1 Objectif

Création de rapports avec Tanagra.

Le reporting est un vrai critère de différenciation entre les logiciels de data mining à vocation professionnelle et ceux issus de la recherche. Pour un praticien (ex. chargé d'études), il est important de pouvoir récupérer facilement le fruit de son travail dans un traitement de texte ou dans un diaporama. L'affaire devient particulièrement intéressante lorsqu'il dispose déjà d'une sortie au format tableur. En effet les résultats se présentent souvent sous la forme de divers tableaux et, éventuellement, de graphiques. Le nec plus ultra est de pouvoir définir à l'avance des maquettes de rapports que l'on nourrit simplement à l'issue des calculs et que l'on peut imprimer directement. Pour le chercheur qui développe des outils, tout cela est bien beau, mais ce n'est absolument pas valorisable académiquement. Je me vois très mal pour ma part proposer un article dans une revue montrant que je suis capable d'intégrer automatiquement des camemberts 3D dans un fichier PDF. De fait, les outils élaborés par les chercheurs se contentent souvent de sorties textes, certes complètes, mais peu présentables en l'état dans des rapports destinés à être diffusés à large échelle. Les sorties de R ou de Weka en sont un exemple édifiant.

Tanagra, créé par un enseignant chercheur, s'inscrit dans la même démarche. Rien n'a été initialement prévu pour le reporting. Et pourtant, paradoxalement, il propose dans un des ses menus (DIAGRAM / CREATE REPORT) un outil de création de rapports. C'est la conséquence heureuse d'un choix technologique effectué lors de l'écriture du cahier des charges du logiciel.

Revenons un peu en arrière pour comprendre la démarche. Lorsque j'avais écrit SIPINA (version 3.x), je me suis rendu compte que la construction des fenêtres d'affichage des résultats me prenait énormément de temps, plus que l'écriture des algorithmes de calculs. Dans mon optique, ce n'était pas une bonne chose car cela me détournait de ma principale préoccupation : comprendre les méthodes, les implémenter, les évaluer, en parler. Lorsque j'ai réfléchi aux spécifications de Tanagra, je me suis dit qu'il fallait absolument définir une fenêtre d'affichage standardisée, forcément avec des sorties textes, mais avec néanmoins une présentation relativement attrayante. Et là, j'ai redécouvert le HTML. C'est un peu amusant à dire, surtout en 2003. Le HTML permet de faire un effort minimum de description des sorties, une seule méthode dans la classe de calcul suffit (un peu comme Weka pour ceux qui sont allés voir le code source), tout en obtenant une présentation avenante. De plus, il est possible de mettre en évidence les informations importantes à lire en priorité. Par exemple, rien que pouvoir attribuer des codes couleurs à des tranches de p-value est infiniment précieux.

Par la suite, j'ai réalisé que le choix du HTML allait s'avérer doublement judicieux. En effet, c'est un standard largement répandu. **Sans effort de programmation supplémentaire,** nous pouvons d'une part récupérer les sorties dans le tableur Excel ; d'autre part, nous pouvons exporter les fenêtres de visualisation dans un fichier externe et visualiser les résultats dans un navigateur web, indépendamment du logiciel Tanagra. De fait, leur diffusion est largement facilitée.

Ce sont ces fonctionnalités de « reporting » de Tanagra que nous présentons dans ce didacticiel.

2 Données

Nous utilisons le fichier « heart_disease_male_for_reporting.xls ». C'est un fichier que nous avons beaucoup pratiqué. Il s'agit de diagnostiquer la présence ou non d'une maladie cardiaque (DISEASE) à partir des caractéristiques du patient (âge, etc.).

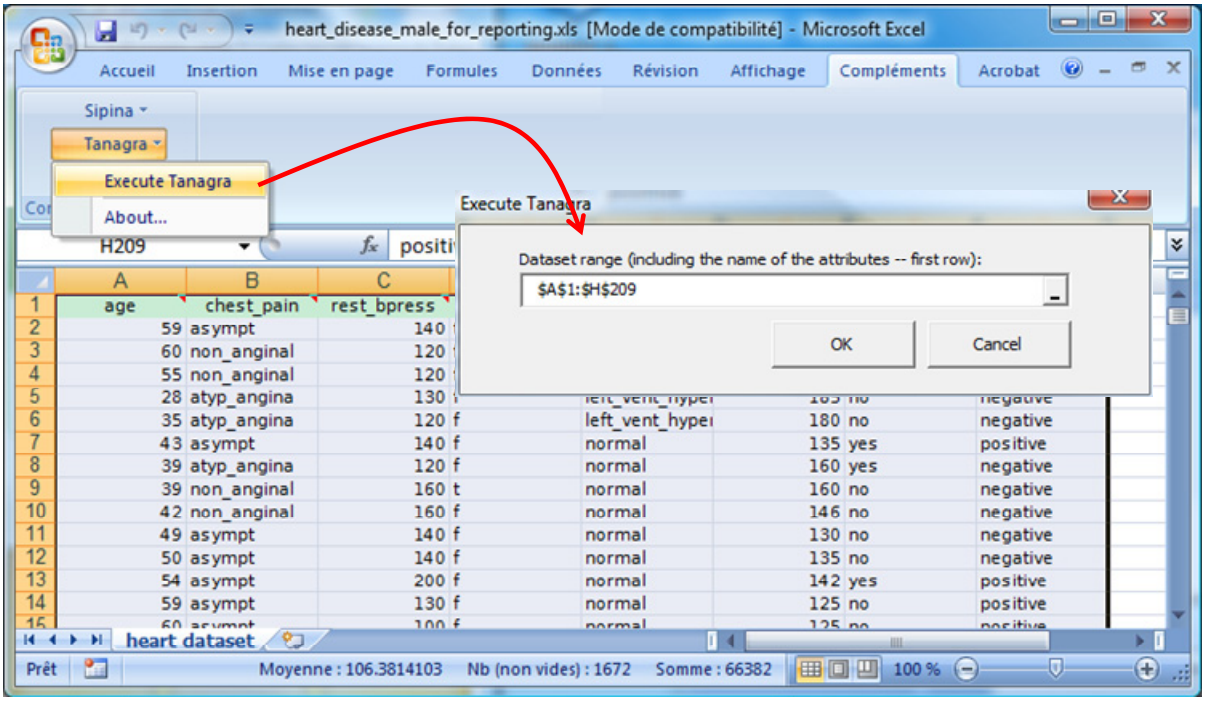
Nous ne nous attarderons pas sur les interprétations. Notre objectif est de montrer les capacités de reporting de Tanagra. Nous réaliserons les tâches suivantes : statistiques descriptives comparatives pour caractériser les deux sous populations (malades vs. non malades) ; subdivision des données en apprentissage et test ; construction et évaluation d'un arbre de décision, construction et évaluation de la régression logistique, précédée d'un recodage des variables ; la régression logistique sera aussi évaluée à l'aide d'une courbe ROC.

Les résultats de l'étude complète seront exportés dans un rapport que l'on peut consulter dans un navigateur web, indépendamment du logiciel Tanagra.

3 Création de rapports avec Tanagra

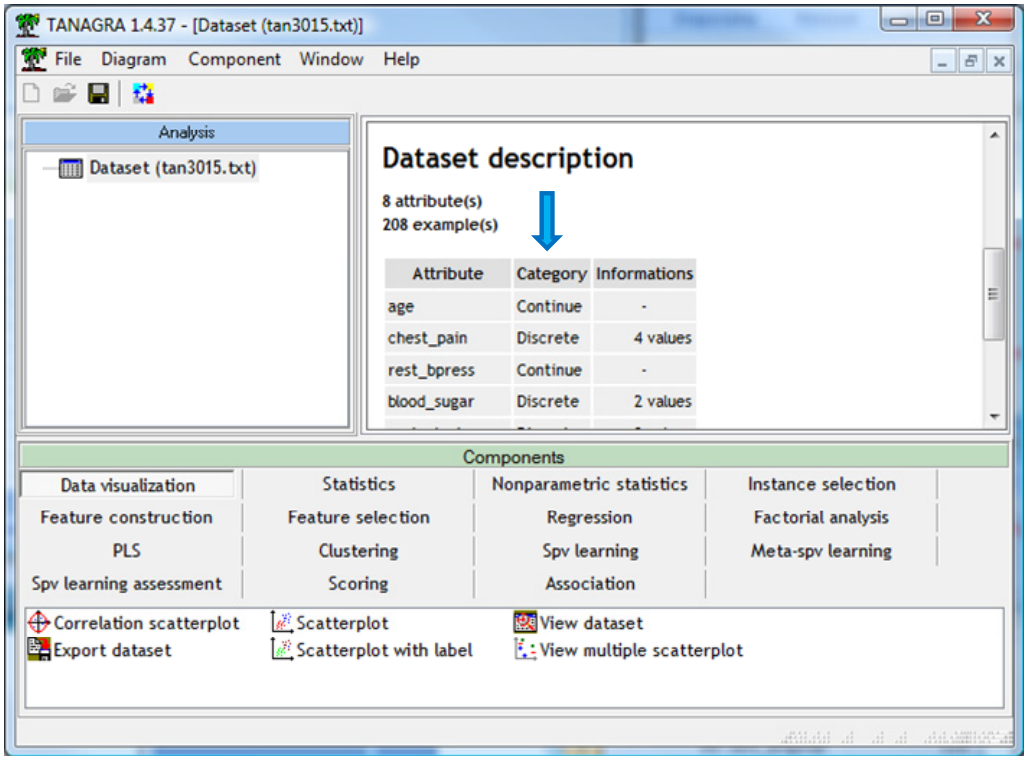
3.1 Chargement des données

Le plus simple est d'ouvrir le fichier dans le tableur Excel. Puis, après avoir sélectionné la plage de données, nous les envoyons vers Tanagra via le menu COMPLEMENTS / TANAGRA / EXECUTE TANAGRA[1].



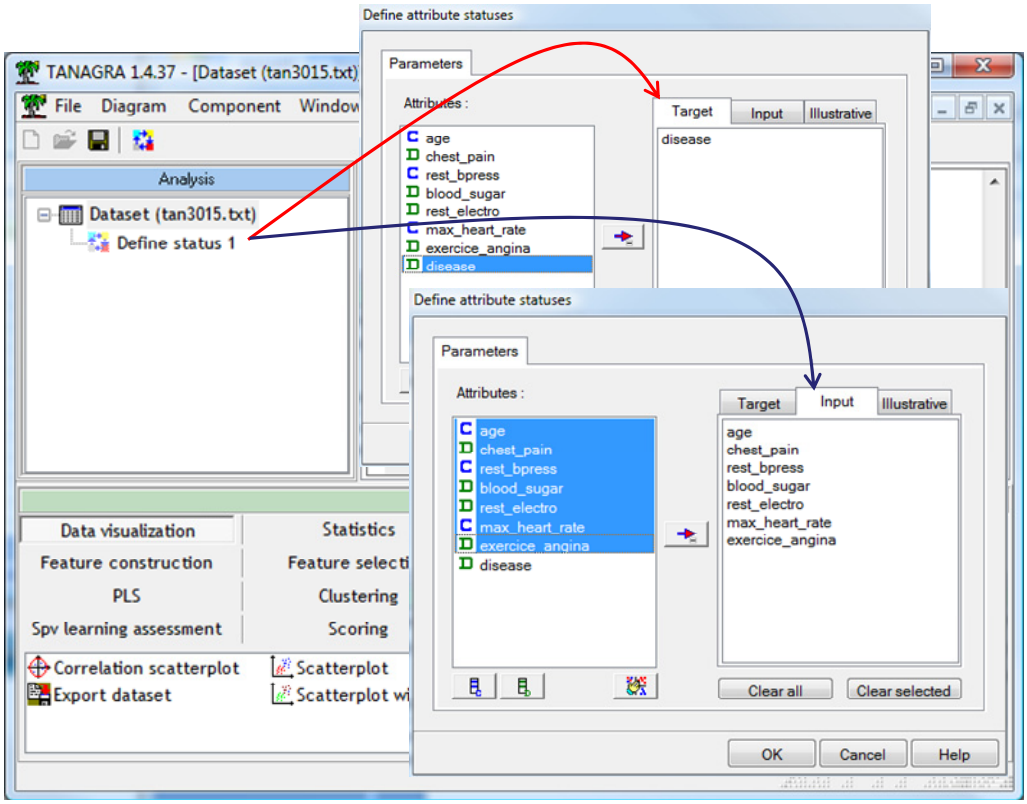
Tanagra est automatiquement démarré. Nous vérifions que 208 individus et 8 colonnes ont bien été chargés. On notera également le typage automatique des variables effectué par Tanagra.

