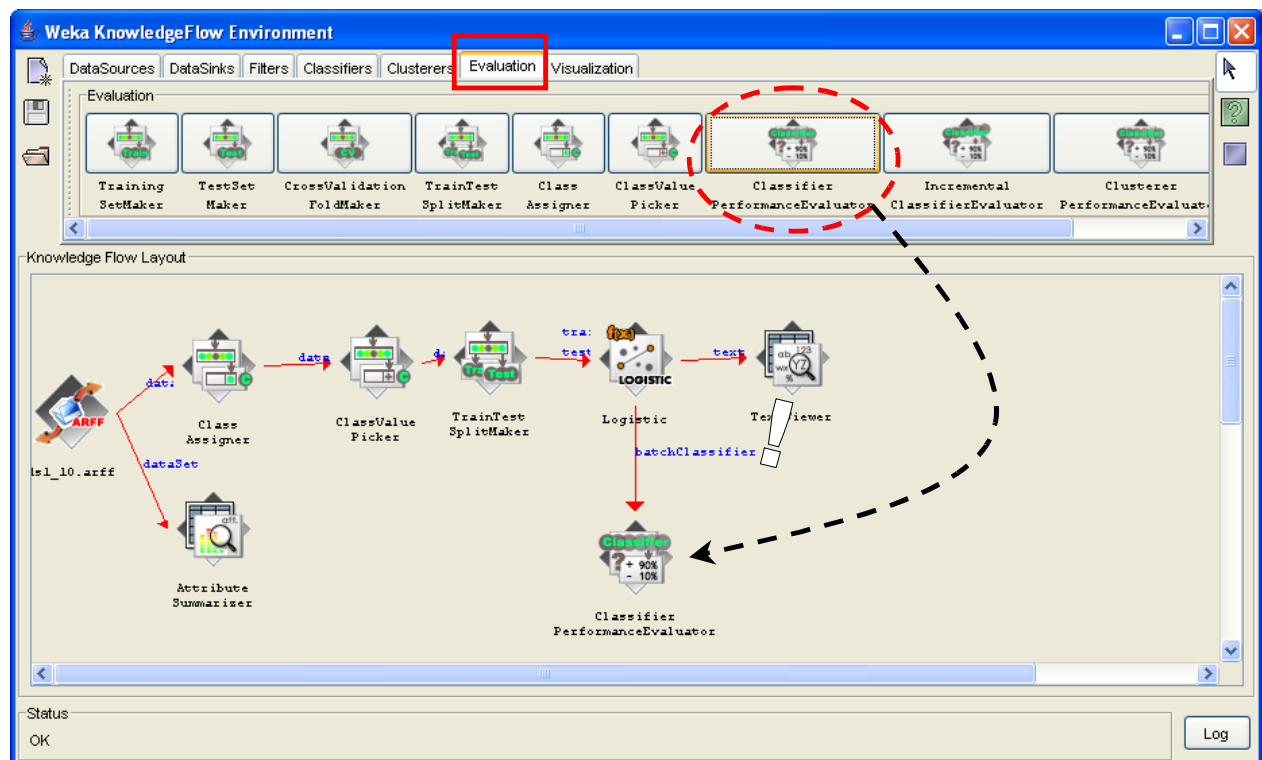
Scheme: Logistic
Relation: dsl_10.arff

Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
Variable    Coeff.
1           0.1294
2           0.3301
3           0.7174
4           0.3972
5           0.164
6          -0.0313
7          -0.0689
8           0.7603
9          -0.2524
10          0.0683
Intercept  -4.2745

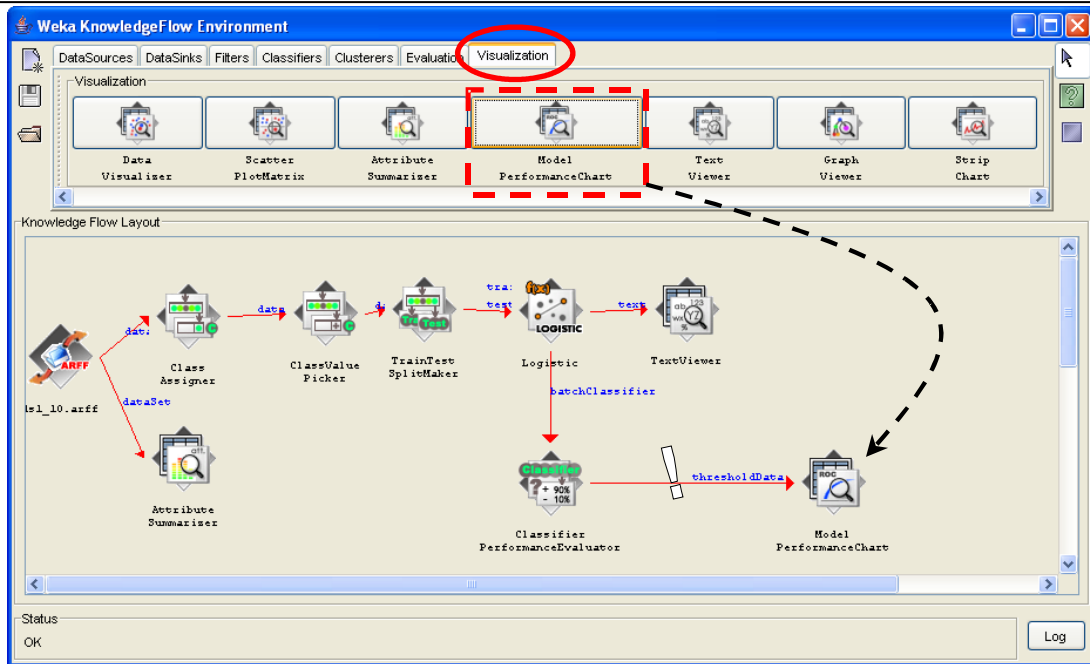
```

Construire la courbe ROC

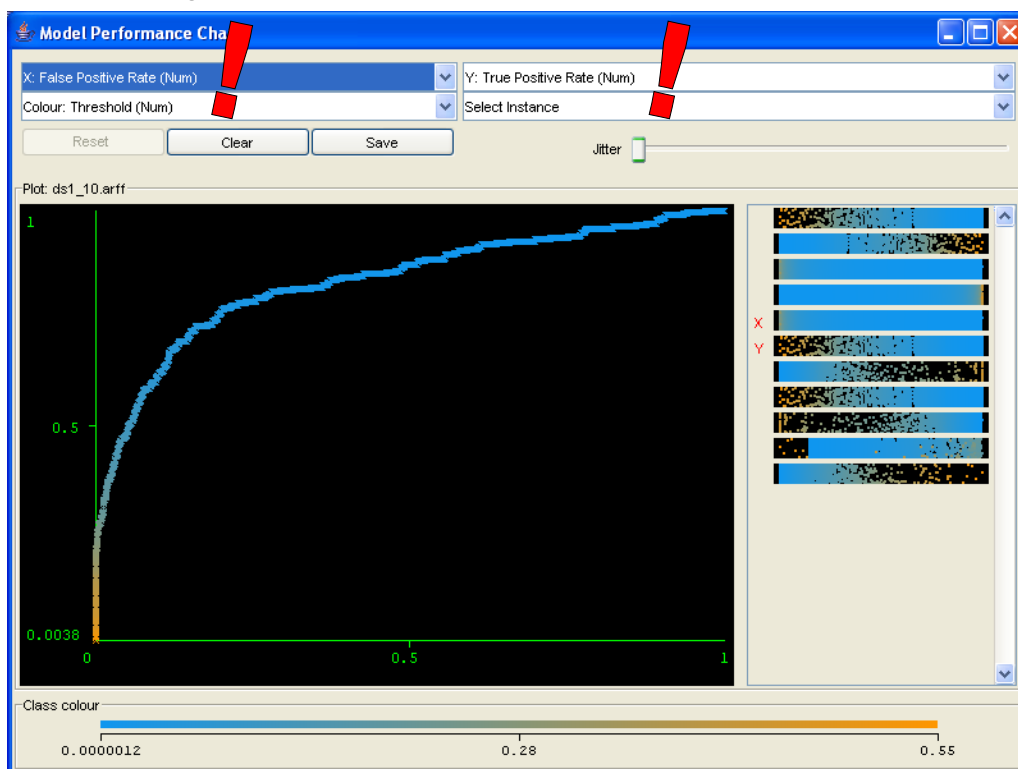
Pour évaluer un apprentissage supervisé, il faut rajouter un composant CLASSIFIER PERFORMANCE EVALUATOR (palette EVALUATION) dans le diagramme. Nous lui connectons la sortie BATCH CLASSIFIER du composant LOGISTIC.



