Objectif

Construire une courbe ROC (Receiver Operating Characteristics) dans TANAGRA.

Une courbe ROC permet de comparer des algorithmes d'apprentissage indépendamment (1) de la distribution des modalités de la variable à prédire dans le fichier test, (2) des coûts de mauvaises affectations.

La courbe ROC est avant tout définie pour les problèmes à deux classes (les positifs et les négatifs), elle indique la capacité du classifieur à placer les positifs devant les négatifs. Elle met en relation dans un graphique les taux de faux positifs (en abscisse) et les taux de vrais positifs (en ordonnée).

Sa construction s'appuie donc sur les probabilités d'être positif fournies par les classifieurs. Il n'est pas nécessaire que ce soit réellement une probabilité, une valeur quelconque dite « score » permettant d'ordonner les individus suffit amplement.

Il est possible de dériver un indicateur synthétique à partir de la courbe ROC, il s'agit de l'AUC (Area Under Curve – Aire Sous la Courbe), elle indique la probabilité d'un individu positif d'être classé devant un individu négatif. Il existe une valeur seuil, si l'on classe les individus au hasard, l'AUC sera égal à 0.5.

Dans ce didacticiel, nous construisons les courbes ROC de deux méthodes d'apprentissage que nous voulons comparer sur un problème de détection de maladies cardio-vasculaires : l'analyse discriminante linéaire (LDA) et les machines à vecteur de supports (SVM).

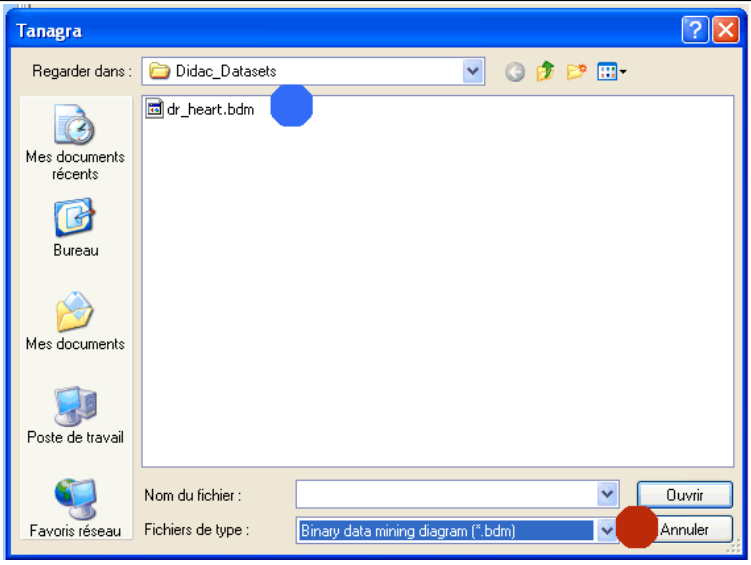
Fichier

Nous utilisons le fichier HEART déjà présenté dans d'autres didacticiels, il s'agit de prédire l'occurrence d'une maladie cardio-vasculaire à partir des caractéristiques des patients.

Construire une courbe ROC

Charger le fichier de données

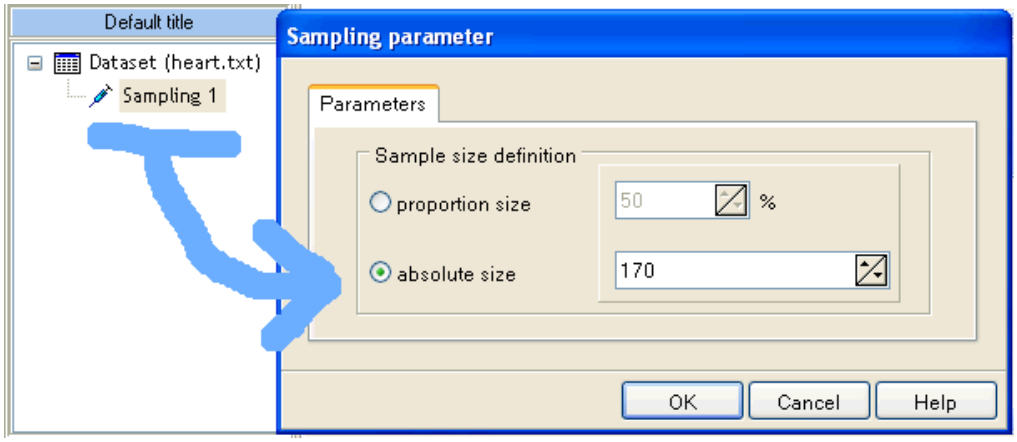
Ouvrir le fichier de données DR_HEART.BDM à partir du menu « File / Open ».