[1] Voir http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour l'installation et l'utilisation de la macro complémentaire dans Excel 2007 et 2010 ; pour Excel 1997 à 2003, voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html

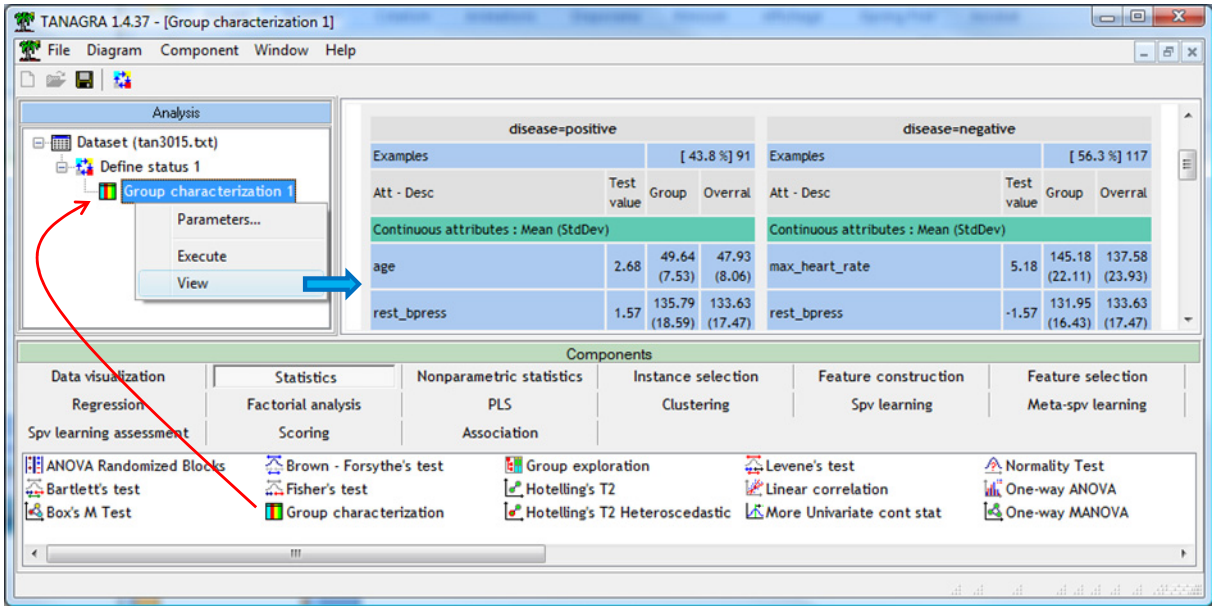


3.2 Description des classes

Dans un problème d'apprentissage supervisé, calculer les statistiques descriptives conditionnellement aux groupes est souvent instructif. Il s'agit d'une description univariée certes, mais elle nous donne déjà une idée de la teneur des résultats que nous obtiendrons par la suite. Nous introduisons le composant DEFINE STATUS (barre d'outils) pour spécifier le rôle des variables. Nous plaçons DISEASE en TARGET, les autres variables en INPUT.



Nous insérons le composant GROUP CHARACTERIZATION (onglet STATISTICS) dans le diagramme.



Voyons le détail des résultats.

Results

Description of "disease"

disease=positive				disease=negative			
Examples			[43.8 %] 91	Examples			[56.3 %] 117
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
age	2.68	49.64 (7.53)	47.93 (8.06)	max_heart_rate	5.18	145.18 (22.11)	137.58 (23.93)
rest_bpress	1.57	135.79 (18.59)	133.63 (17.47)	rest_bpress	-1.57	131.95 (16.43)	133.63 (17.47)
max_heart_rate	-5.18	127.81 (22.72)	137.58 (23.93)	age	-2.68	46.61 (8.24)	47.93 (8.06)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			
chest_pain=asympt	8.47	[73.5 %]	82.4 % 49.0 %	exercise_angina=no	8.35	[77.2 %]	89.7 % 65.4 %
exercice_angina=yes	8.35	[83.3 %]	65.9 % 34.6 %	chest_pain=atyp_angina	6.75	[90.8 %]	50.4 % 31.3 %
blood_sugar=t	2.09	[68.8 %]	12.1 % 7.7 %	chest_pain=non_anginal	3.22	[80.6 %]	24.8 % 17.3 %
rest_electro=st_t_wave_abnormality	1.54	[56.7 %]	18.7 % 14.4 %	blood_sugar=f	2.09	[58.3 %]	95.7 % 92.3 %
chest_pain=typ_angina	0.74	[60.0 %]	3.3 % 2.4 %	rest_electro=left_vent_hyper	1.08	[80.0 %]	3.4 % 2.4 %
rest_electro=normal	-1.00	[42.2 %]	80.2 % 83.2 %	rest_electro=normal	1.00	[57.8 %]	85.5 % 83.2 %
rest_electro=left_vent_hyper	-1.08	[20.0 %]	1.1 % 2.4 %	chest_pain=typ_angina	-0.74	[40.0 %]	1.7 % 2.4 %
blood_sugar=f	-2.09	[41.7 %]	87.9 % 92.3 %	rest_electro=st_t_wave_abnormality	-1.54	[43.3 %]	11.1 % 14.4 %
chest_pain=non_anginal	-3.22	[19.4 %]	7.7 % 17.3 %	blood_sugar=t	-2.09	[31.3 %]	4.3 % 7.7 %
chest_pain=atyp_angina	-6.75	[9.2 %]	6.6 % 31.3 %	exercice_angina=yes	-8.35	[16.7 %]	10.3 % 34.6 %
exercice_angina=no	-8.35	[22.8 %]	34.1 % 65.4 %	chest_pain=asympt	-8.47	[26.5 %]	23.1 % 49.0 %

Chez les individus malades (DISEASE = POSITIVE), l'âge moyen est plus élevé (49.64 vs. 47.93 dans la totalité du fichier[2]. Ils ont un MAX_HEART_RATE plus faible en revanche (127.81 vs. 137.58). Concernant les variables prédictives catégorielles, nous observons une surreprésentation de CHEST_PAIN = ASYMPT (ils sont 82.4% à avoir cette caractéristique dans ce groupe vs. 49% dans la totalité du fichier ; 73.5% des individus chest_pain = asympt se retrouvent dans ce groupe des

[2] Voir http://tutoriels-data-mining.blogspot.com/2008/04/interprter-la-valeur-test.html pour une lecture approfondie du composant GROUP CHARACTERIZATION et la compréhension du critère VALEUR TEST.

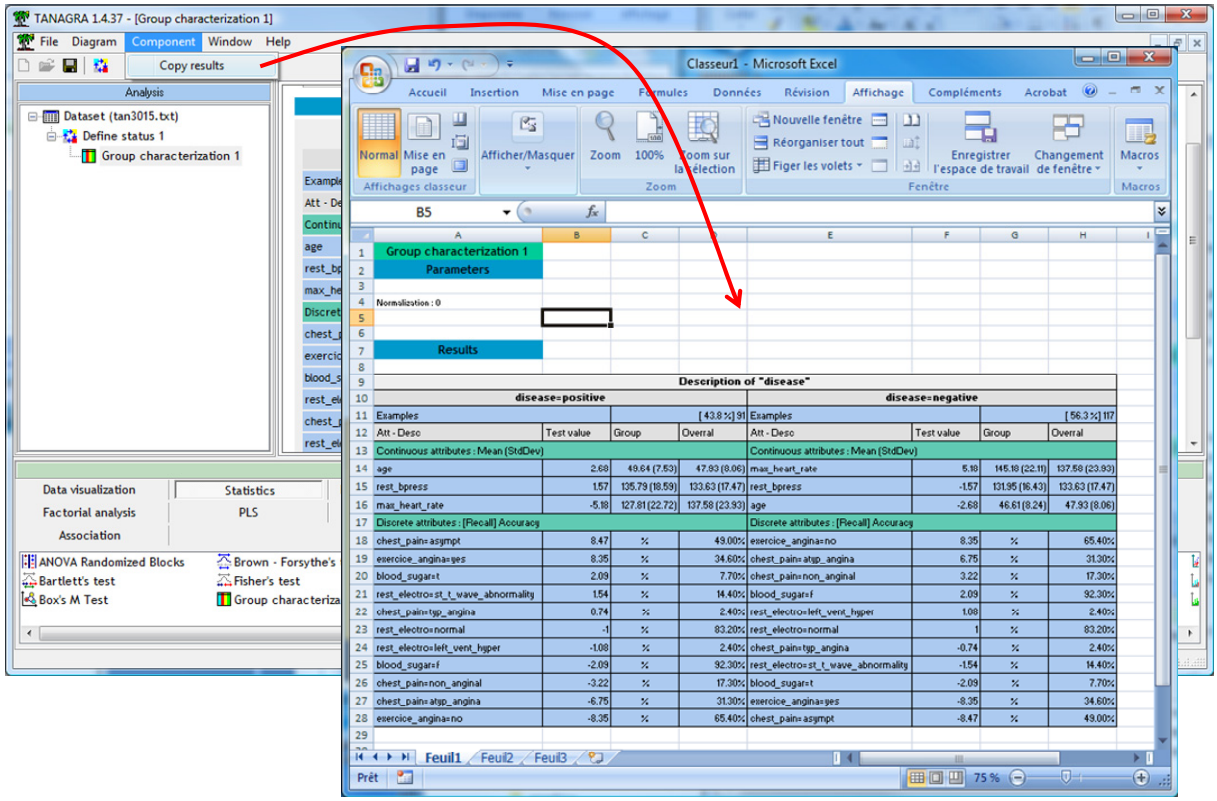
malades) et de EXERCICE_ANGINA = YES. A contrario, nous avons une sous représentation de CHEST_PAIN = ATYP_ANGINA et EXERCICE_ANGINA = NO.

Nous observons exactement les caractéristiques inverses chez les personnes non malades (DISEASE = NEGATIVE). Ce n'est pas étonnant puisque la variable à prédire est binaire.

Il est donc possible d'obtenir une vraie différenciation entre les groupes. Il est à prévoir que les variables mises en avant ici joueront un rôle important dans la modélisation (arbre de décision, régression logistique) que l'on mettra en place.