Puis nous rajoutons le composant MODEL PERFORMANCE CHART (palette VISUALIZATION), nous lui relions le composant précédent en veillant à utiliser la connexion THRESHOLD DATA (très important !).



Il reste alors à relancer de nouveau l'exécution du diagramme (START LOADING de ARFF LOADER), puis visualiser les résultats en activant le menu SHOW PLOT du composant MODEL PERFORMANCE CHART. Une fenêtre apparaît, la courbe ROC est la courbe par défaut (le **taux de faux positifs** en abscisse et le **taux de vrais positifs** en ordonnée). Nous pouvons modifier le graphique en choisissant d'autres variables en abscisse et en ordonnée.



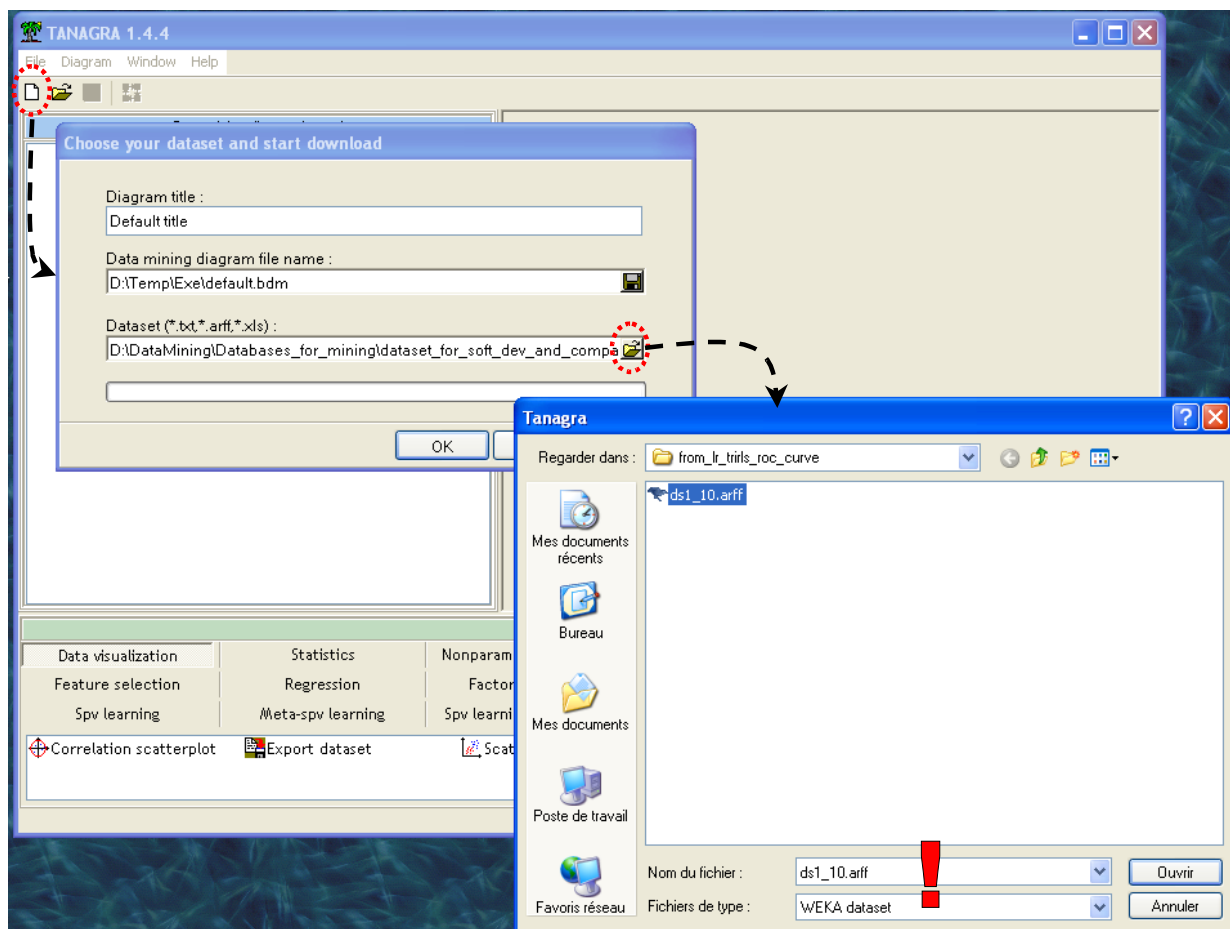
WEKA est manifestement un outil très puissant. Mais comme nous avons pu le constater dans ce didacticiel, il faut bien le connaître pour en exploiter toutes les possibilités.

Construire la courbe ROC avec TANAGRA

TANAGRA nous est maintenant familier. Au lancement du logiciel, l'interface est subdivisée en trois parties : à gauche l'espace dans lequel nous définirons la séquence de traitements ; en bas les palettes de composants ; au centre l'espace dévolu à l'affichage des résultats.

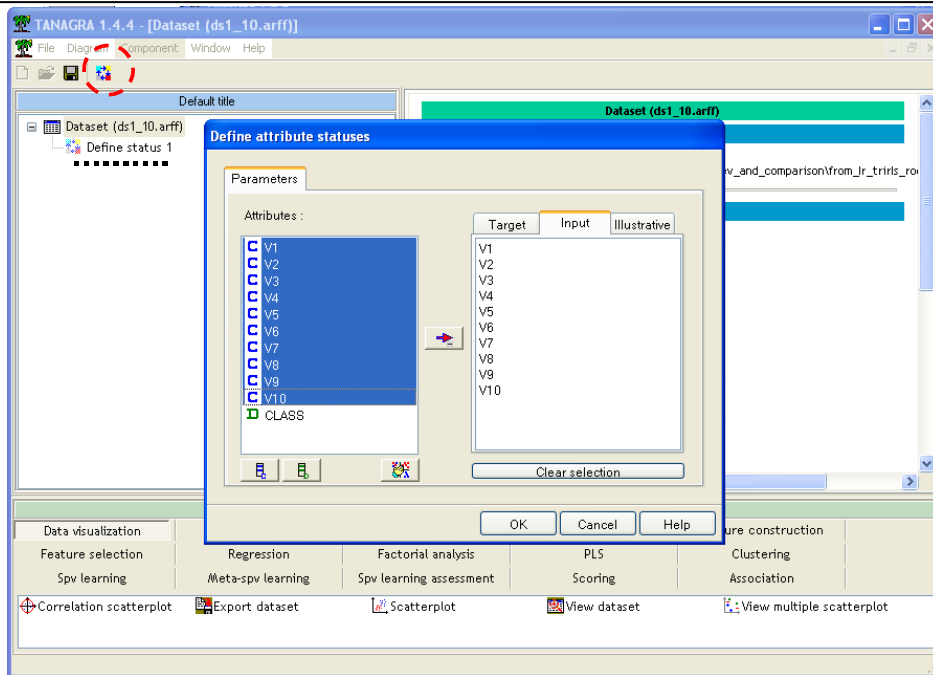
Importer les données

Nous activons le menu FILE / NEW pour construire un nouveau diagramme de traitements. Nous sélectionnons le type de fichier ARFF pour charger les mêmes données que précédemment.

