

Objectif

Codage des variables prédictives catégorielles pour la régression logistique.

Lorsque les variables prédictives (variables indépendantes, exogènes, etc.) sont catégorielles, les méthodes d'apprentissage supervisé telles que la régression logistique et l'analyse discriminante ne peuvent pas être mises en œuvre directement. Il est nécessaire de recoder les variables.

La technique la plus connue est certainement le codage disjonctif complet ou codage 0/1. Chaque modalité de la variable originelle devient une variable binaire qui prend la valeur 1 lorsque la modalité est présente pour un individu, 0 sinon. Puisque la somme de ces nouvelles variables est systématiquement égal à 1, il est d'usage d'omettre la dernière modalité qui peut être déduite des autres c.-à-d. si toutes les variables binaires prennent la valeur 0, on en déduit que l'individu porte la dernière valeur, elle devient la modalité de référence.

Dans ce didacticiel, nous montrons comment utiliser le composant 0_1_ BINARIZE pour transformer une variable catégorielle en une série de variables binaires que nous introduisons dans une régression logistique.

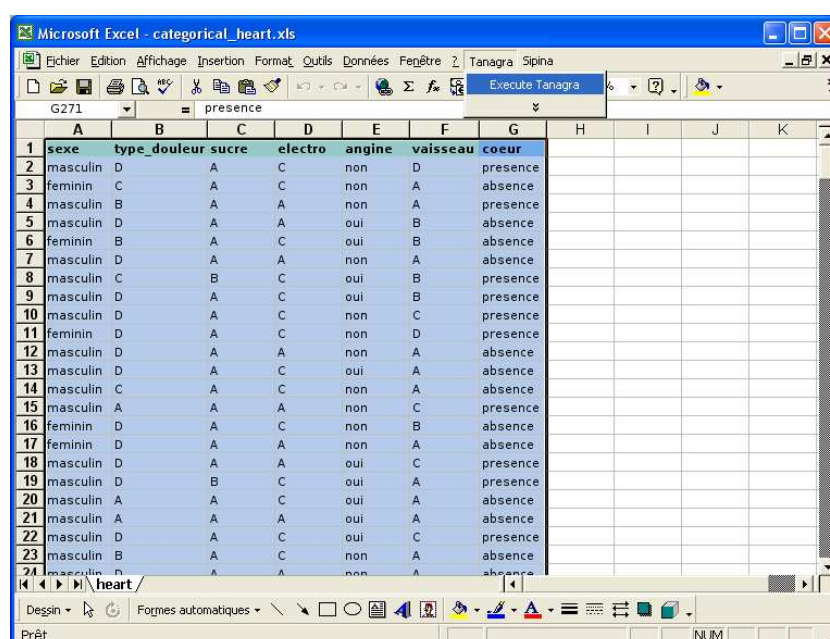
Données

Nous utilisons le fichier CATEGORICAL_HEART.XLS. Il s'agit de prédire la variable COEUR à l'aide d'une série de variables prédictives catégorielles (SEXE...VAISSEAU). Le fichier comporte 270 observations.