3.3 Récupération des résultats

Ce tableau des statistiques descriptives conditionnelles, lorsqu'on sait le lire, est très intéressant. On doit pouvoir le récupérer facilement pour l'intégrer dans un rapport. C'est à ce stade que le format standardisé HTML nous est d'une grande aide. En effet il est reconnu par la grande majorité des outils d'édition. Dans Tanagra, nous actionnons le menu COMPONENT / COPY RESULTS. Les résultats sont copiés dans le presse-papiers de Windows. Dans le tableur Excel, nous ajoutons une nouvelle feuille dans notre classeur. Nous collons le tableau en provenance de Tanagra (CTRL + V). Les valeurs sont inscrites dans les bonnes cellules, la structure de tableau est respectée. De toute manière, nous pouvons y apporter toutes les modifications qui nous semblent souhaitables.



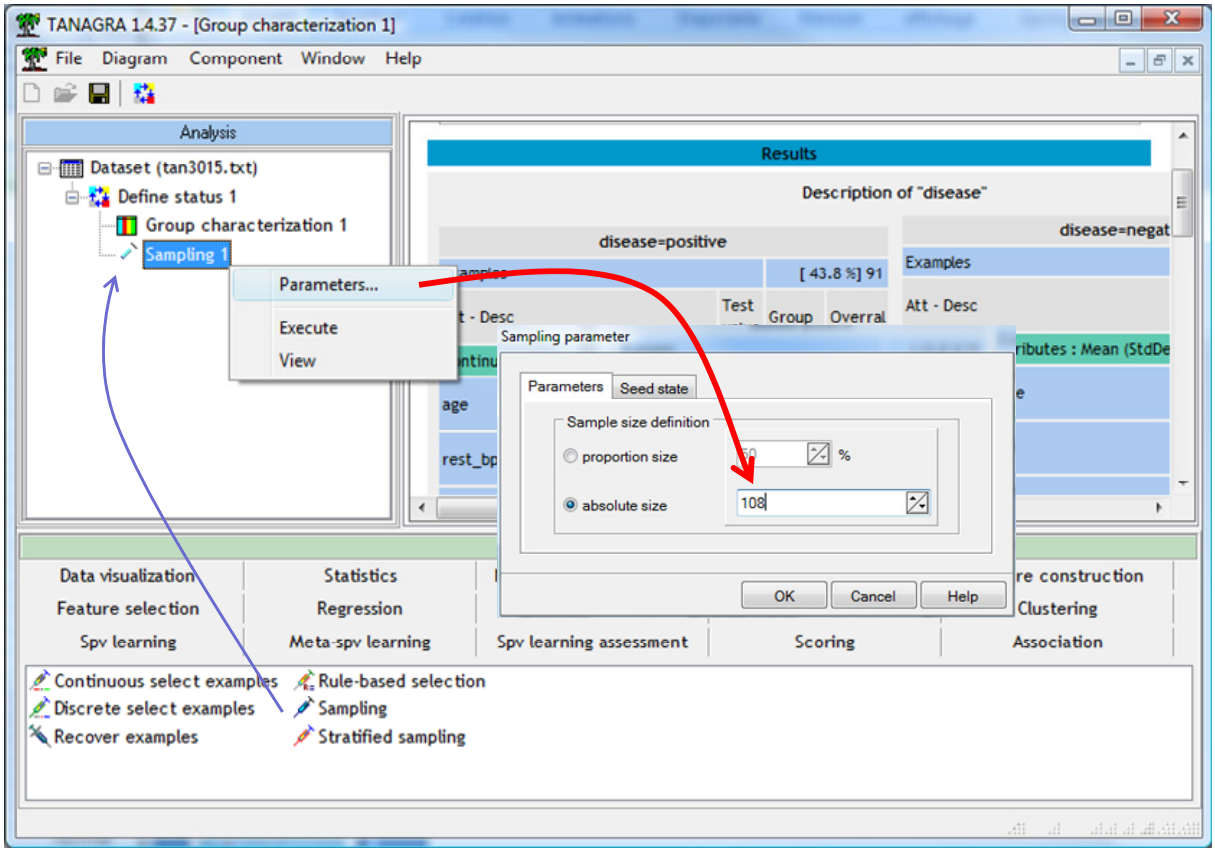
Bien entendu, via Excel, nous pouvons porter les résultats dans n'importe quel outil (traitement de texte, powerpoint, etc.).

3.4 Subdivision des données en échantillons d'apprentissage et de test

Nous souhaitons modéliser la maladie en fonction de la description des patients. Pour obtenir une indication non biaisée des performances en prédiction, il est conseillé de subdiviser les données en

échantillon d'apprentissage, il servira à la construction du modèle, et de test, il servira à son évaluation.

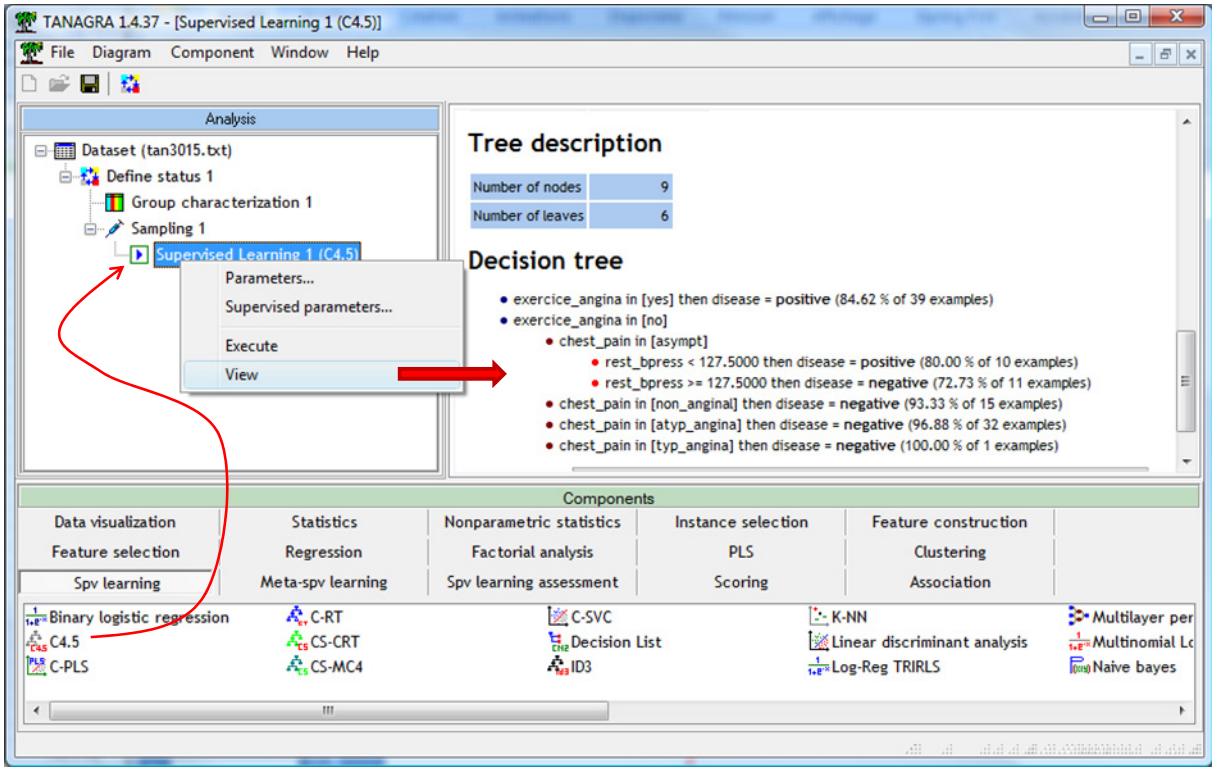
Nous introduisons le composant SAMPLING (onglet INSTANCE SELECTION) dans le diagramme. Nous cliquons sur le menu contextuel PARAMETERS. Nous sélectionnons 108 observations pour l'apprentissage, les 100 restants serviront pour le test.



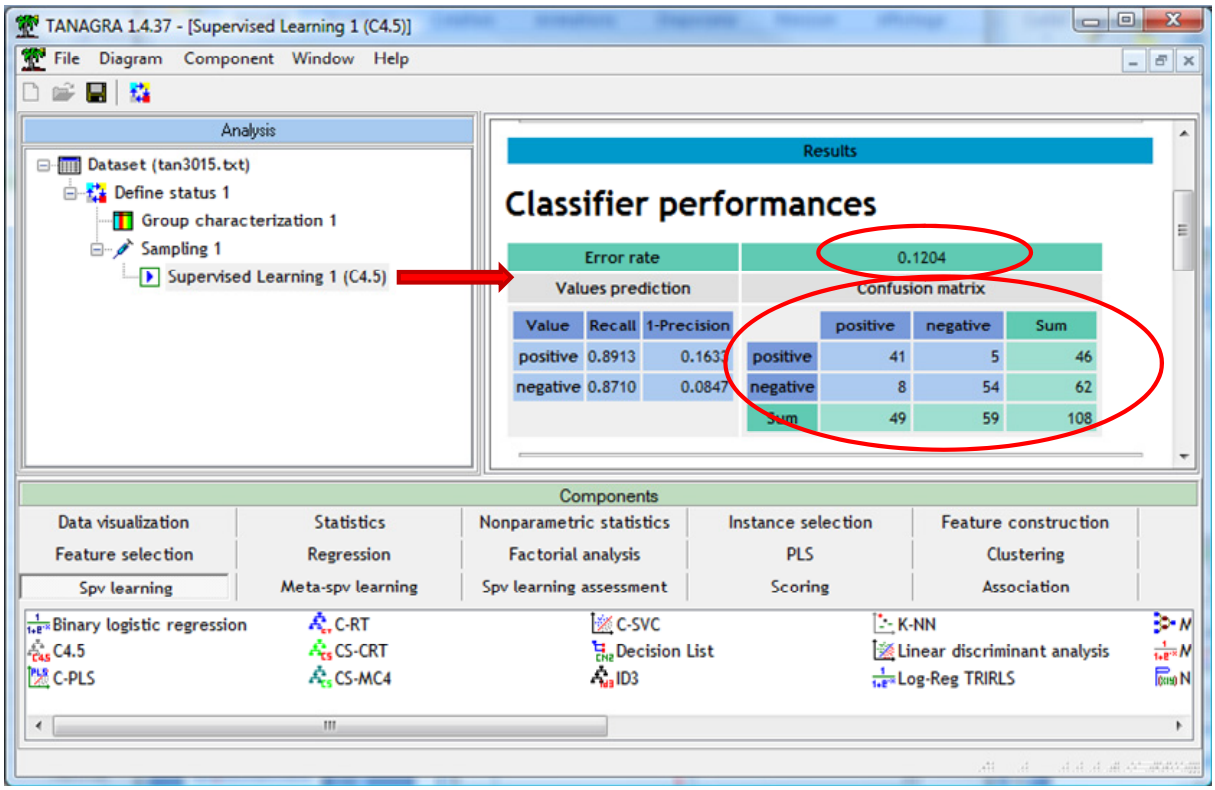
Nous validons, puis nous cliquons sur le menu VIEW.