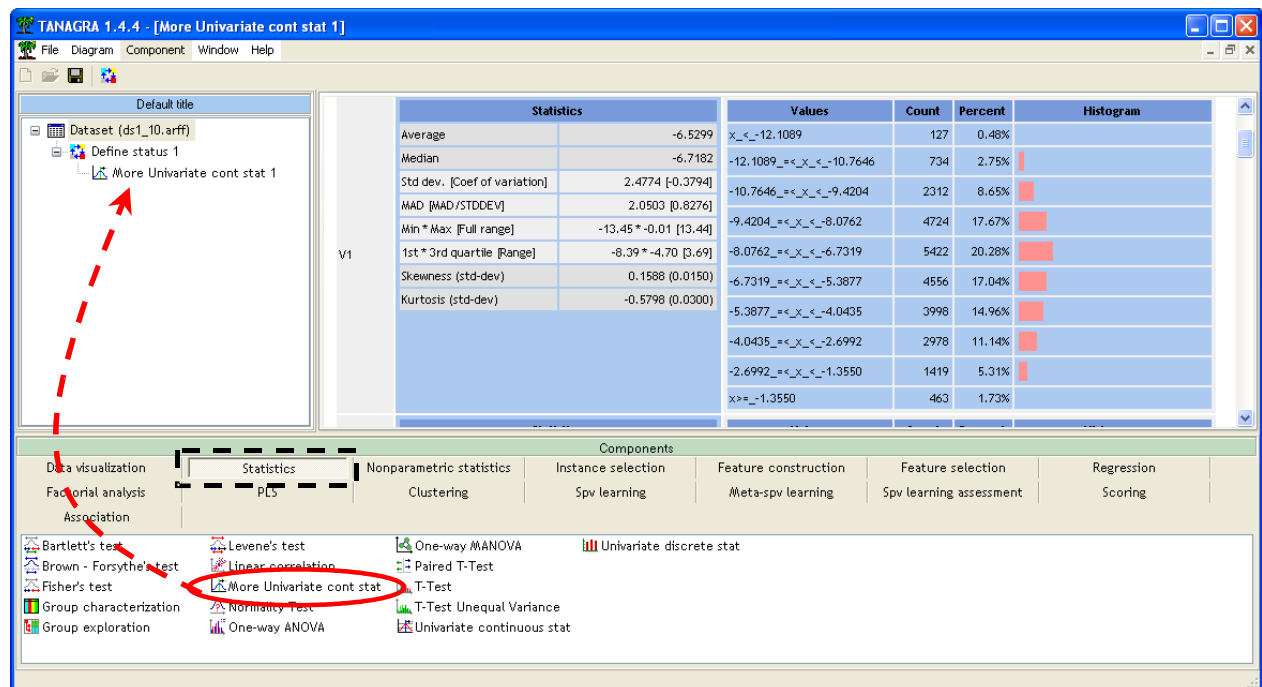


Statistiques descriptives

Pour réaliser les statistiques descriptives sur les variables prédictives, nous devons placer le composant DEFINE STATUS (nous utilisons le raccourci dans la barre d'icône), puis nous plaçons en INPUT toutes les variables continues.

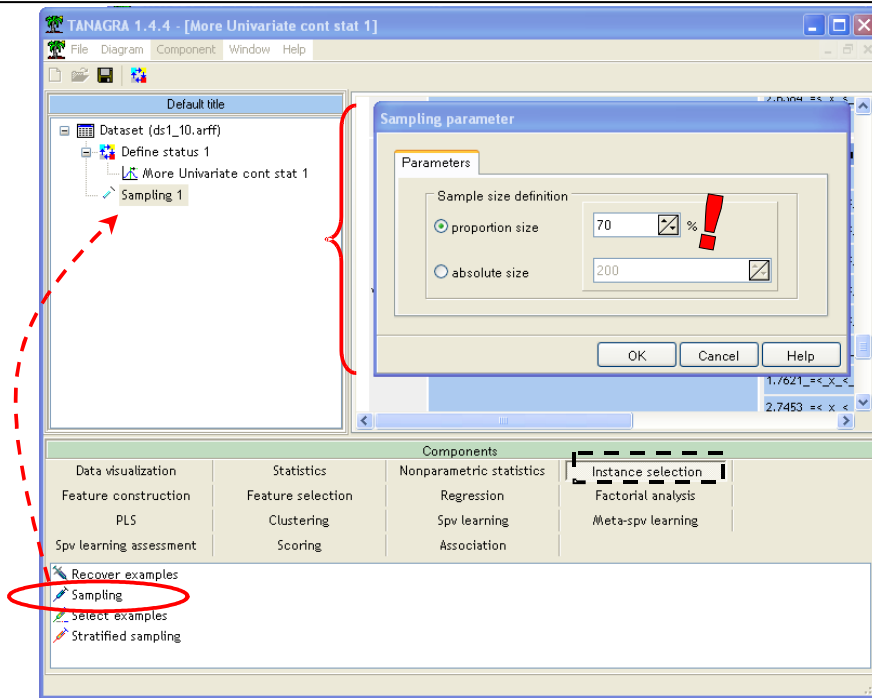


Nous rajoutons alors dans le diagramme le composant MORE UNIVARIATE CONT STAT (palette STATISTICS) que nous activons avec le menu VIEW. La fenêtre principale retrace alors la description détaillée de chaque variable INPUT, la distribution de fréquence et les principaux indicateurs statistiques.



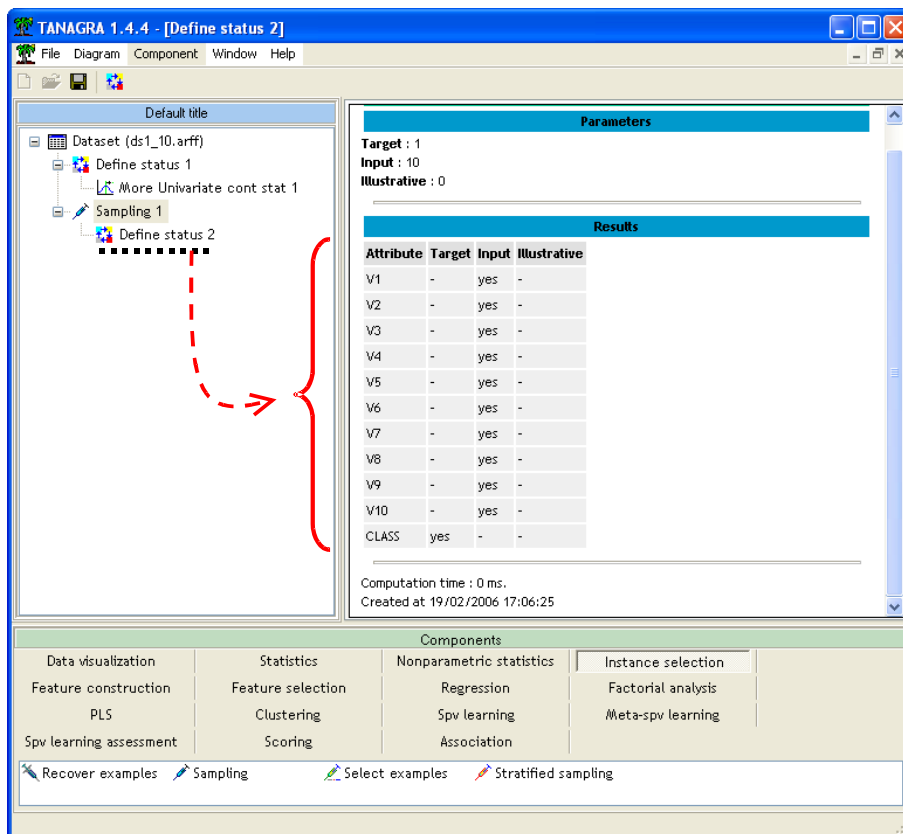
Subdiviser les données en « apprentissage » et « test »

Revenons sur le premier composant du diagramme (la racine, représentant les données). Nous insérons le composant SAMPLING (palette INSTANCE SELECTION), puis nous définissons les tailles respectives des ensembles de données en activant le menu PARAMETERS.



Choisir la variable à prédire et les descripteurs

Nous devons maintenant définir la variable à prédire (TARGET) et les variables prédictives (INPUT). Nous plaçons, à la suite du composant précédent, l'icône DEFINE STATUS (le mieux est de passer par le raccourci dans la barre d'outil). Nous obtenons le résultat suivant en cliquant sur le menu VIEW du composant.