Codage des variables prédictives catégorielles

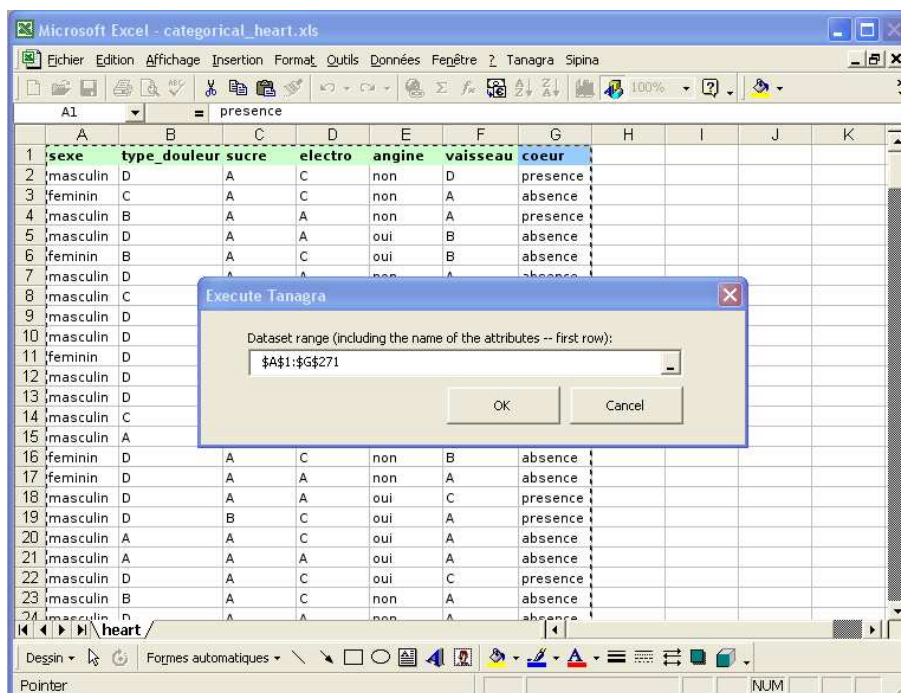
Création du diagramme

Le plus simple est d'ouvrir le fichier dans le tableur EXCEL. Si la macro complémentaire TANAGRA.XLA a été installée (voir le didacticiel associé sur le site web, cette macro est disponible depuis la version 1.4.11 de TANAGRA), un nouveau menu est disponible dans le tableur. Nous devons sélectionner les données puis activer le menu TANAGRA / EXECUTE TANAGRA.