Scinder les données en « apprentissage – test »

Afin d'obtenir une estimation non biaisée des performances des méthodes d'apprentissage, nous allons scinder notre ensemble de données en deux parties : 170 observations pour la construction des modèles de prédiction ; 100 observations pour leur évaluation.

Pour ce faire, nous utilisons le composant SAMPLING, nous le paramétrons comme suit.

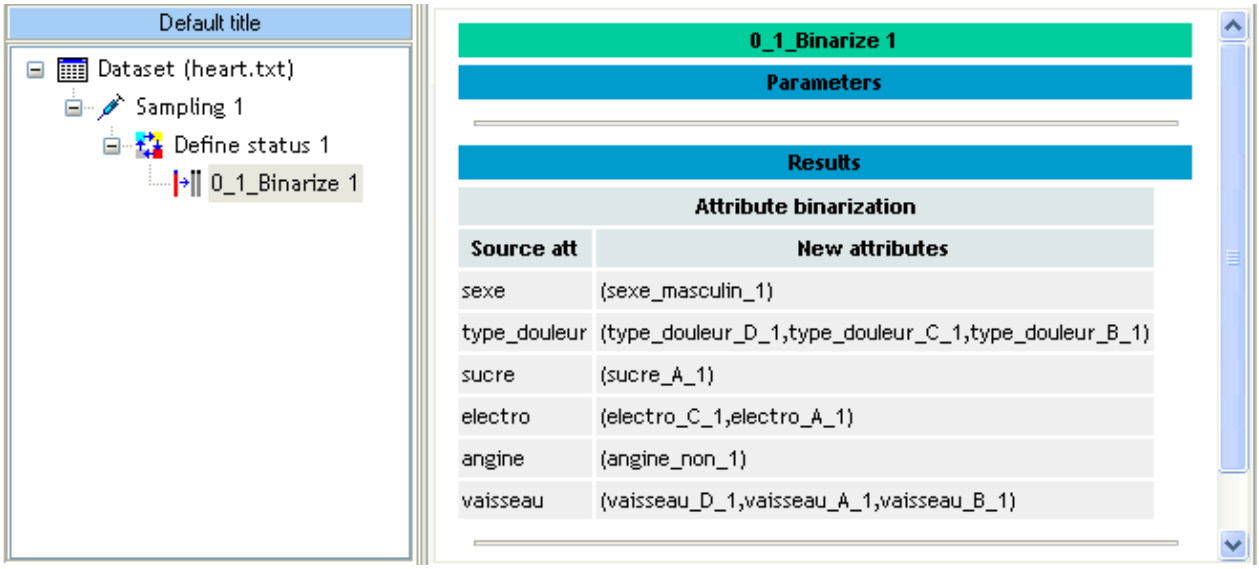


Préparer les données

Les deux méthodes que nous voulons comparer n'acceptent que les descripteurs continus. Afin de rendre possible l'apprentissage, nous décidons de procéder à un codage disjonctif (0/1 pour la présence/absence d'une modalité de la variable) des descripteurs discrets.

Nous utilisons pour cela le composant 0_1_BINARIZE après avoir sélectionné tous les attributs discrets, mis à part la variable CŒUR bien sûr qui est la variable à prédire dans notre problème.