Apprentissage supervisé

L'étape suivante consiste à placer la régression logistique dans le diagramme. Cette opération se fait en deux étapes dans TANAGRA : placer dans un premier temps une enveloppe « meta-apprentissage » (SUPERVISED LEARNING de la palette META-SPV LEARNING), puis y intégrer le composant LOG-REG TRIRLS (palette SPV LEARNING). Ce composant d'apprentissage provient de la bibliothèque de KOMAREK (<http://komarix.org/ac/lr>). Par rapport à nos propres implémentations, il est particulièrement rapide. Vous le constatez en lançant l'exécution (menu VIEW du composant), les résultats s'affichent dans la fenêtre dédiée.

The screenshot shows the TANAGRA 1.4.4 interface. On the left, a workflow diagram includes 'Dataset (ds1_10.arff)', 'Define status 1', 'More Univariate cont stat 1', 'Sampling 1', 'Define status 2', and 'Supervised Learning 1 (Log-Reg TRIRLS)'. A red dashed arrow points from this component to the results window on the right.

The results window, titled 'Supervised Learning 1 (Log-Reg TRIRLS)', displays the following performance metrics:

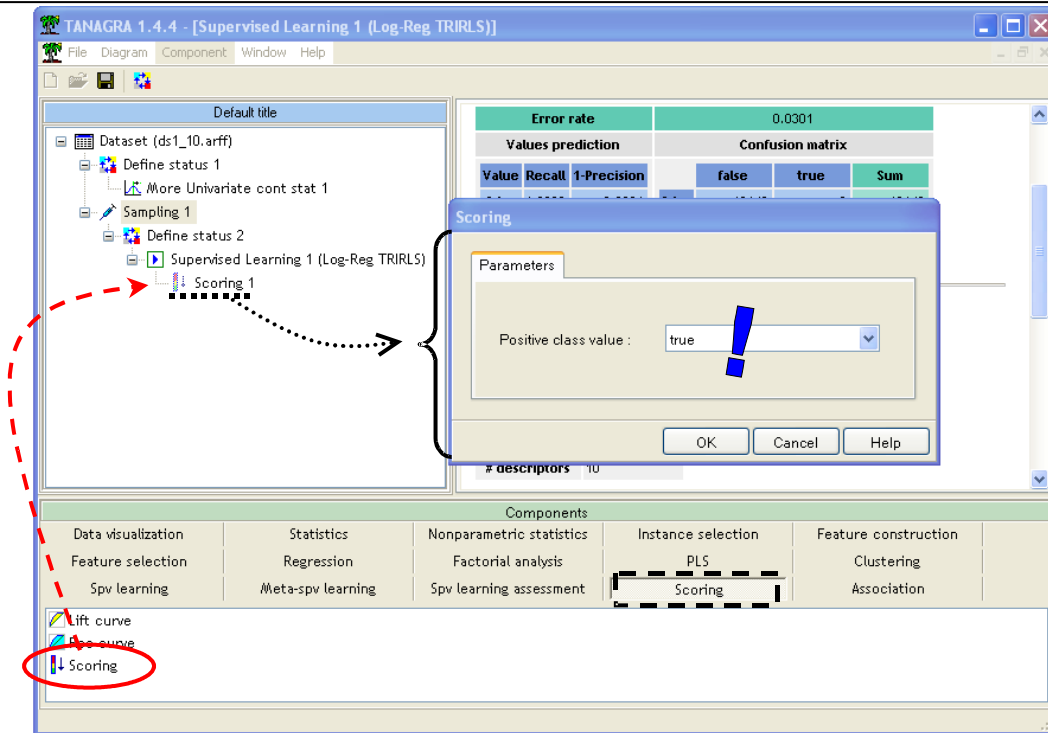
Error rate		Confusion matrix			
0.0301					
Values prediction					
Value	Recall	1-Precision	false	true	Sum
false	1.0000	0.0301	18149	0	18149
true	0.0000	1.0000	564	0	564
Sum			18713	0	18713

At the bottom, the 'Components' palette is visible, with 'Log-Reg TRIRLS' circled in red. Other components include Binary logistic regression, C4.5, C-RT, C-SVC, Decision List, ID3, K-NN, Linear discriminant analysis, Multilayer perceptron, Multinomial Logistic Regression, Naive bayes, Prototype-NN, Radial basis function, and SVM.

Il ne faut pas s'alarmer outre mesure de l'aspect peu engageant de la matrice de confusion, il ne faut pas oublier que la distribution des classes est très déséquilibrée, le plus important pour nous est que les « positifs » soient réellement placés devant les négatifs lorsque nous voudrions les classer selon leur score d'appartenance aux positifs.

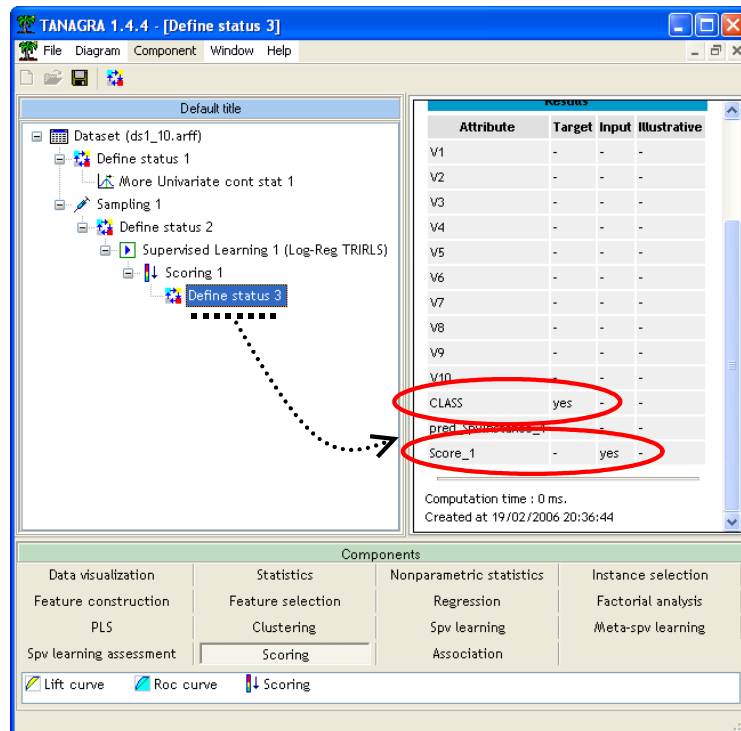
Calculer les scores

Justement, nous voulons spécifier la modalité de la variable à prédire puis calculer les scores. Nous ajoutons le composant SCORING (palette SCORING) dans la chaîne de traitements, nous la configurons (menu PARAMETERS) de manière à préciser que les « TRUE » correspondent aux positifs.