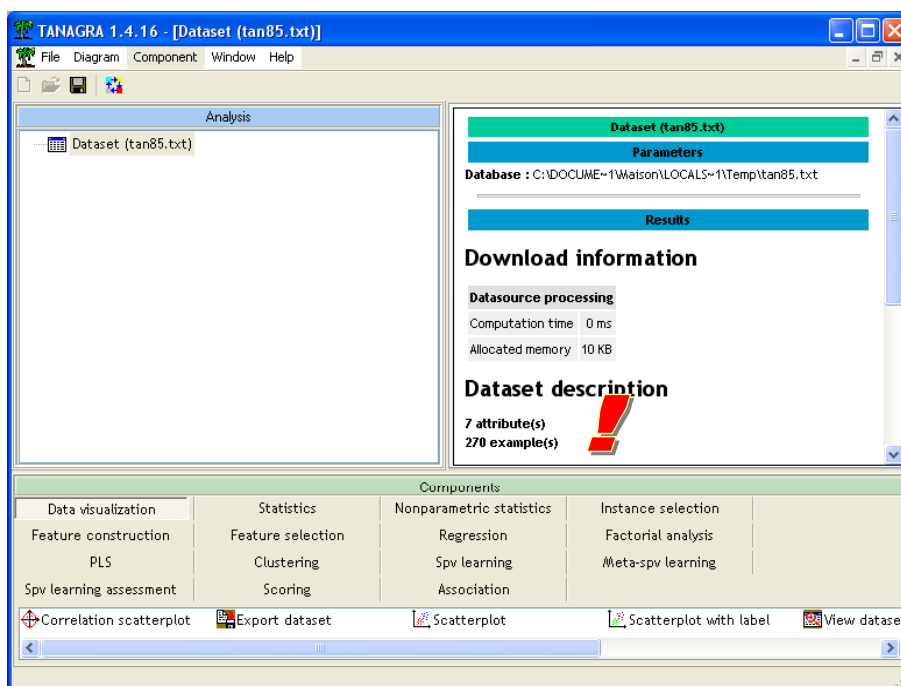


	A	B	C	D	E	F	G	H	I	J	K
1	sexe	type_douleur	sucre	electro	angine	vaisseau	coeur				
2	masculin	D	A	C	non	D	presence				
3	feminin	C	A	C	non	A	absence				
4	masculin	B	A	A	non	A	presence				
5	masculin	D	A	A	oui	B	absence				
6	feminin	B	A	C	oui	B	absence				
7	masculin	D	A	A	non	A	absence				
8	masculin	C	B	C	oui	B	presence				
9	masculin	D	A	C	oui	B	presence				
10	masculin	D	A	C	non	C	presence				
11	feminin	D	A	C	non	D	presence				
12	masculin	D	A	A	non	A	absence				
13	masculin	D	A	C	oui	A	absence				
14	masculin	C	A	C	non	A	absence				
15	masculin	A	A	A	non	C	presence				
16	feminin	D	A	C	non	B	absence				
17	feminin	D	A	A	non	A	absence				
18	masculin	D	A	A	oui	C	presence				
19	masculin	D	B	C	oui	A	presence				
20	masculin	A	A	C	oui	A	absence				
21	masculin	A	A	A	oui	A	absence				
22	masculin	D	A	C	oui	C	presence				
23	masculin	B	A	C	non	A	absence				
24	masculin	B	A	A	non	A	absence				

Une boîte de confirmation apparaît. Nous vérifions que les coordonnées de la plage de cellules sont correctes et nous validons.



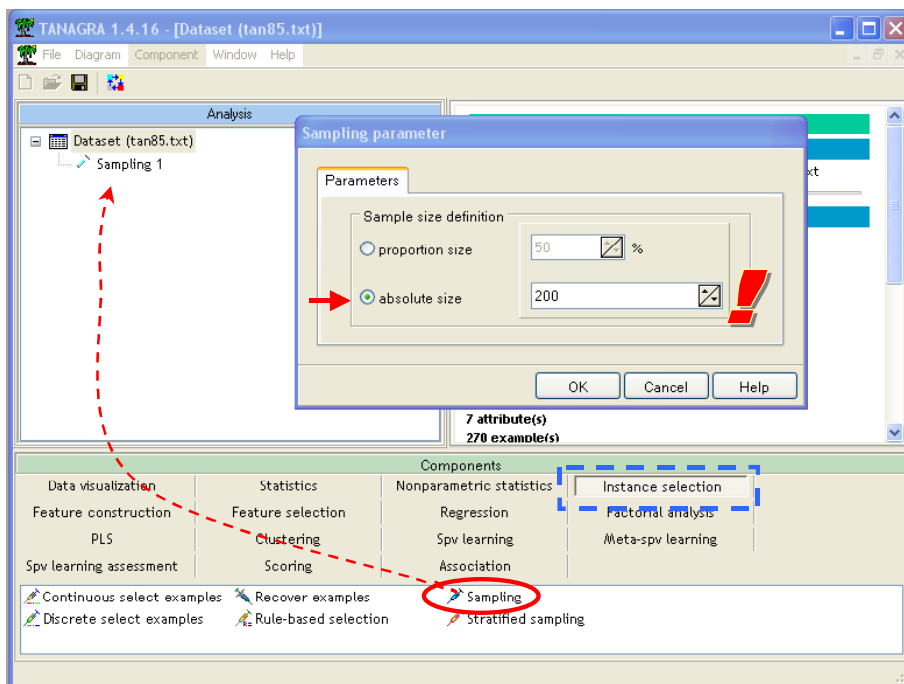
TANAGRA est automatiquement démarré, nous constatons que 270 observations et 7 variables sont disponibles.