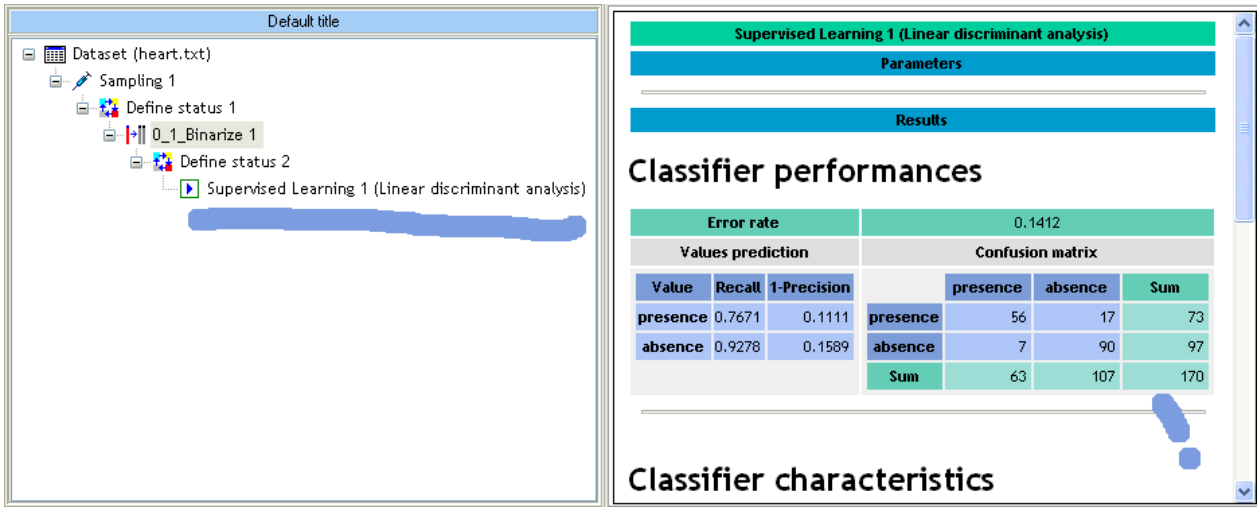
Nous obtenons le diagramme de traitement suivant.



LDA

Il est maintenant possible de lancer l'apprentissage de la LDA : placez en INPUT tous les attributs continus et en TARGET l'attribut à prédire CŒUR.

Les résultats sont consignés dans la figure ci-dessous. Nous constatons bien que le modèle de prédiction a été construit à partir des données d'apprentissage (170 observations).

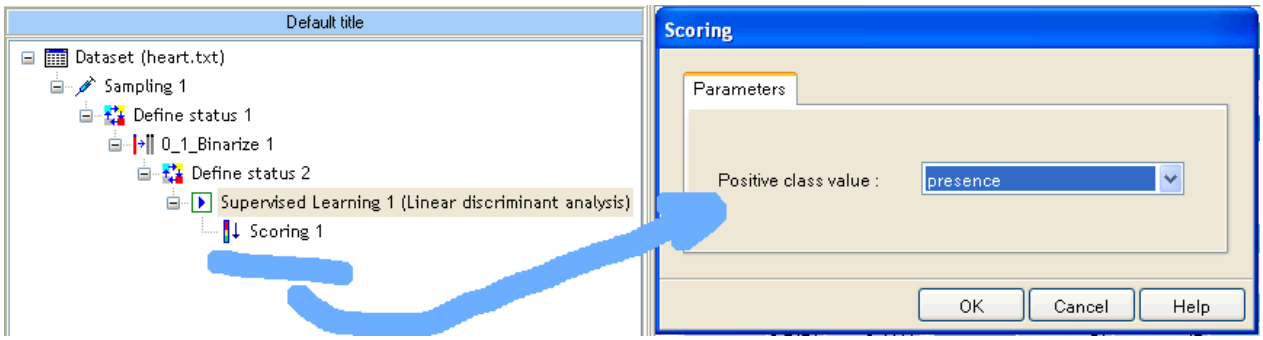


SCORING pour la LDA

L'étape suivante consiste à attribuer un score à tous les individus de l'ensemble de données, qu'ils fassent partie de l'ensemble d'apprentissage ou de l'ensemble test.