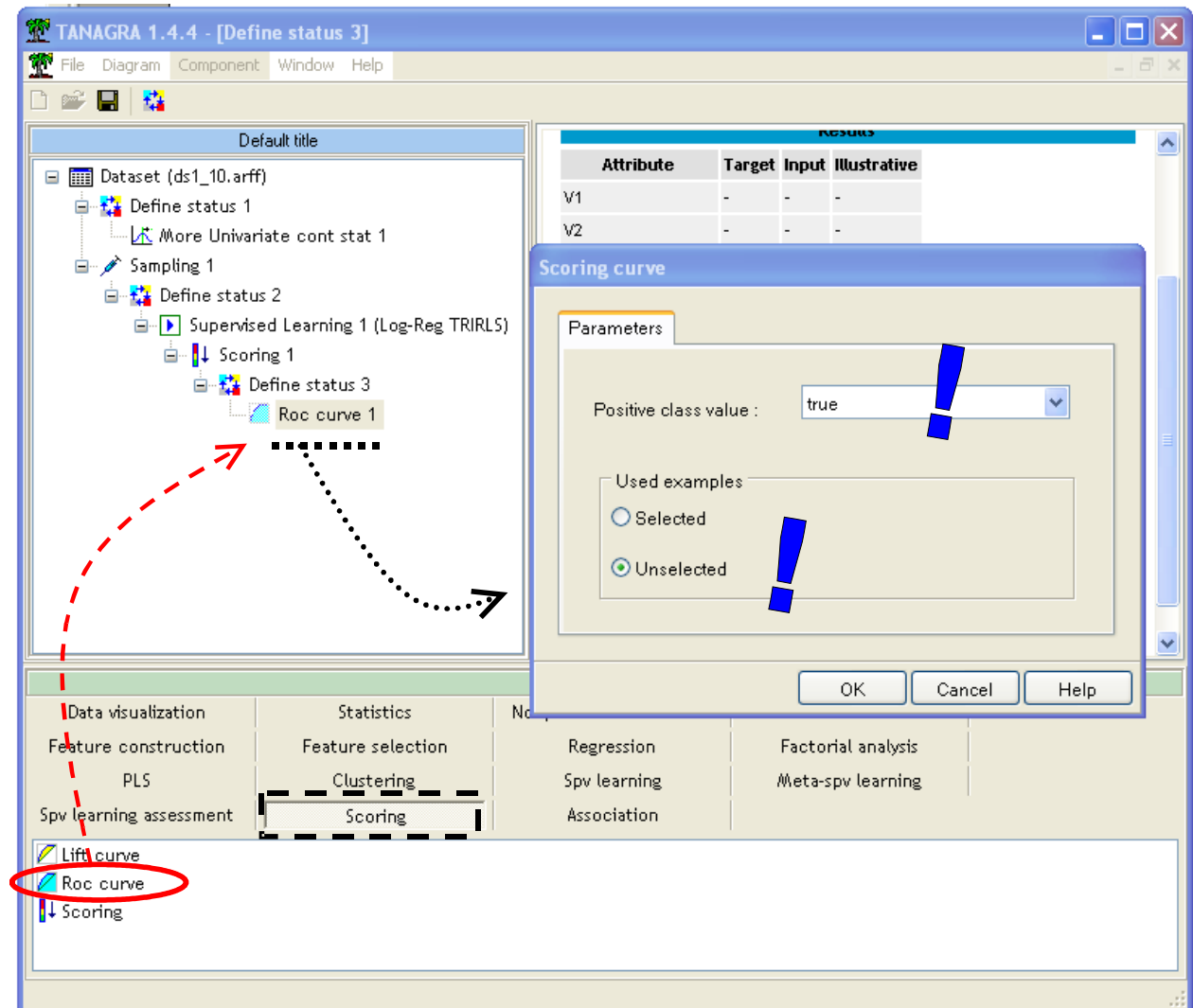


Construire la courbe ROC

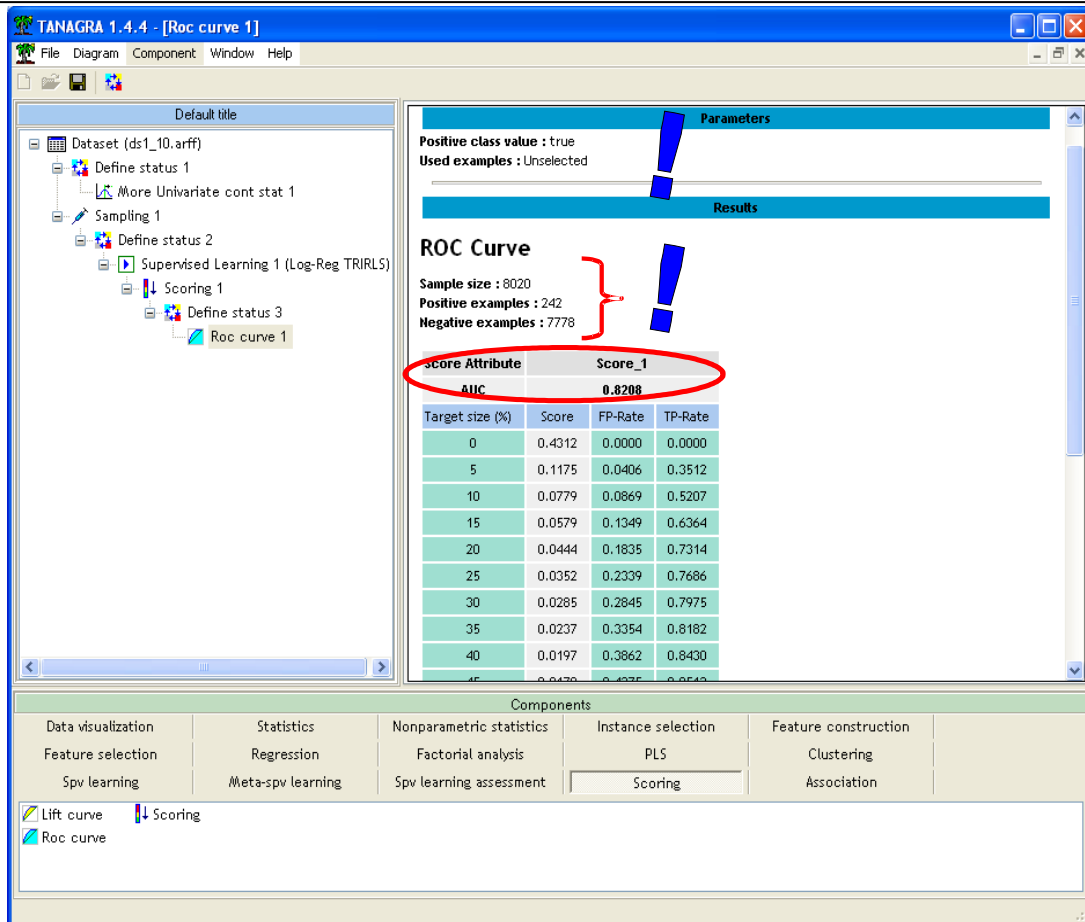
Il ne reste plus qu'à calculer la courbe ROC. Nous insérons de nouveau un composant DEFINE STATUS dans le diagramme, en TARGET nous plaçons l'attribut CLASS, en INPUT, l'attribut SCORE_1 générée par le composant de scoring. Notons qu'il est possible de définir plusieurs attributs INPUT, cela permet de comparer les performances de plusieurs stratégies d'apprentissage.



Il ne reste plus qu'à placer le composant ROC CURVE dans le diagramme, préciser que la classe positive correspond à la modalité TRUE de la variable à prédire et que le calcul doit être réalisé sur les individus non-sélectionnés, c.-à-d. n'ayant pas participé à l'apprentissage.



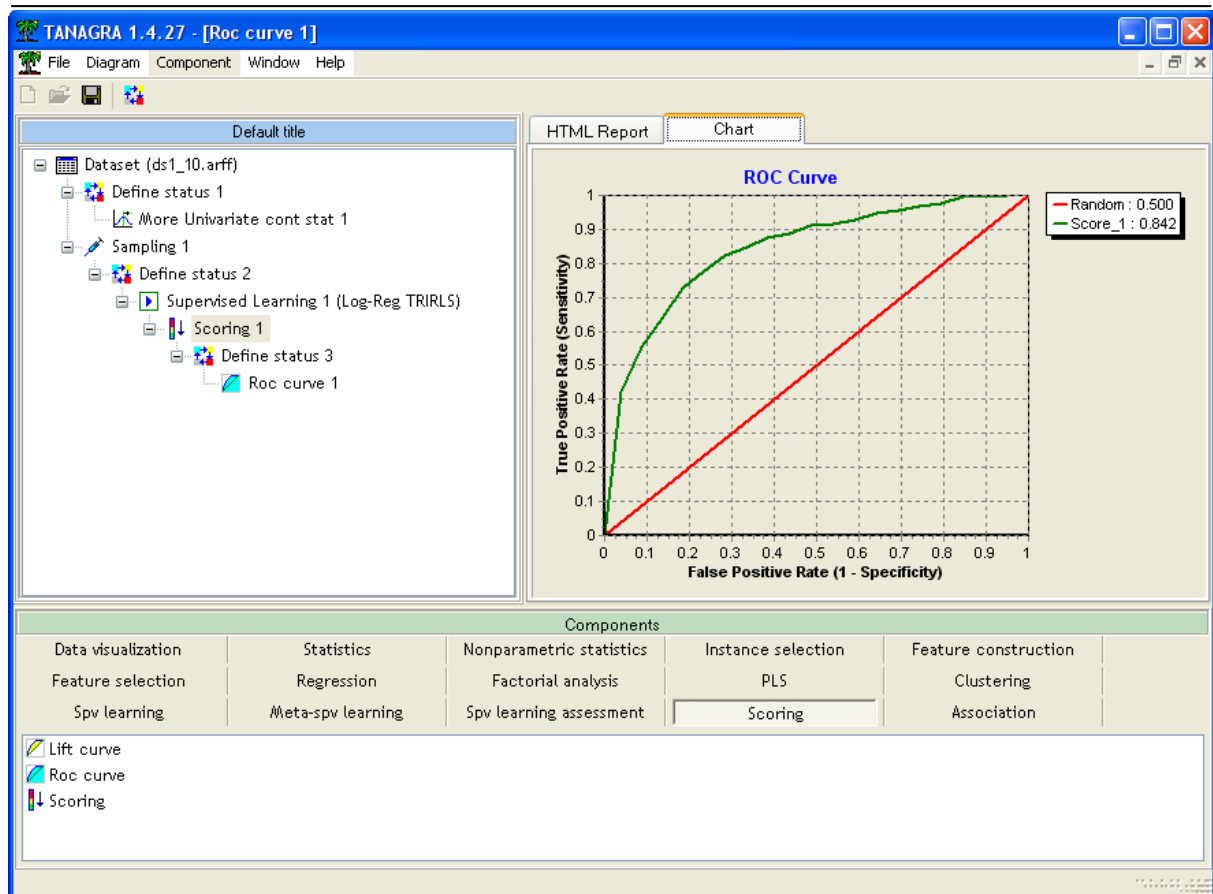
Le résultat s'affiche dans la fenêtre de résultats en activant le menu VIEW du composant. La courbe elle-même n'est pas dessinée, en revanche nous disposons des informations adéquates pour réaliser le graphique dans un tableur quelconque. D'autres indicateurs (taille de l'échantillon utilisé, nombre de positifs, AUC, etc.) sont disponibles.