3.5 Arbre de décision C4.5

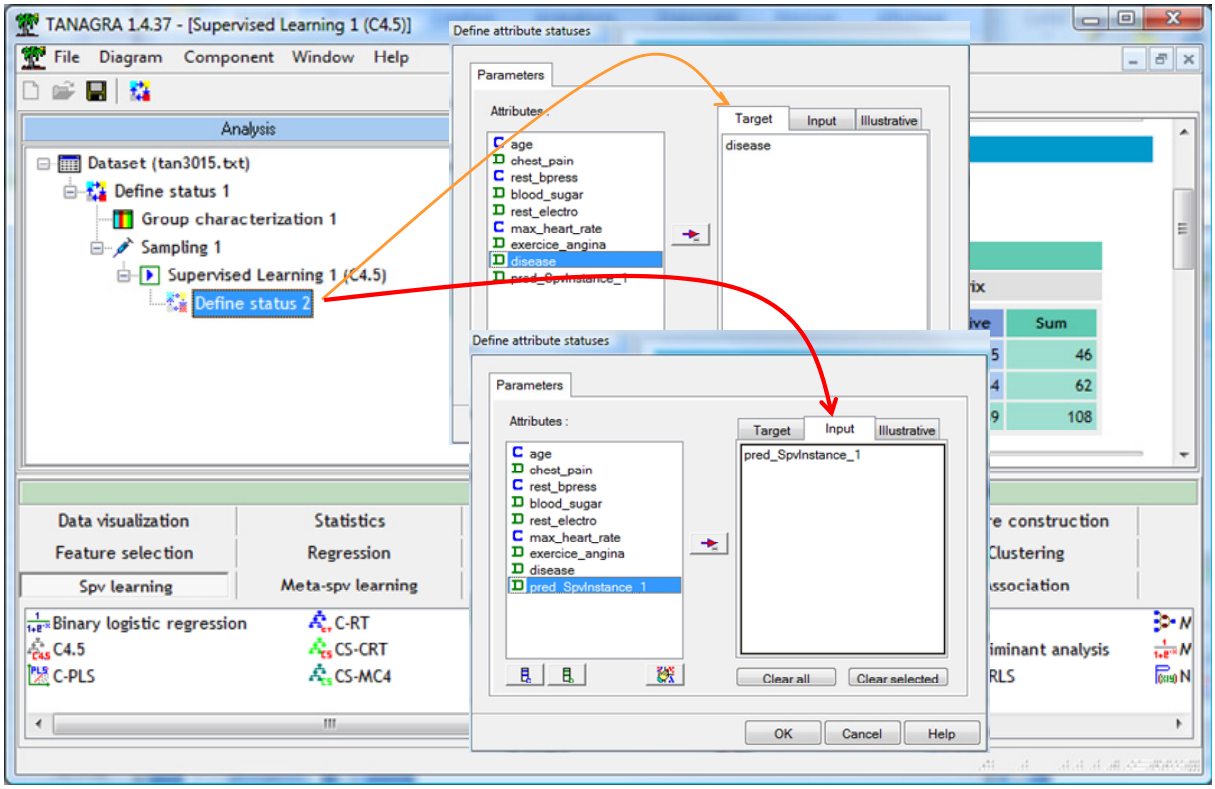
Modélisation. Nous souhaitons implémenter la méthode C4.5 (Quinlan, 1993) (onglet SPV LEARNING). Nous l'insérons à la suite de SAMPLING. Nous actionnons directement le menu VIEW. Nous obtenons l'arbre ci-dessous. Effectivement, les variables CHEST-PAIN et EXEERCICE_ANGINA sont intégrés dans le modèle prédictif. Une troisième variable, REST_BPRESS, qui ne se distingue pas de manière univariée, fait son entrée dans l'arbre. Pouvoir déceler les interactions est le principal avantage des techniques multivariées.



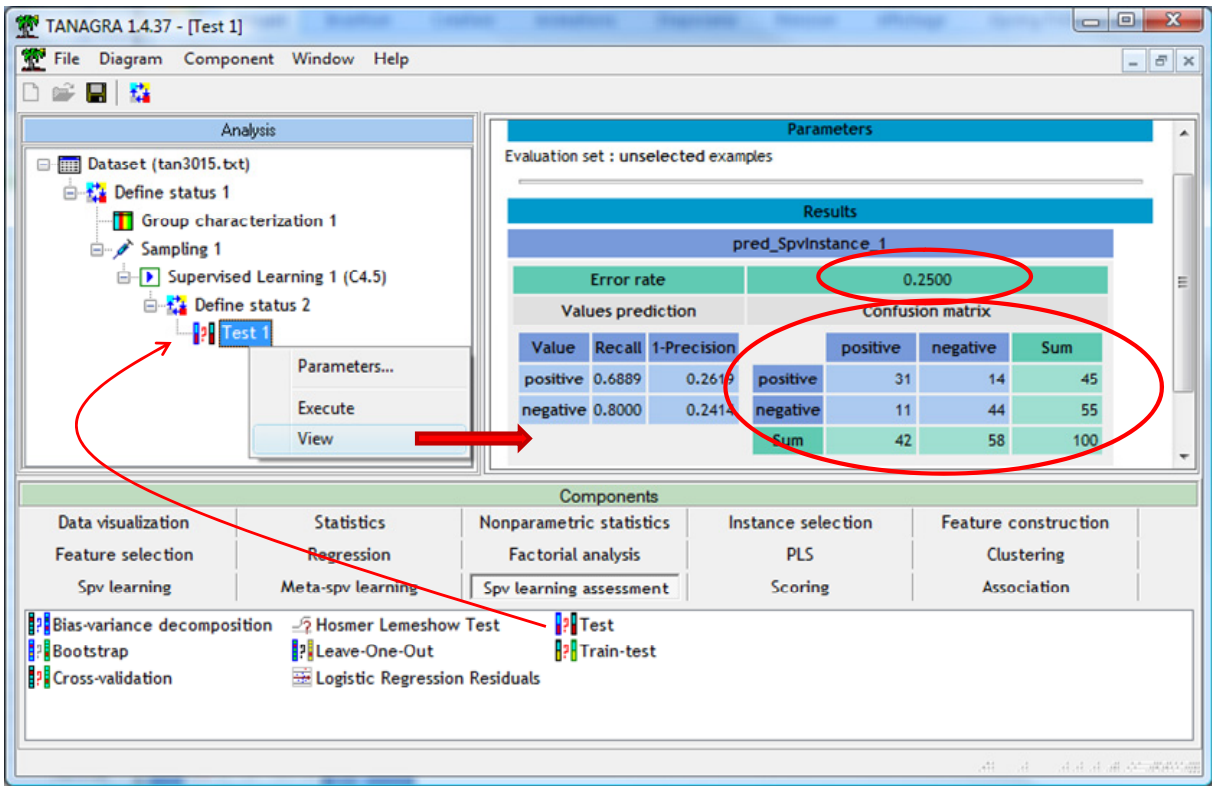
Evaluation. Il indique un taux d'erreur en resubstitution de 12.04%. Nous savons que ce chiffre ne reflète pas les performances en généralisation de l'arbre, surtout s'agissant de la méthode C4.5 réputée pour sa propension au sur apprentissage.



Nous insérons de nouveau le composant DEFINE STATUS dans le diagramme. Nous plaçons en TARGET la variable cible observée DISEASE, en INPUT, la variable prédite par l'arbre de décision PRED_SPV_INSTANCE_1.



Enfin, nous ajoutons le composant TEST (onglet SPV LEARNING ASSESSMENT). Nous cliquons sur VIEW. Par défaut, il est paramétré pour calculer la matrice de confusion sur les données inactives c.-à-d. l'échantillon test.

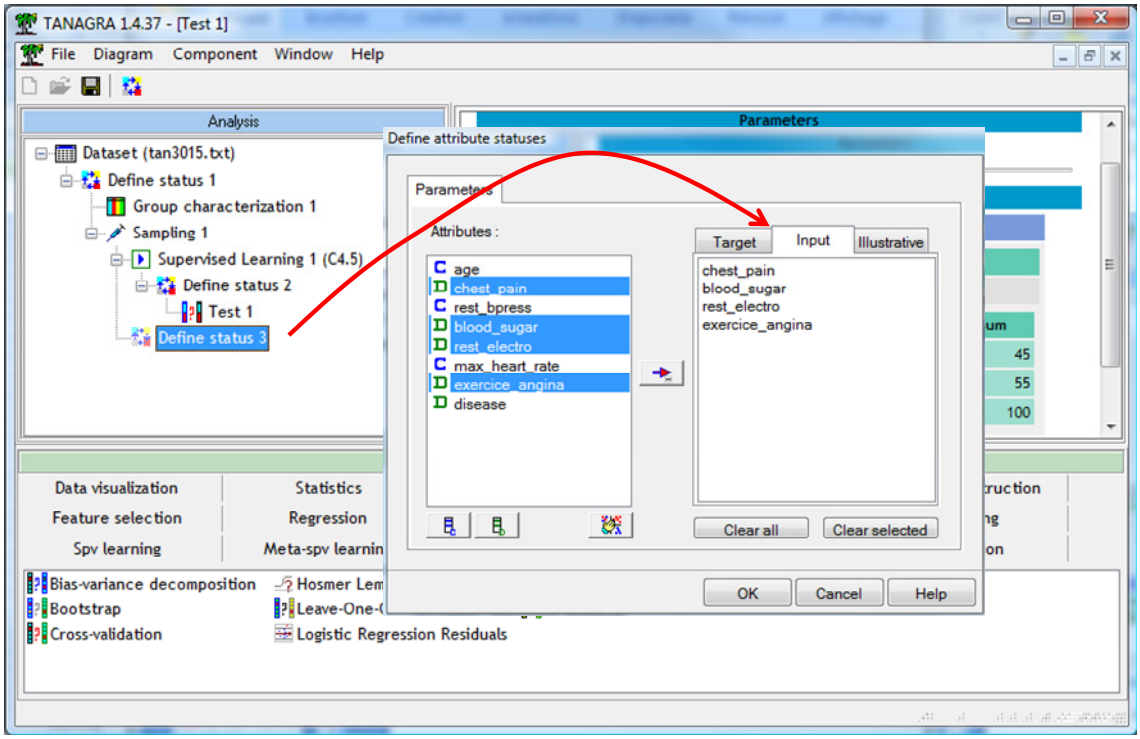


Le « vrai » taux d'erreur en généralisation de l'arbre est de 25%, plus du double de l'erreur en resubstitution. Comme quoi il fallait réellement se méfier.