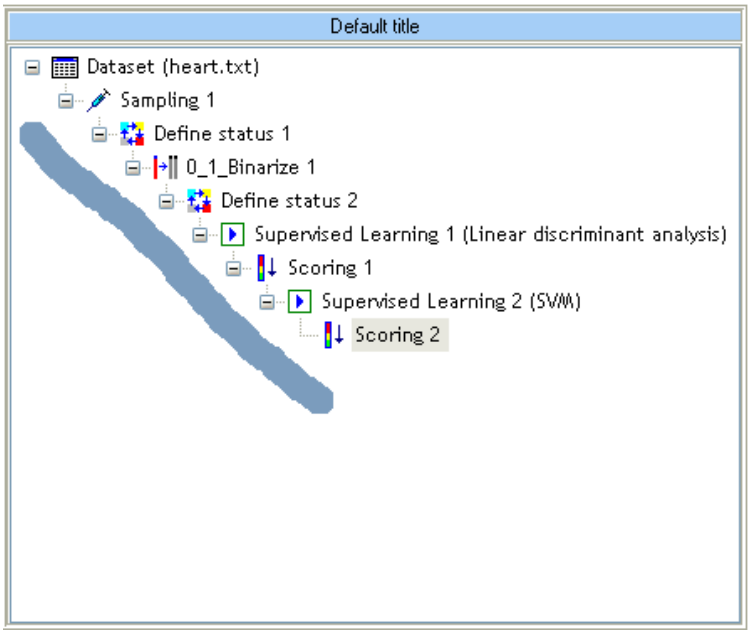
Placer le composant SCORING et spécifier la modalité de la variable à prédire que l'on veut désigner comme la modalité positive.

Nous notons qu'il est donc possible de reproduire ce processus sur les problèmes où la variable à prédire prend plus de 2 modalités. Il suffit de préciser dans le composant SCORING, laquelle nous voulons désigner comme modalité positive.



SCORING pour les SVM

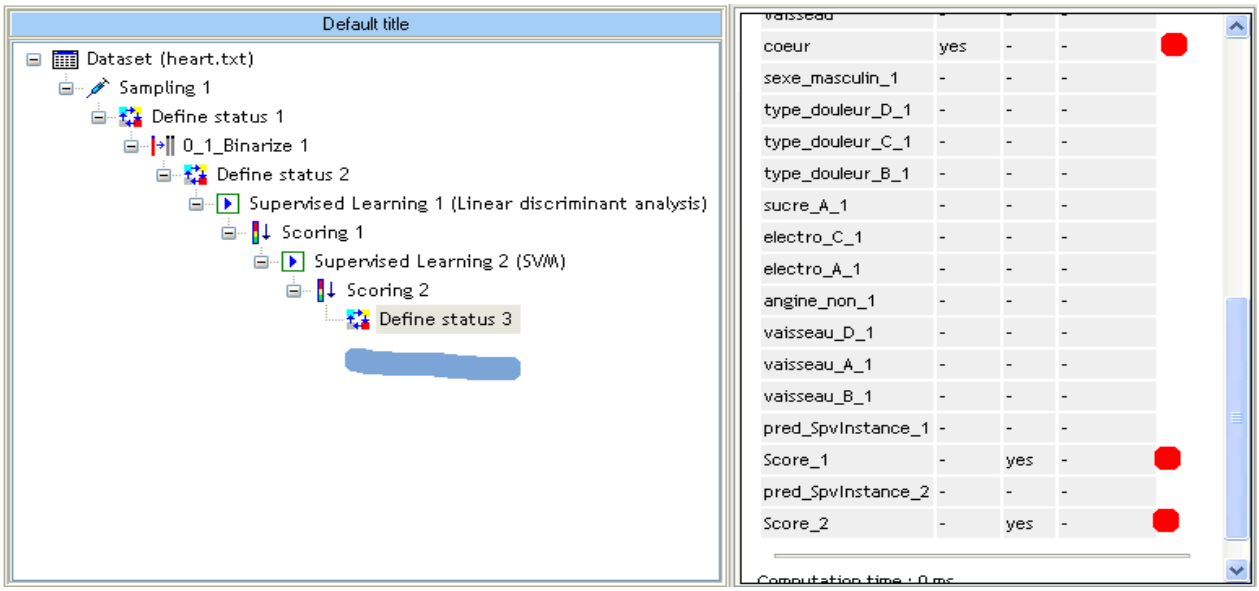
Réitérons le même processus pour les SVM en plaçant les composants adéquats dans le diagramme de traitement.