TANAGRA permet de mettre en œuvre assez rapidement la construction de la courbe ROC. A noter le rôle central de DEFINE STATUS avant le composant ROC CURVE, il permet de comparer différents critères de classement des observations selon leur probabilité d'être positif : ces critères peuvent être calculés à partir d'une méthode d'apprentissage, comme c'est le cas ici, ils peuvent également correspondre à n'importe quelle variable déjà présente dans la base.

Remarque : Depuis la version 1.4.21, Tanagra fournit directement une représentation graphique de la courbe ROC¹

¹ La partition apprentissage – test étant aléatoire, la valeur AUC est forcément très légèrement différente dans cette copie d'écran plus récente. Mais nous sommes bien dans des ordres de grandeurs identiques, c'est ce qui importe.

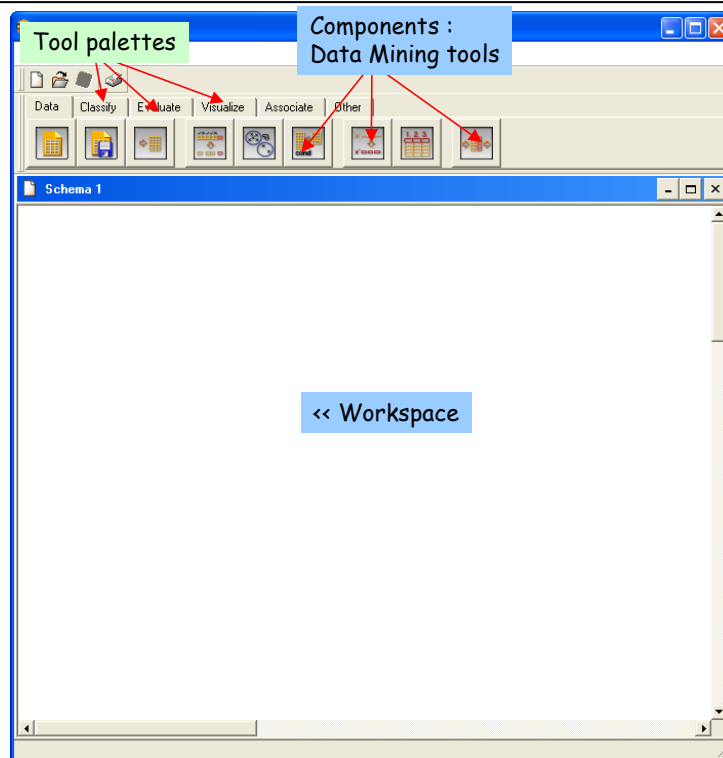


Construire la courbe ROC avec ORANGE

Nous avons choisi de présenter ORANGE en dernier car il se démarque considérablement des deux logiciels précédents par sa simplicité d'utilisation, sans pour autant faire de concessions sur la précision des calculs. Il est particulièrement adapté pour le type de traitement que nous avons choisi dans ce didacticiel. En revanche, le prix à payer pour cette simplicité est que certaines options, définies par défaut, ne sautent pas aux yeux, il importe alors de bien savoir à quels endroits du logiciel elles sont précisées pour que l'utilisateur comprenne exactement ce qu'il est en train de faire.

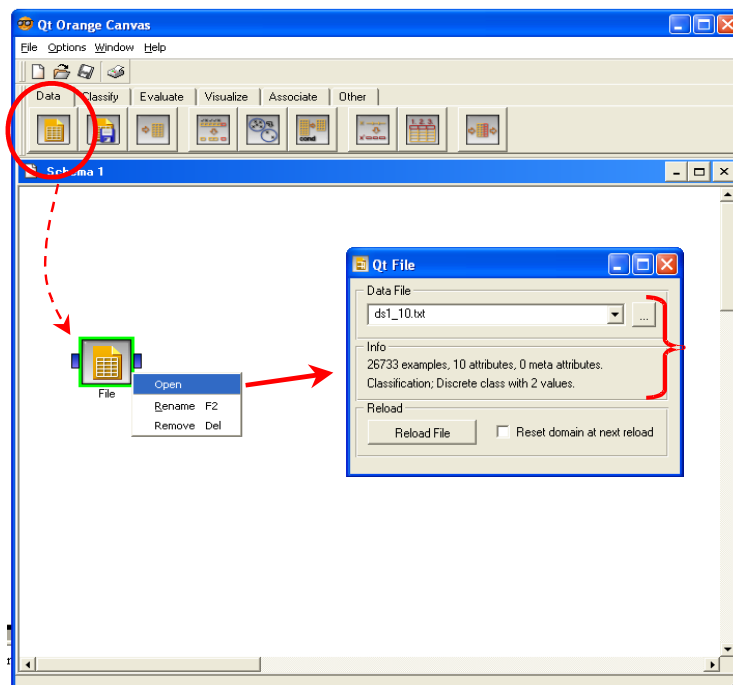
Lancement du logiciel

Tout comme WEKA, ORANGE propose une interface composée de deux parties distinctes : un espace pour définir les traitements ; une palette d'outils située dans la partie haute de la fenêtre principale.



Importer les données

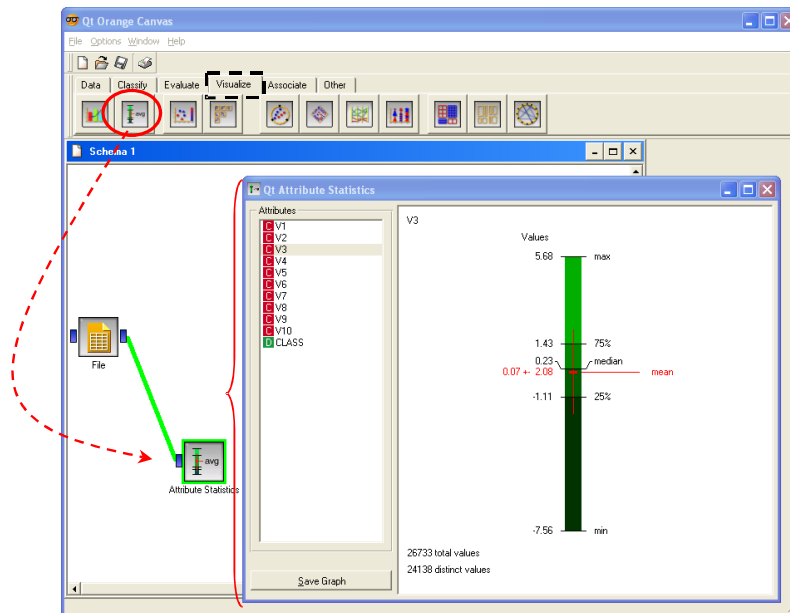
ORANGE peut importer des données au format texte (séparateur tabulation). Il suffit de cliquer sur l'outil pour qu'il s'insère automatiquement dans l'espace de travail.