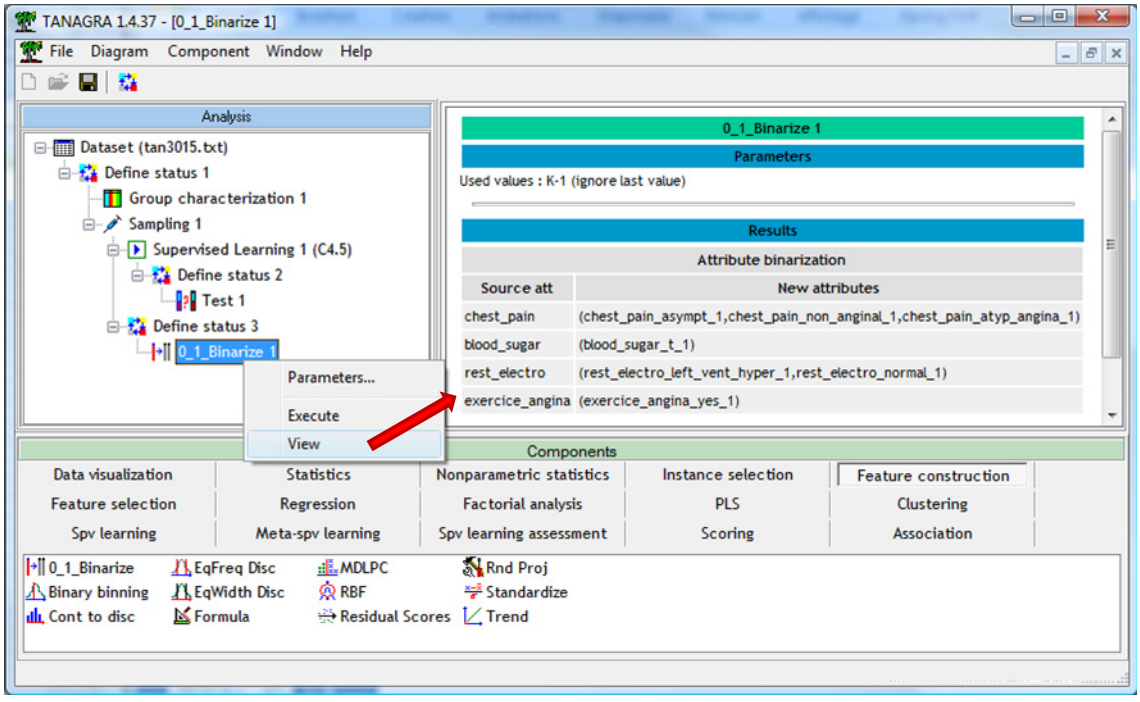
3.6 Régression logistique

Modélisation. Sachant à quoi s'en tenir avec C4.5, nous voulons évaluer le comportement de la régression logistique sur nos données. Elle ne peut pas être mise en œuvre directement car il y a des variables catégorielles parmi les prédictives. Nous allons les recoder en 0/1.

Nous insérons le composant DEFINE STATUS pour désigner les variables à recoder. Nous plaçons en INPUT : CHEST_PAIN, BLOOD_SUGAR, REST_ELECTRO, EXERCICE_ANGINA.

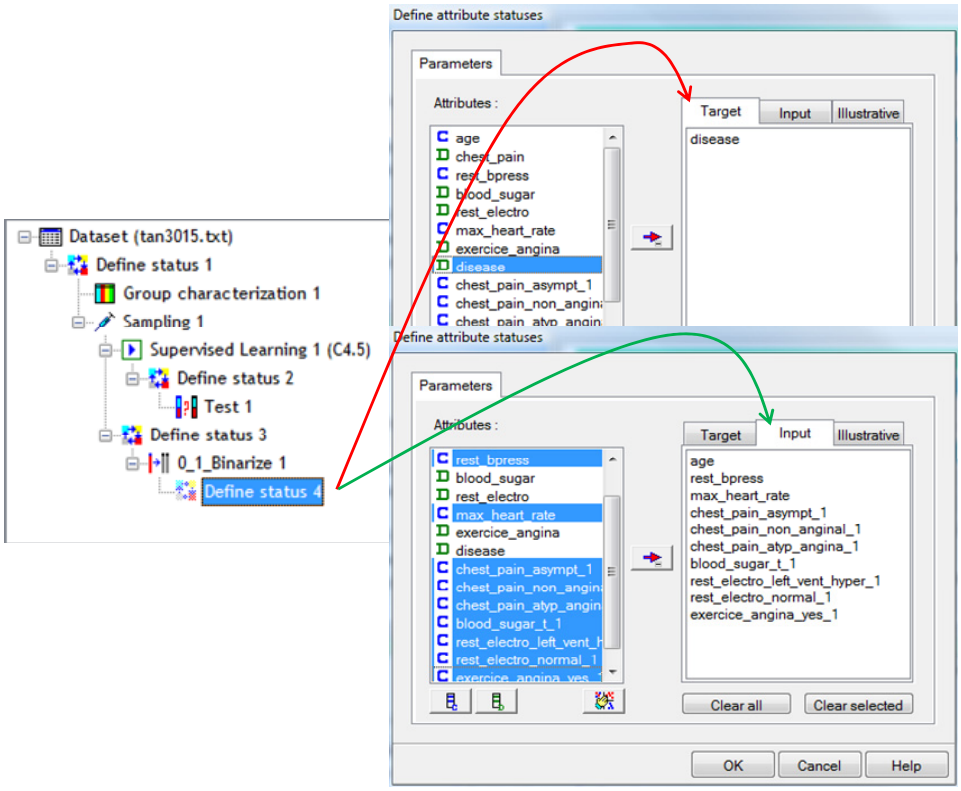


Nous ajoutons à la suite le composant 0_1_BINARIZE (onglet FEATURE CONSTRUCTION). Nous cliquons directement sur VIEW.

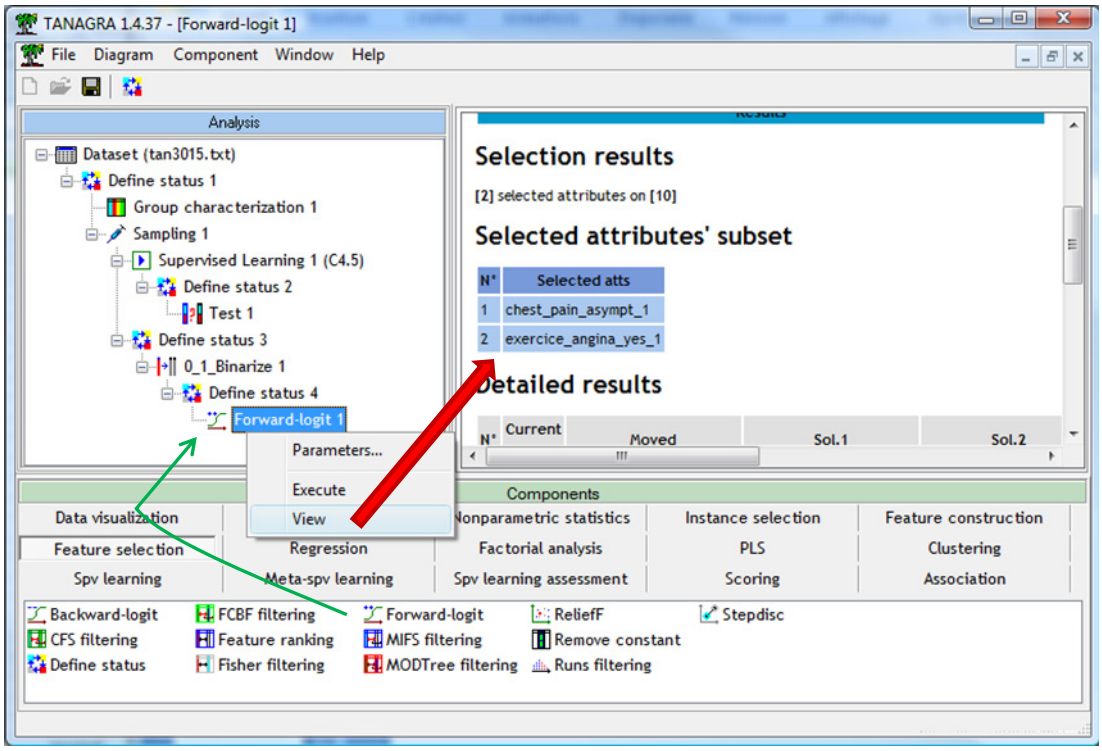


Pour une variable à K modalités, Tanagra crée (K-1) indicatrices. La dernière modalité sert de référence. Elle n'est pas indiquée dans la liste des variables générées.

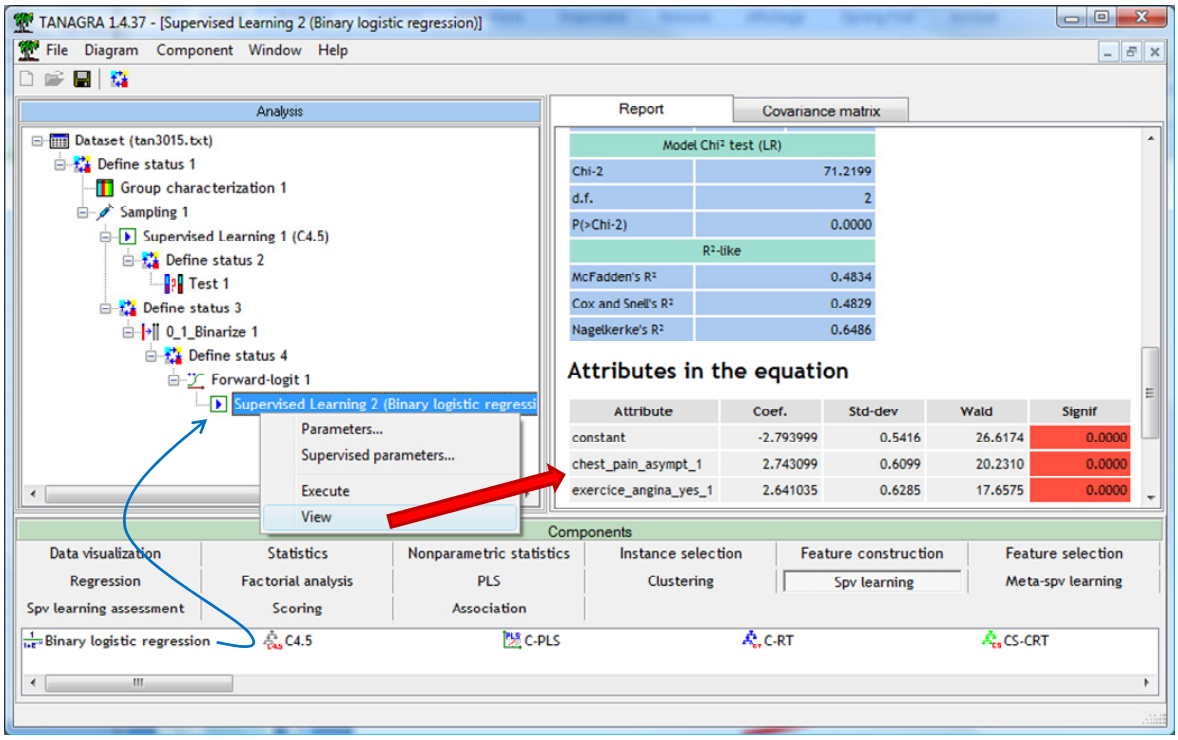
Nous pouvons initier l'apprentissage. Un DEFINE STATUS sert à indiquer la variable cible (TARGET = DISEASE) et les prédictives (INPUT = toutes les variables numériques).



Nous plaçons le composant FORWARD LOGIT (onglet FEATURE SELECTION) pour la sélection des variables pertinentes.