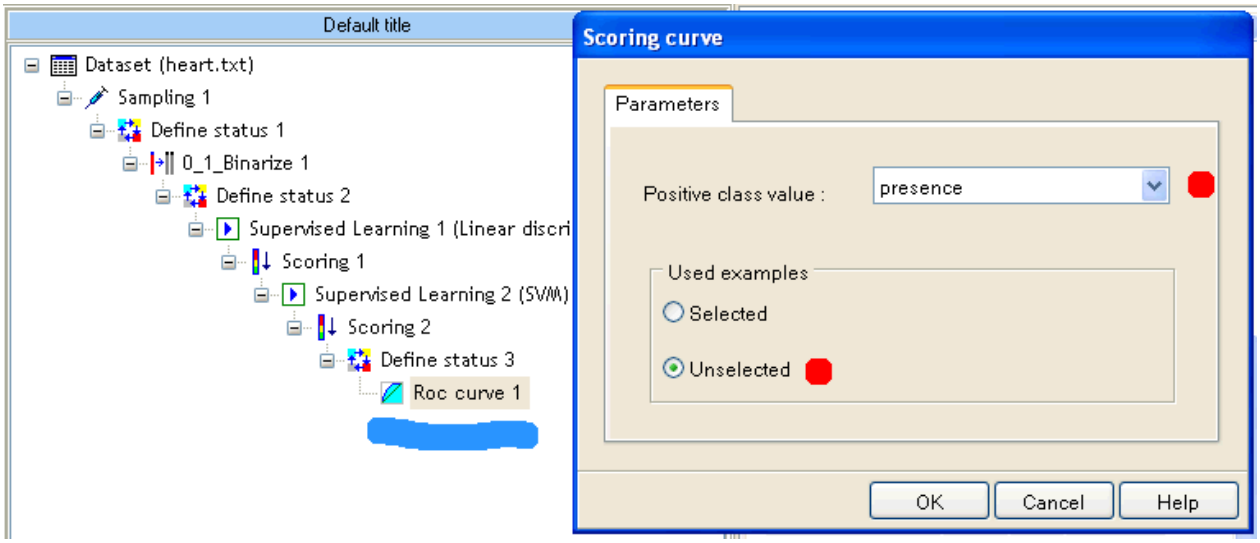
Construire les courbes ROC

Pour construire les courbes ROC, nous devons spécifier l'attribut à prédire et les variables de classement des individus (les scores).

Placez le composant DEFINE STATUS, mettez en TARGET l'attribut CŒUR, et en INPUT les attributs SCORE_1 et SCORE_2. Ce procédé permet d'étendre les comparaisons, il est même possible d'utiliser des scores qui auraient été attribués par des experts de manière ad hoc.



Il reste alors à placer le composant ROC CURVE dans le diagramme de traitements, il y deux paramètres à régler : la modalité de la variable à prédire qui représentera la modalité « positive » ; l'ensemble de données qui servira à construire la courbe.



Les résultats sont regroupés dans un tableau récapitulatif. Nous disposons :

- De l'indicateur AUC calculée de manière très simplifiée à l'aide de la méthode des trapèzes.

- Pour chaque taille de cible (Vrais positifs + Faux positifs), nous disposons des taux de faux positifs et taux de vrais positifs.
- Attention, dans la majorité des cas, les scores ne sont pas comparables d'une méthode à l'autre.