Statistiques descriptives

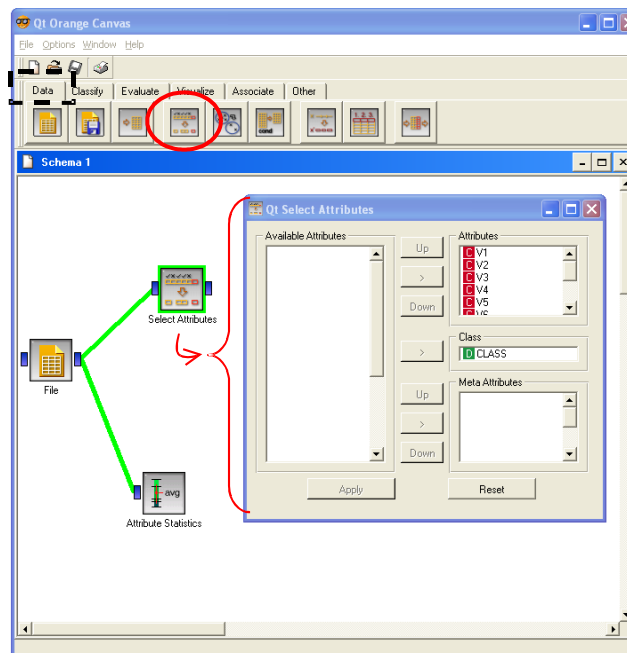
ORANGE propose de séduisants outils graphiques pour représenter les distributions de données. Le temps de calcul est en revanche un peu élevé. Nous insérons le composant dans

le diagramme, nous lui connectons la source de données, le calcul est lancé automatiquement. Le résultat apparaît dans une fenêtre flottante lorsque nous activons le menu contextuel OPEN du composant.



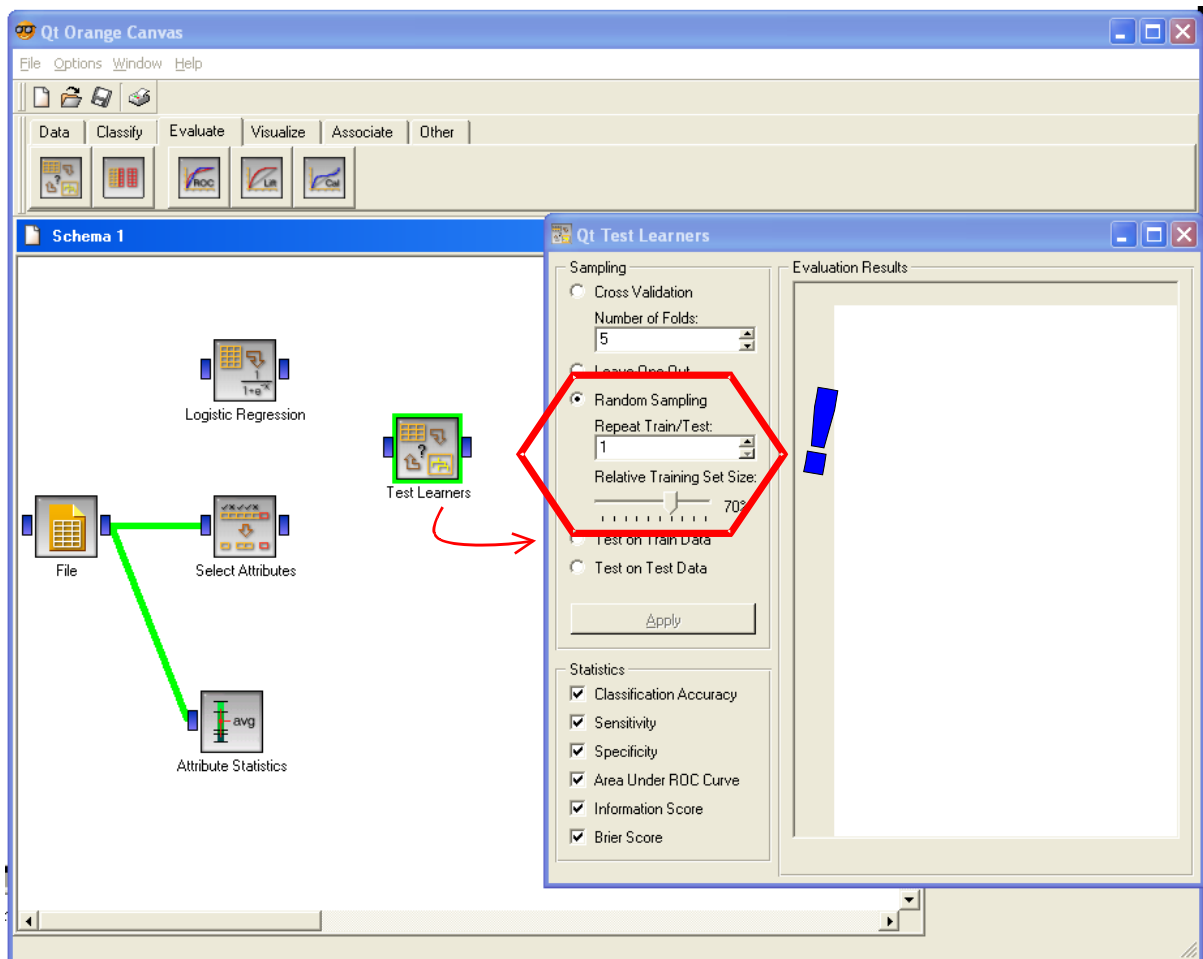
Définir la variable à prédire

Par défaut, tout comme dans WEKA, la variable à prédire est la dernière colonne. Il est cependant possible de modifier la sélection en plaçant le composant SELECT ATTRIBUTES (palette DATA). Nous nous contentons de vérifier que la sélection adéquate a été réalisée sur nos données.