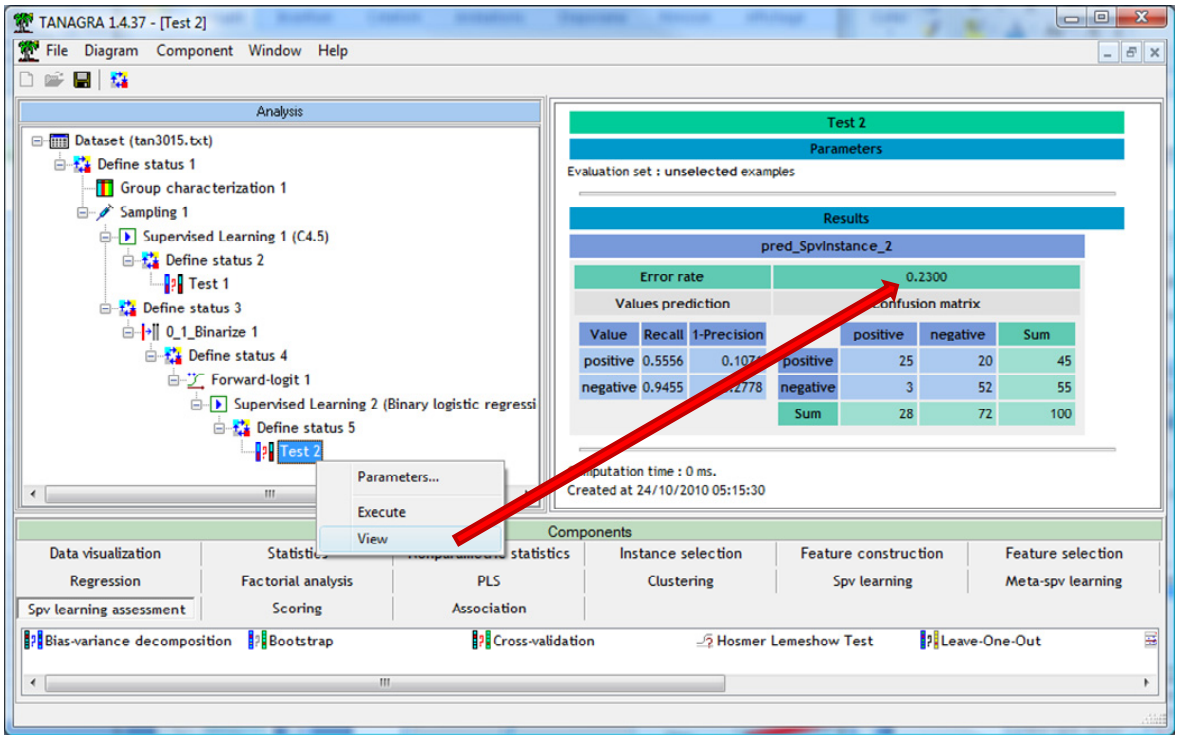


Deux indicatrices sont sélectionnées : CHEST_PAIN = ASYMPT et EXERCICE_ANGINA = YES (*tiens donc ! revoyons les statistiques descriptives conditionnelles*, section 3.2). Il ne reste plus qu'à lancer la régression logistique (BINARY LOGISTIC REGRESSION, onglet SPV LEARNING).



La méthode indique un taux d'erreur en resubstitution de 18.52%. Damned, serait-elle moins performante que l'arbre C4.5 ?

Evaluation. De nouveaux, nous appliquons le classifieur sur l'échantillon test. Nous plaçons tour à tour les composants DEFINE STATUS (target = disease, input = pred_spvinstance_2) et TEST.

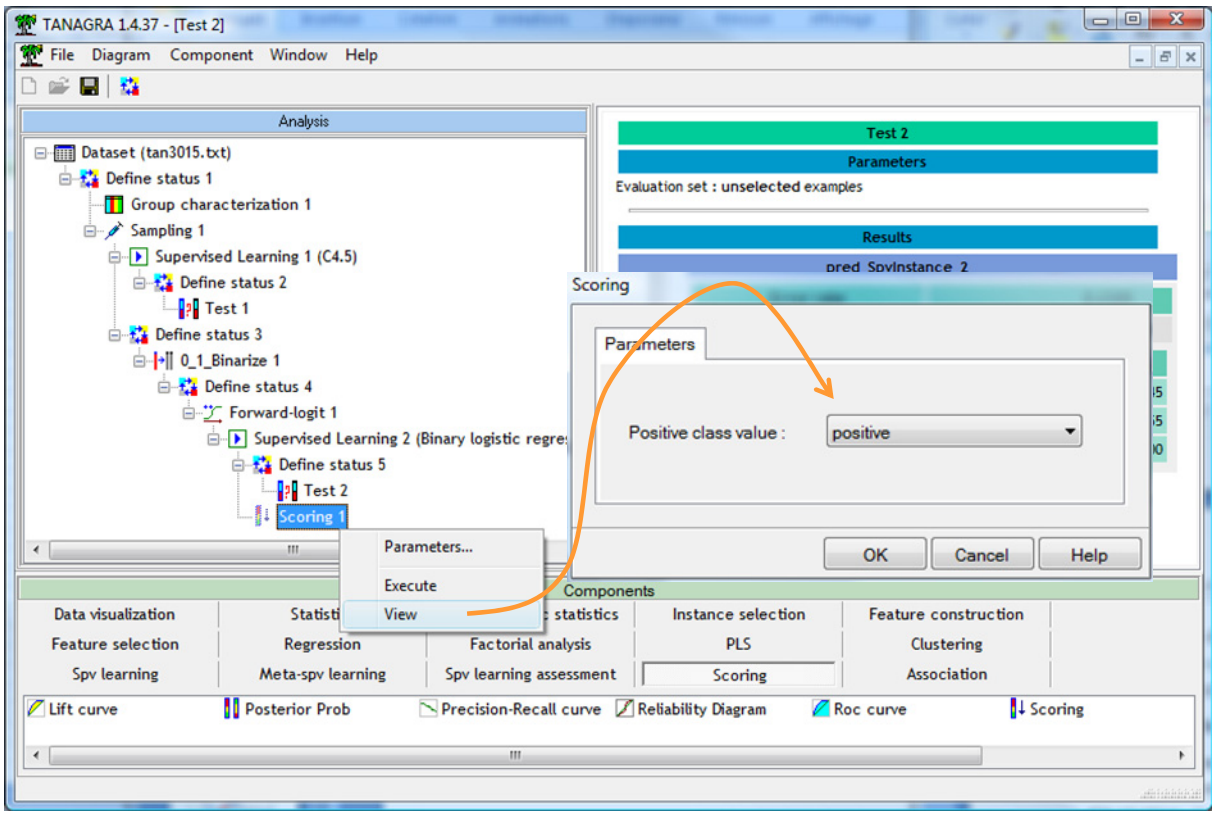


Le « vrai » taux d'erreur du modèle est de 23%, légèrement meilleure que C4.5 globalement, mais surtout avec un comportement très différent : la sensibilité est moindre (25/45 = 56% vs. 69%), au profit d'une précision plus intéressante (25/28 = 89% vs. 74%).

3.7 Elaboration de la courbe ROC

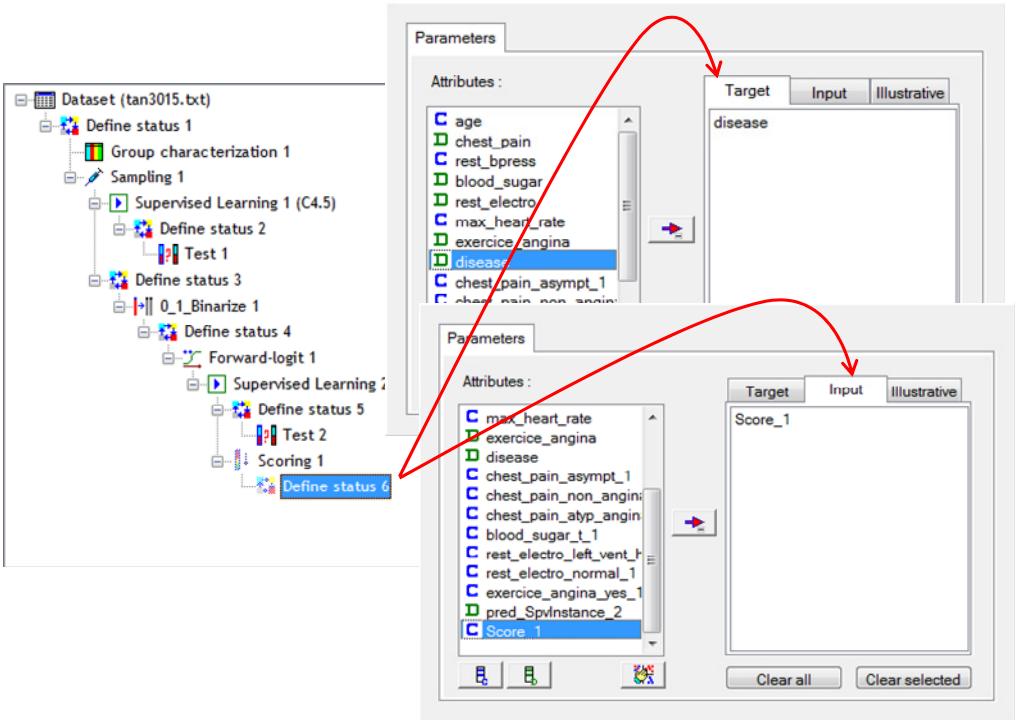
Contrairement aux arbres, la régression logistique fournit une bonne approximation de la probabilité d'être positif des observations. Nous allons mettre à profit cette qualité pour utiliser un autre outil d'évaluation : la courbe ROC. Elle présente l'avantage d'être plus complète, ne dépendant ni d'un système de coûts (implicitement unitaire et symétrique dans le taux d'erreur), ni de la prévalence des positifs dans l'échantillon de données (le taux d'erreur considère que le fichier utilisé est représentatif).

Dans un premier temps, nous calculons le score de chaque observation avec le composant SCORING (onglet SCORING). Nous le paramétrons de la manière suivante (DISEASE = POSITIVE sont les « positifs »).

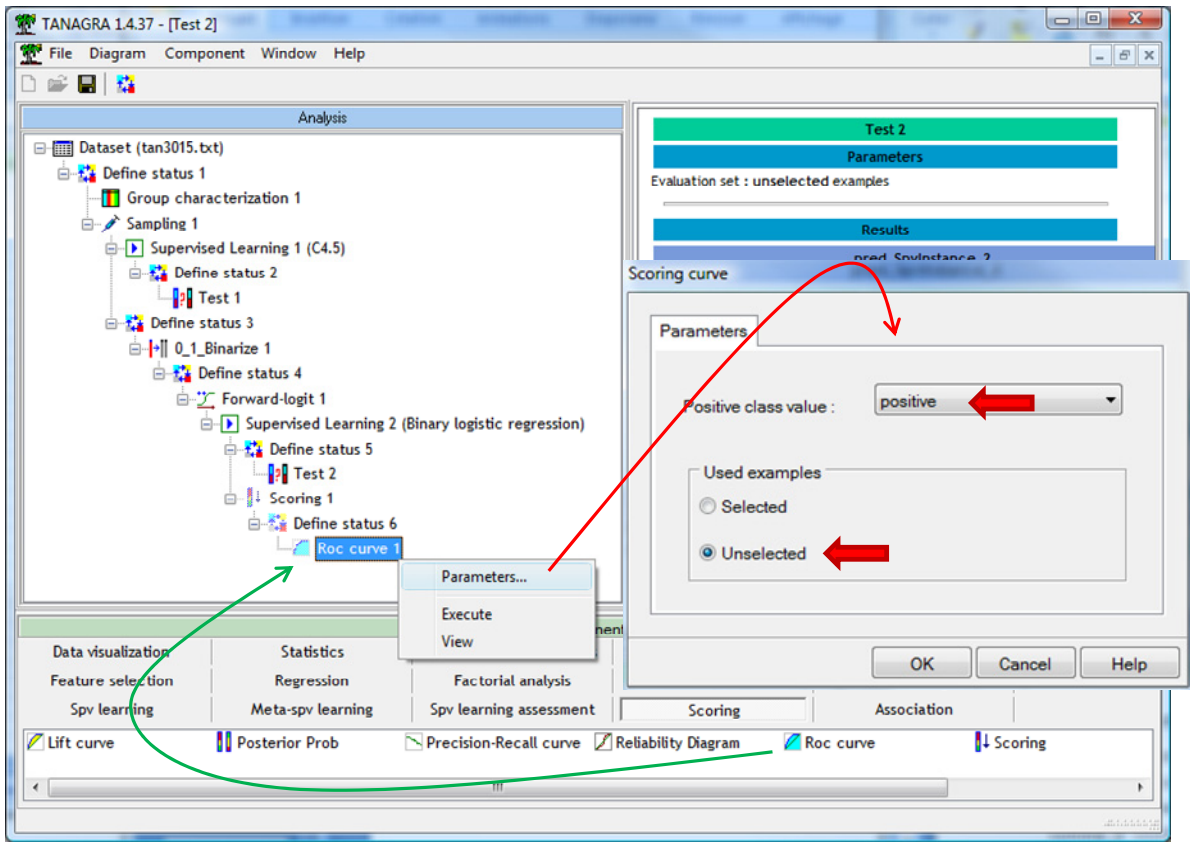


Nous validons et nous cliquons sur VIEW. Une nouvelle colonne est ajoutée aux données.