Roc curve 1

Parameters

Positive class value : presence
Used examples : Unselected

Results

ROC Curve

Sample size : 100
Positive examples : 47
Negative examples : 53

Score Attribute	Score_1			Score_2		
AUC	0.9065			0.8884		
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	0.5069	0.0000	0.0000	0.9639	0.0000	0.0000
5	0.5059	0.0000	0.1064	0.9374	0.0000	0.1064
10	0.5054	0.0000	0.2128	0.9127	0.0000	0.2128
15	0.5040	0.0189	0.2979	0.8728	0.0189	0.2979
20	0.5027	0.0189	0.4043	0.7869	0.0189	0.4043
25	0.5023	0.0189	0.5106	0.7357	0.0189	0.5106
30	0.5018	0.0189	0.6170	0.6366	0.0377	0.5957
35	0.5004	0.0566	0.6809	0.5392	0.0566	0.6809
40	0.4997	0.1132	0.7234	0.4996	0.0943	0.7447
45	0.4990	0.1509	0.7872	0.4192	0.1887	0.7447
50	0.4987	0.2075	0.8298	0.3535	0.2453	0.7872
55	0.4984	0.2642	0.8723	0.3200	0.3019	0.8298
60	0.4979	0.3396	0.8936	0.2562	0.3396	0.8936
65	0.4971	0.3962	0.9362	0.2345	0.3962	0.9362
70	0.4966	0.4528	0.9787	0.1861	0.4717	0.9574
75	0.4963	0.5472	0.9787	0.1461	0.5660	0.9574
80	0.4960	0.6415	0.9787	0.1258	0.6415	0.9787
85	0.4954	0.7170	1.0000	0.1038	0.7358	0.9787
90	0.4944	0.8113	1.0000	0.0621	0.8113	1.0000
95	0.4937	0.9057	1.0000	0.0435	0.9057	1.0000
100	0.4930	1.0000	1.0000	0.0270	1.0000	1.0000

Dans ce cas précis, nous constatons que la LDA et les SVM proposent des performances similaires (à cause du tirage aléatoire des individus, il est possible que vous obteniez une grille de valeurs légèrement différente).

Construire le graphique ROC

Les résultats sont présentés sous forme de tableau dans TANAGRA. Il est possible de présenter la courbe ROC sous forme de graphique en exportant la grille ci-dessus dans le tableur de votre choix.

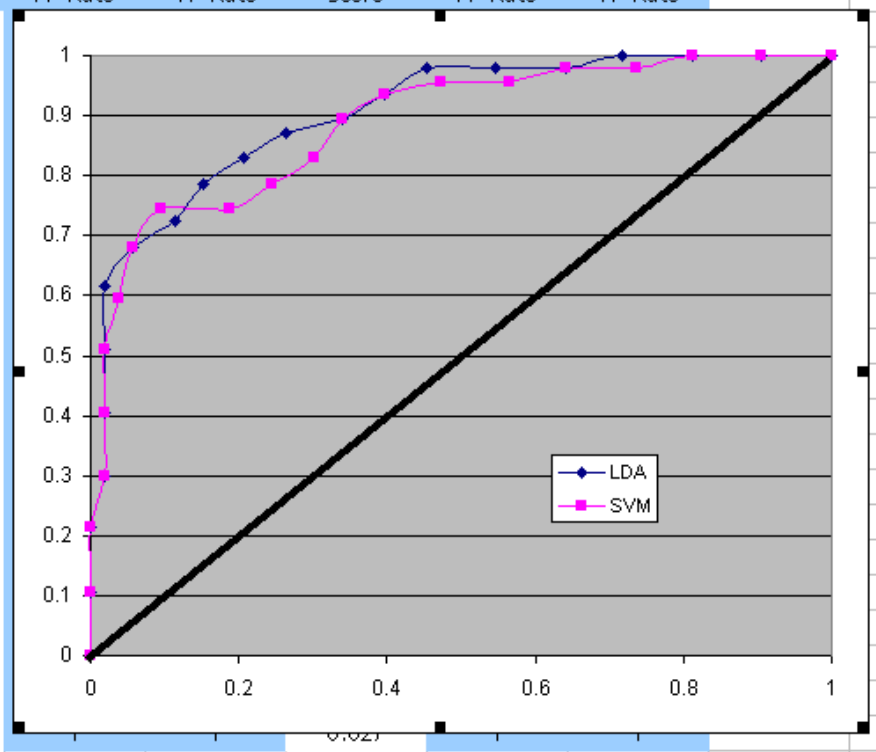
Cliquez sur le menu COMPONENT / COPY RESULTS et copiez les données dans un tableur. L'élaboration du graphique est relativement aisé.