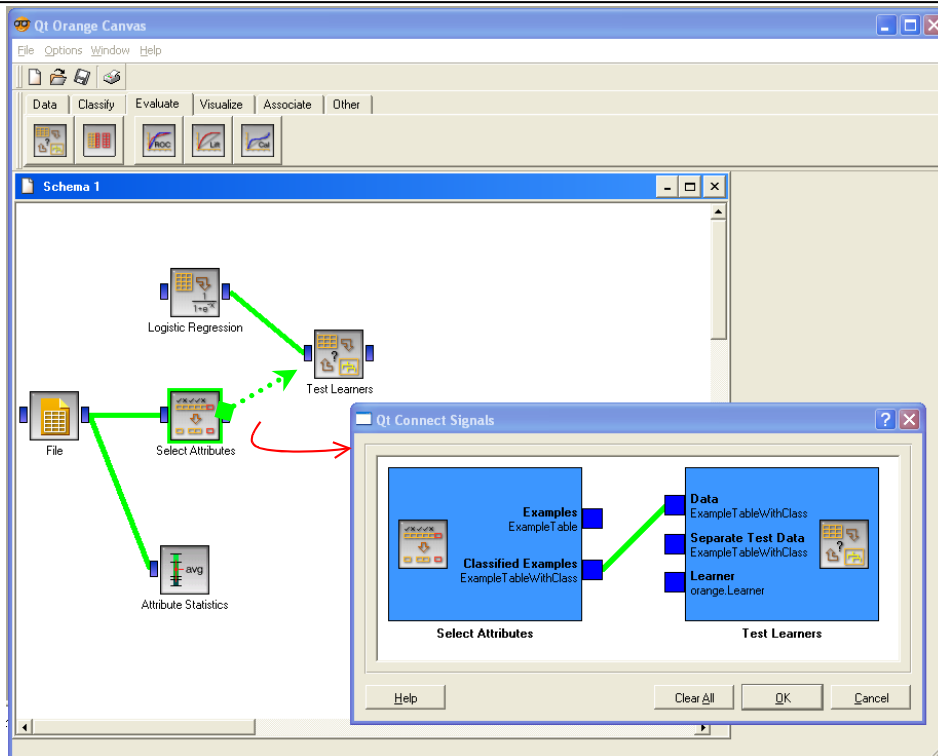


Méthode d'apprentissage et l'outil d'évaluation

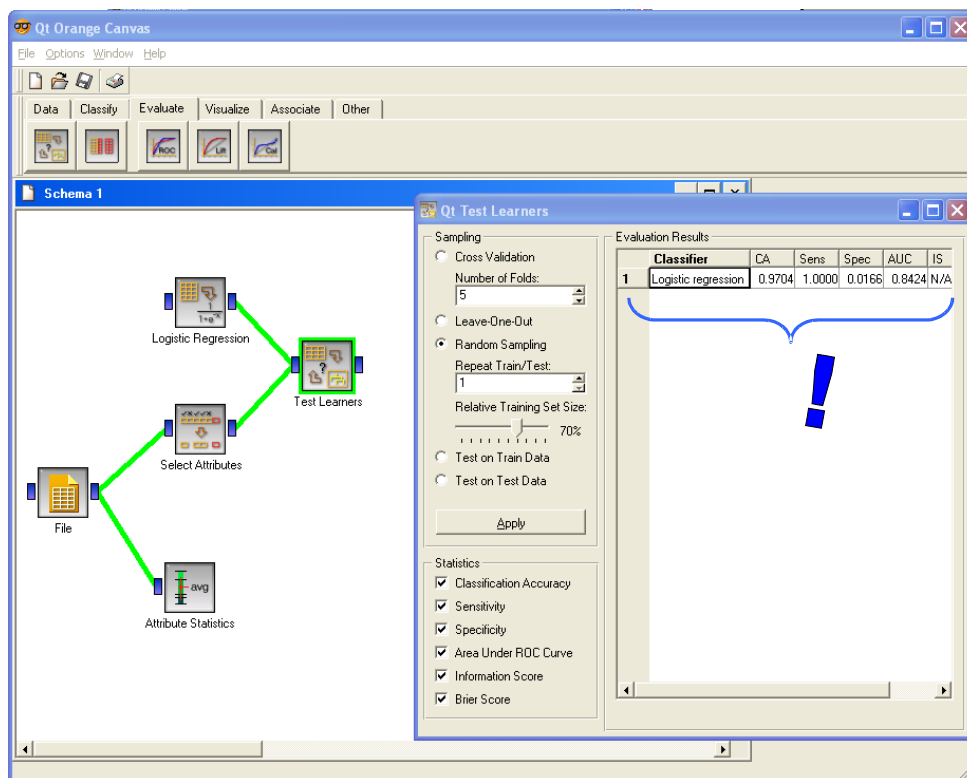
Nous plaçons le composant LOGISTIC REGRESSION de la palette CLASSIFY, il n'y pas de paramétrage particulier à spécifier. Puis nous plaçons le composant TEST LEARNER de la palette EVALUATE. Sur ce dernier composant, il est nécessaire de préciser que l'évaluation sera basée sur une subdivision « apprentissage » - « test » des données.



Il ne reste plus qu'à connecter LOGISTIC REGRESSION à TEST LEARNERS, puis SELECT ATTRIBUTES à TEST LEARNERS. Lors de cette dernière connexion, une boîte de confirmation apparaît, permettant de préciser qu'il s'agit bien de notre ensemble de données.



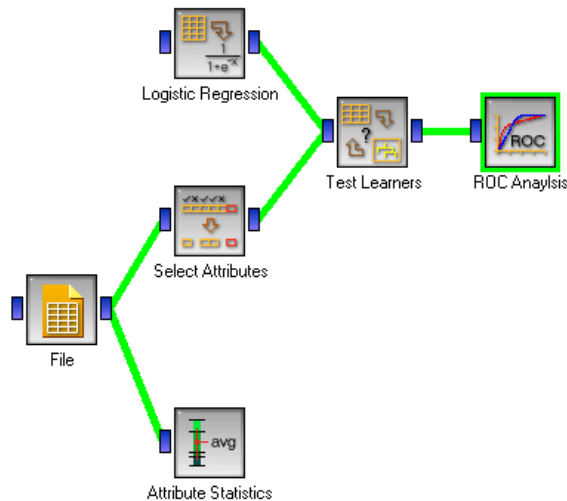
Après confirmation, le calcul est directement réalisé. Il faut alors activer le menu OPEN de TEST LEARNERS pour visualiser les résultats.



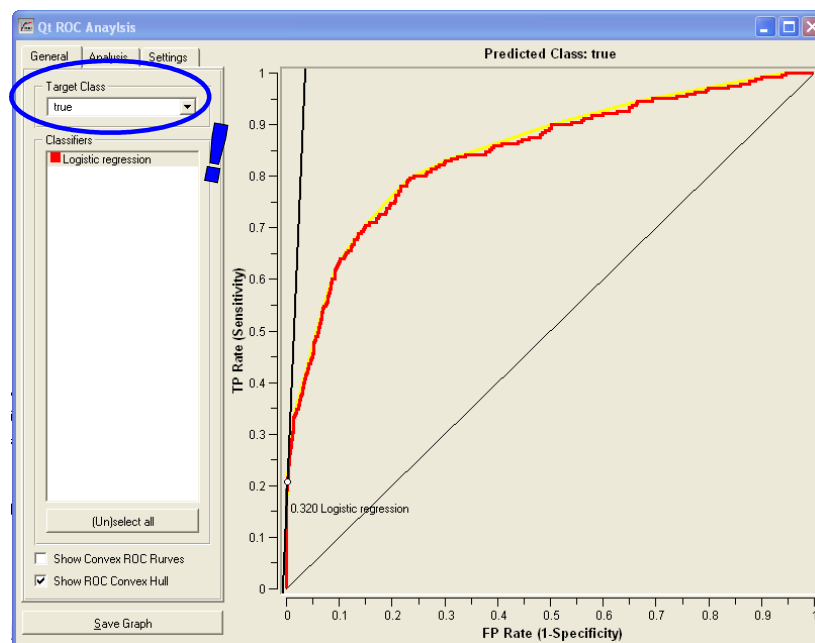
Nous constatons que plusieurs indicateurs sont calculés, dont l'AUC. Nous n'avons pas trouvé en revanche si ces valeurs considéraient les TRUE ou les FALSE comme des « positifs ».

Construire la courbe ROC

Plus intéressant pour nous, ORANGE dispose d'un outil graphique pour l'élaboration de la courbe ROC : c'est le composant ROC ANALYSIS de la palette EVALUATE. Nous l'insérons dans le diagramme puis nous lui connectons le composant TEST LEARNERS.



En activant le menu OPEN de ROC ANALYSIS, nous pouvons visualiser la courbe ROC, en précisant de manière explicite cette fois-ci la valeur de l'attribut CLASS correspondant à la modalité positive. Notons que cet outil a été dimensionné pour comparer plusieurs courbes en provenance de plusieurs méthodes d'apprentissage, sa facilité d'utilisation est remarquable.



Par rapport à WEKA et TANAGRA, nous noterons le nombre réduit de composants à placer dans le diagramme pour réaliser la courbe ROC. Nous noterons également que le

même dispositif permet à ORANGE de comparer un grand nombre de classifieurs, il suffit de rajouter une icône par méthode et de les connecter sur le composant TEST LEARNERS.

Conclusion

Chacun à leur manière, ces logiciels ont permis de construire une courbe ROC évaluant la qualité de la régression logistique sur un ensemble de données.

Pour que les résultats soient réellement comparables, il aurait fallu subdiviser le fichier en apprentissage et test puis de lancer les logiciels sur les mêmes ensembles de données.

ORANGE et WEKA adoptent la même stratégie dans ce cas, les données doivent être scindées en deux fichiers, le diagramme comportera alors deux sources de données, ce qui ne pose aucun problème étant donné la présentation sous forme de graphe du diagramme de traitements.

TANAGRA adopte une stratégie différente : une nouvelle colonne, spécifiant l'appartenance des observations à la partie « apprentissage » ou « test », doit être insérée dans le fichier de données. Il faut alors substituer le composant SELECT EXAMPLES au composant SAMPLING dans le diagramme de traitements, le reste n'est pas modifié.