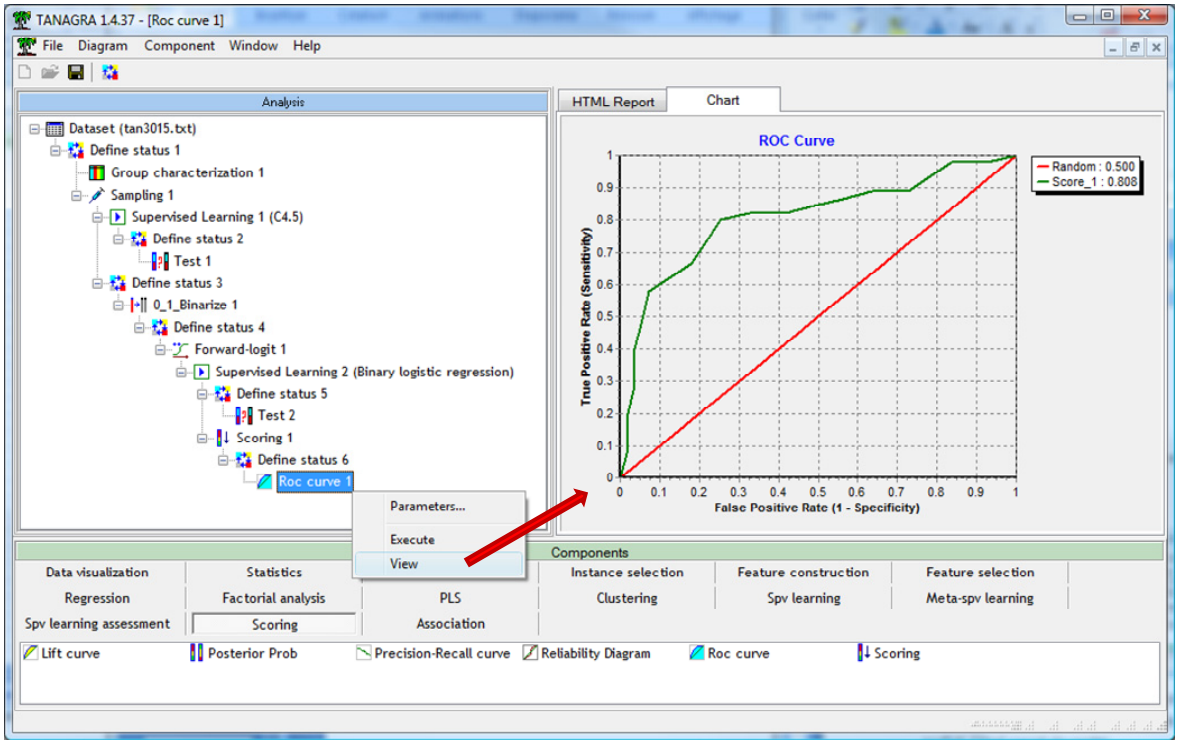
Nous ajoutons DEFINE STATUS dans le diagramme, nous plaçons en TARGET la variable cible DISEASE, en INPUT la colonne score SCORE_1 (nous pouvons en placer plusieurs si nous souhaitons comparer plusieurs scores, y compris ceux obtenus via d'autres approches que les techniques d'apprentissage supervisé).



Nous insérons alors le composant ROC CURVE (onglet SCORING). Nous le paramétrons de manière à cibler les positifs (DISEASE = POSITIVE) et à calculer la courbe sur les individus de l'échantillon test (UNSELECTED).

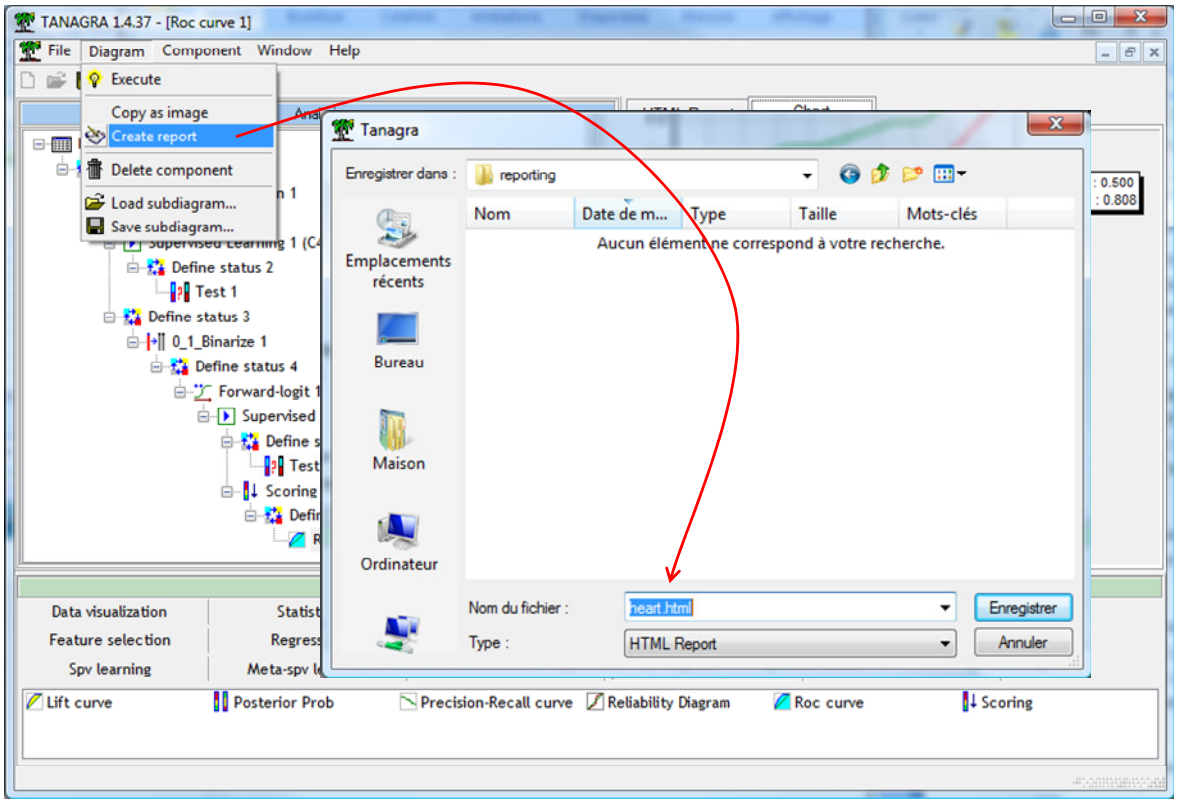


Nous cliquons sur VIEW. Nous obtenons la courbe ROC. L'aire sur la courbe est AUC = 80.8%. Le modèle est plutôt de bonne qualité (si le score était mauvais, nous aurions une AUC = 50% ; s'il était parfait, l'AUC serait de 100%).



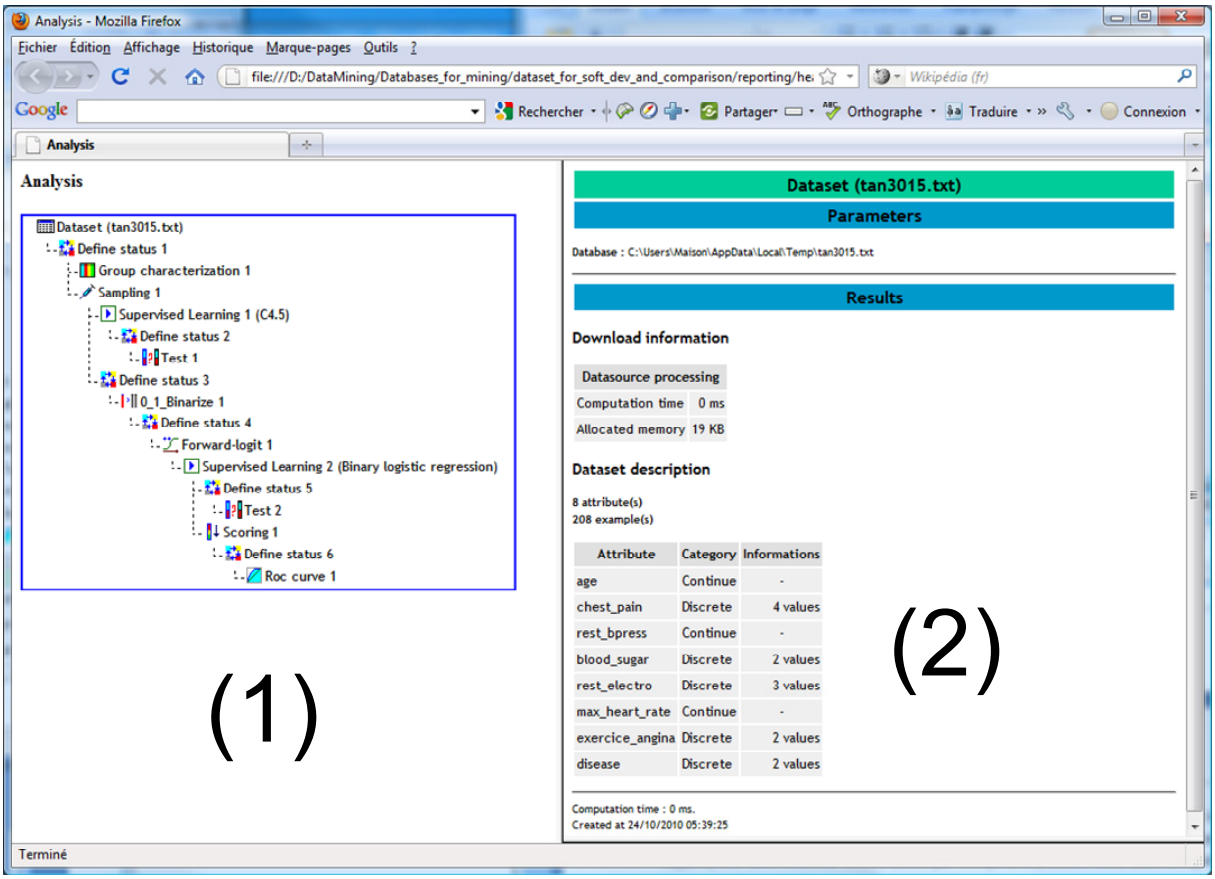
3.8 Construction et visualisation du rapport

Voilà une analyse complète, faite d'exploration, de modélisation et d'évaluation. Nous aimerions partager les résultats avec d'autres personnes. Tanagra sait produire un rapport au format HTML constitué des pages affichées dans chaque fenêtre de visualisation. Lorsqu'il s'agit d'un graphique comme la courbe ROC, il génère une copie d'écran au format JPG intégrée automatiquement dans la page associée à la méthode.