Results

ROC Curve

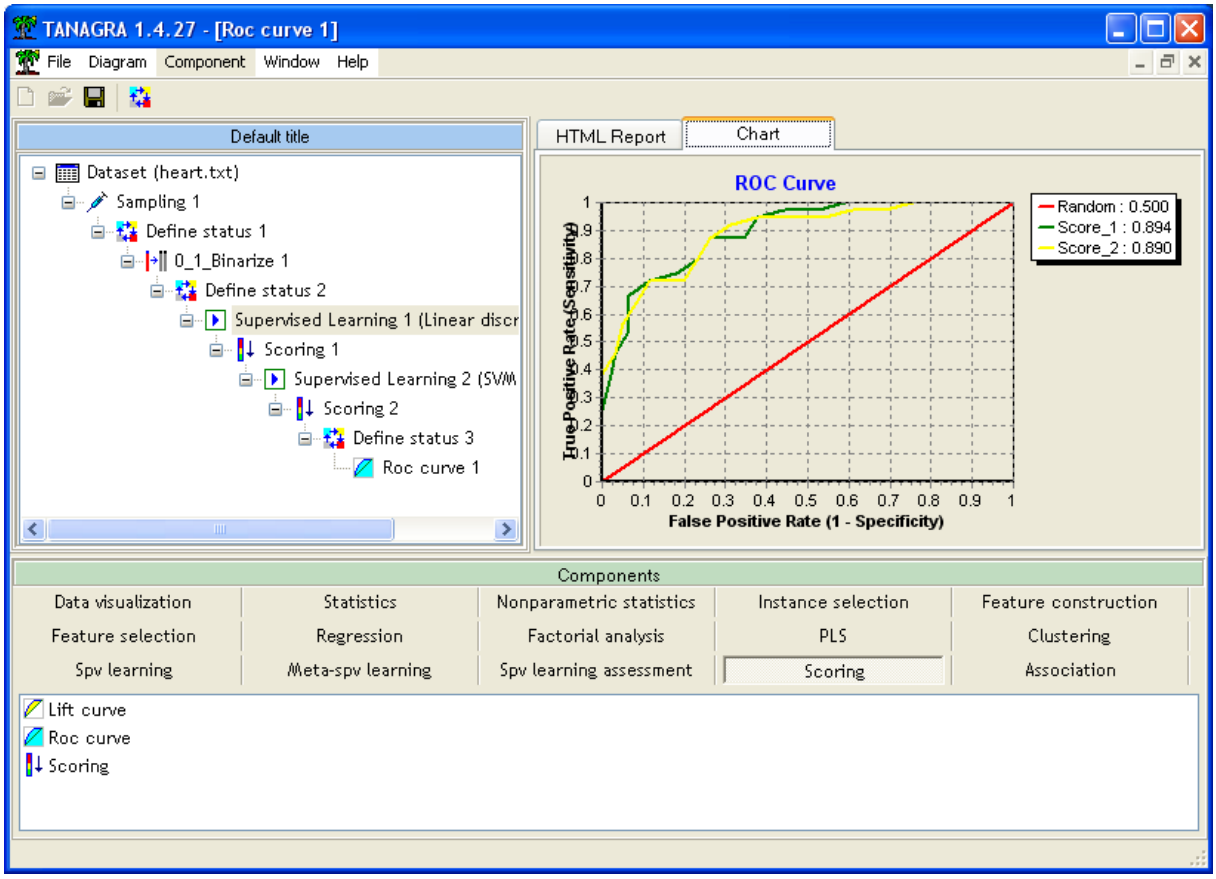
Sample size : 100
Positive examples : 47
Negative examples : 53

Score Attribute	Score_1			Score_2		
AUC	0.9065			0.8884		
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	0.5069					
5	0.5059					
10	0.5054					
15	0.504					
20	0.5027					
25	0.5023					
30	0.5018					
35	0.5004					
40	0.4997					
45	0.499					
50	0.4987					
55	0.4984					
60	0.4979					
65	0.4971					
70	0.4966					
75	0.4963					
80	0.496					
85	0.4954					
90	0.4944					
95	0.4937					
100	0.493					



Computation time : 0 ms.
Created at 29/04/2005 10:49:42

Remarque : Depuis la version 1.4.21 de Tanagra, la courbe ROC est directement fournie. Nous obtenons la représentation suivante[1]



[1] Le tirage aléatoire lors de la partition « apprentissage » - « test » fait que nous obtenons des résultats légèrement différents dans cette copie d'écran plus récente