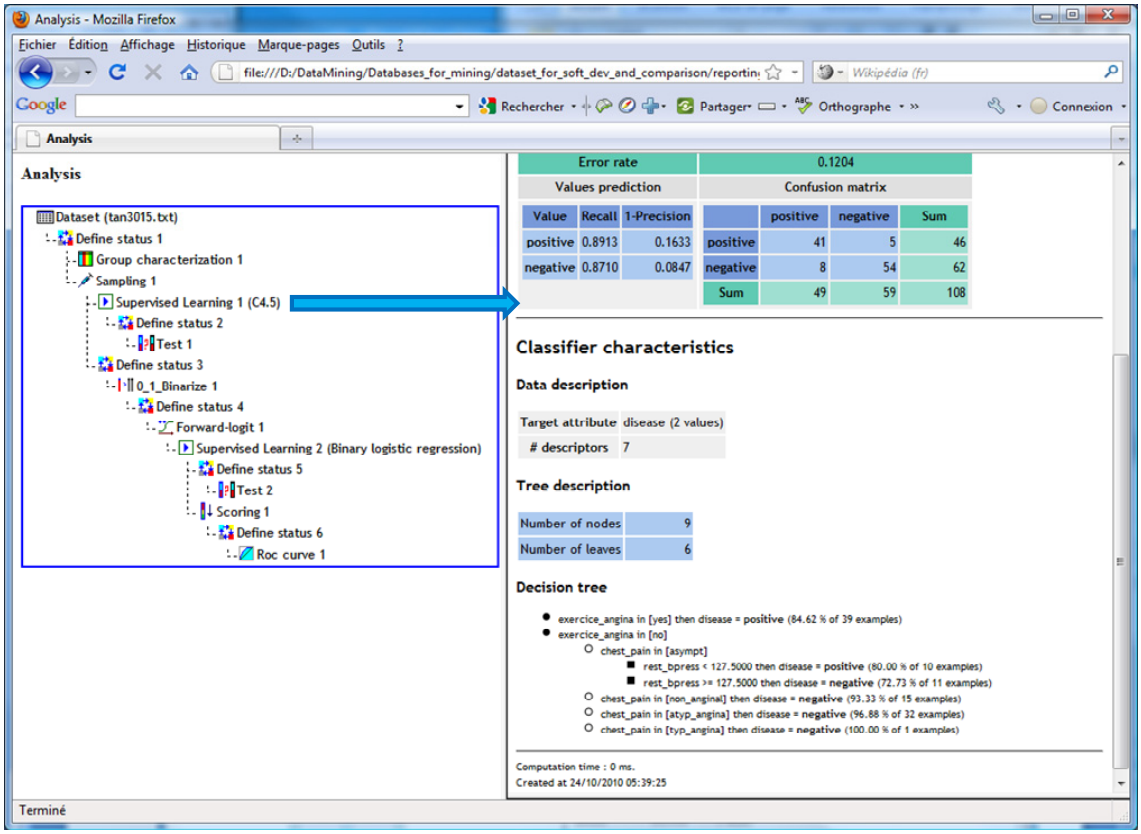
Pour élaborer le rapport, nous actionnons le menu DIAGRAM / CREATE REPORT. Une boîte de dialogue nous demandant de spécifier le nom de fichier apparaît. Il s'agit du fichier maître du rapport, une série d'autres fichiers (deux par composants) seront créés dans le répertoire. Nous inscrivons « **heart.html** ».

Le rapport est généré. Il est automatiquement ouvert dans votre navigateur par défaut (FIREFOX en ce qui concerne ma machine). Il reprend la physionomie de Tanagra. Sur la gauche (1), nous avons le diagramme de traitements. Nous pouvons cliquer sur le composant de notre choix. Les résultats correspondants s'affichent dans la partie droite du navigateur (2).

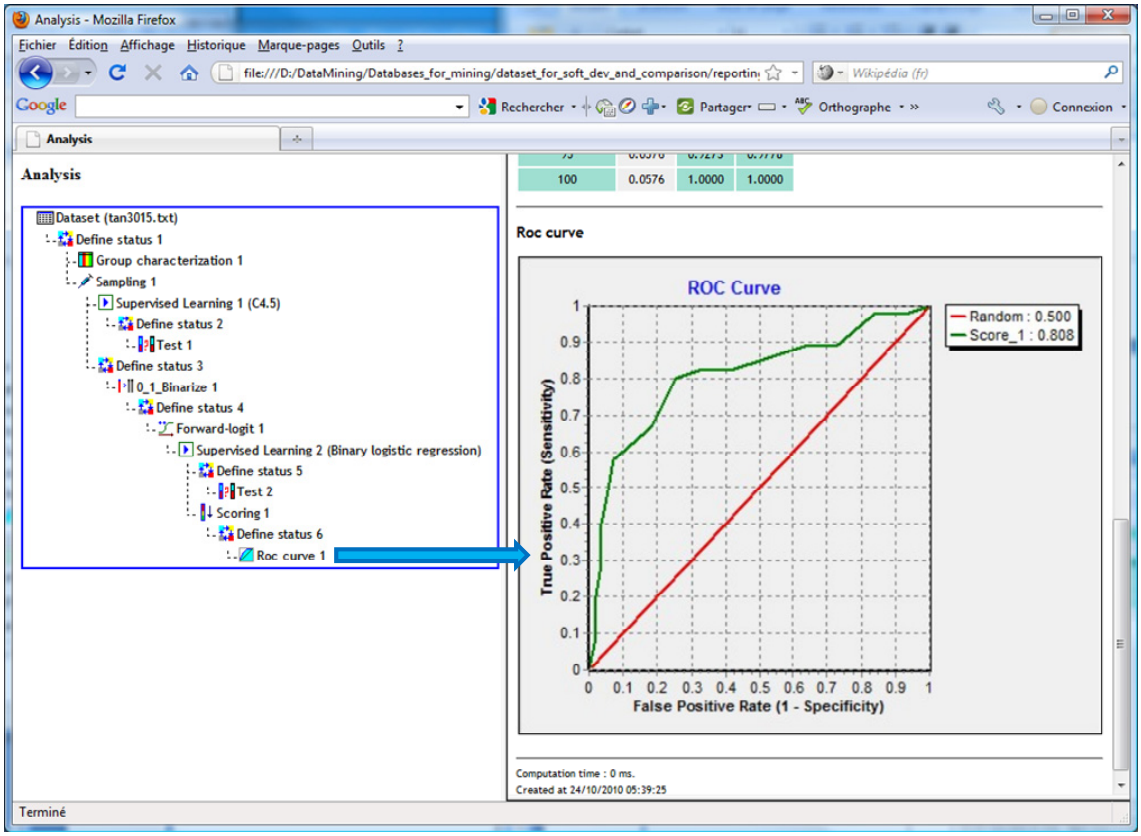


A partir de ce stade, nous pouvons visualiser les résultats indépendamment de Tanagra. C'est une condition primordiale pour une large diffusion du rapport. Un simple navigateur suffit pour consulter les résultats des calculs.

Si l'on choisit de voir l'arbre de décision par exemple, en cliquant sur « Supervised Learning 1 (C4.5) », nous obtenons :



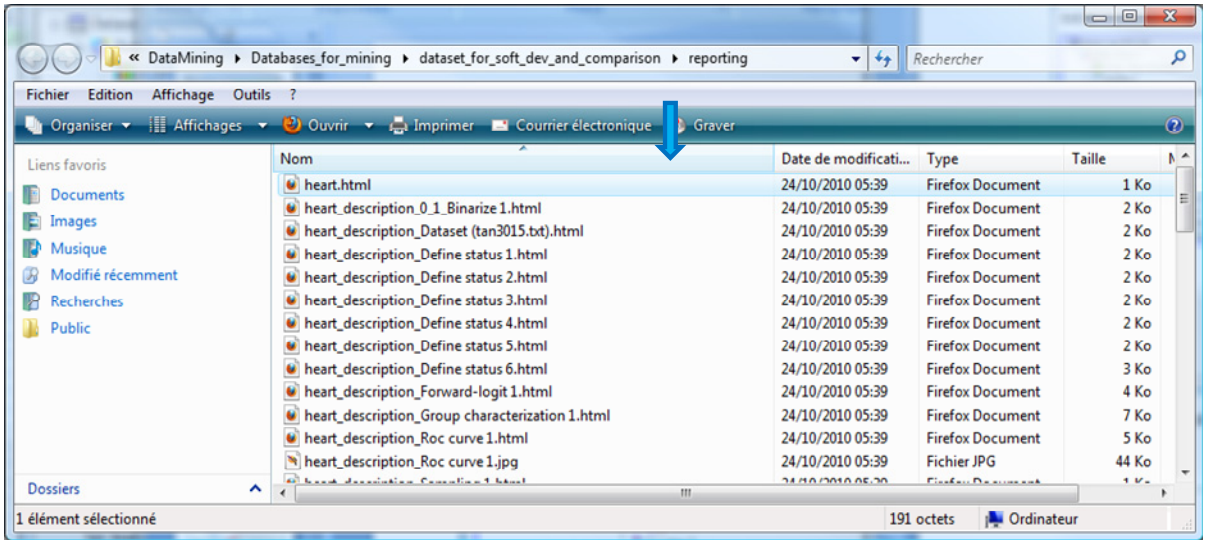
De la même manière, si nous sélectionnons « Roc curve 1 », nous avons :



Le graphique « Roc Curve » est inséré dans la page HTML.

Evidemment, le rapport peut être consulté avec n'importe quel navigateur (Internet explorer, Google Chrome, Safari, etc.). **Comme il est composé d'une série de fichiers HTML, nous pouvons les charger sélectivement dans un traitement de texte et les compléter à notre guise** (pour l'insertion de commentaires supplémentaires par exemple, etc.).

Pour la diffusion, il suffit de copier le contenu du répertoire de sauvegarde. C'est celui du fichier maître « heart.html ». Nous y trouvons tous les fichiers associés au rapport.



4 Conclusion

A partir d'une étude on ne peut plus classique de Data Mining, nous avons montré dans ce didacticiel qu'il était facile de récupérer les sorties de Tanagra dans les outils Office (via le tableur Excel), voire de produire directement un rapport consultable et diffusable indépendamment du logiciel.

Cela a été rendu possible par l'adoption du standard HTML pour la description des sorties. Pour le coup, ce choix a été particulièrement heureux. Ainsi, durant mes travaux dirigés, j'exploite énormément le fait de pouvoir naviguer facilement entre Tanagra et le tableur Excel (envoi des données d'Excel vers Tanagra via la macro complémentaire Tanagra.xla, copie des résultats de Tanagra dans Excel pour post traitements). Les vertus pédagogiques du tableur dans l'enseignement des techniques statistiques et de data mining sont indéniables (pour s'en persuader, voir par exemple la page Excel'Ense de la Revue Modulad – http://www-rocq.inria.fr/axis/modulad/excel.htm).