Subdivision de la base en apprentissage et test

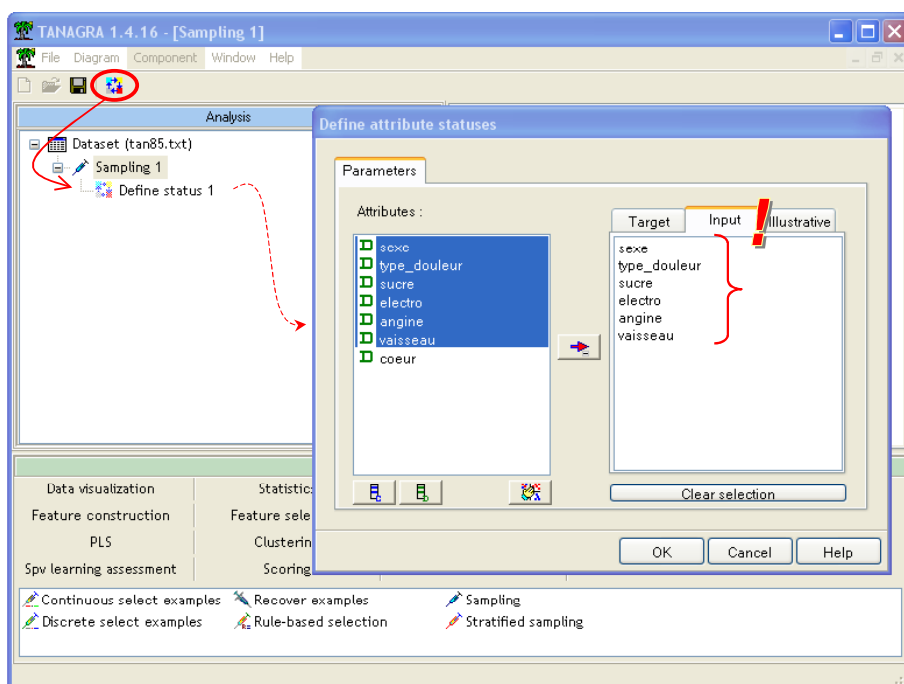
Pour disposer d'une évaluation non biaisée des performances du modèle de prédiction, nous subdivisons la base de données en réservant 200 observations pour l'apprentissage, la construction du modèle, et 70 observations pour le test, l'évaluation des performances du modèle.

Nous insérons à cet effet le composant SAMPLING (onglet INSTANCE SELECTION) dans le diagramme. Nous activons le menu PARAMETERS pour le paramétrer de manière à sélectionner 200 observations. Le reste est donc mis de côté et sera utilisé plus tard lors de l'évaluation.

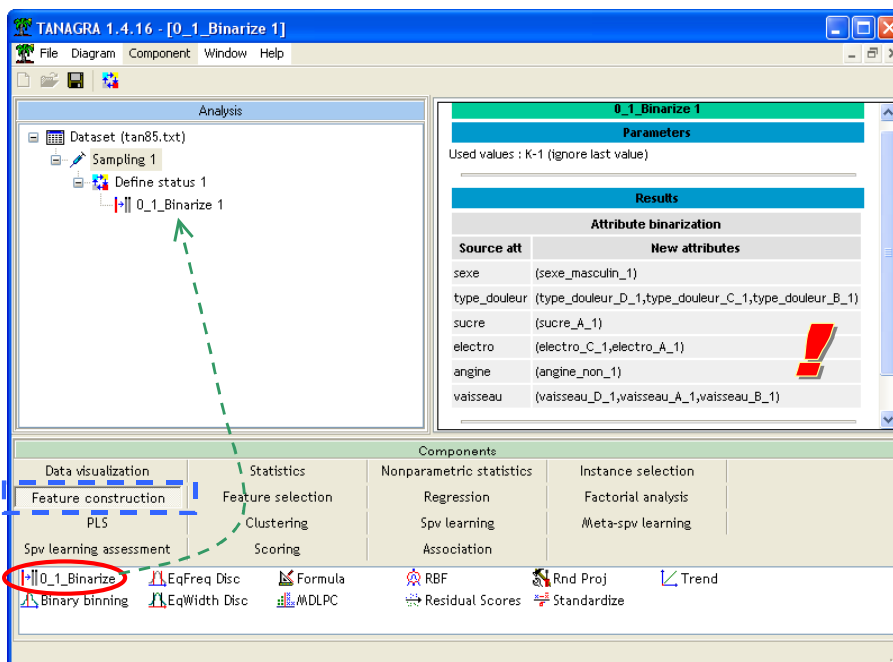


Transformation des variables prédictives catégorielles

Nous voulons prédire les valeurs de la variable CLASS à l'aide d'une régression logistique, les descripteurs à utiliser sont les 6 premières variables de la base. Si nous réalisons directement cette opération, TANAGRA produira une erreur indiquant que l'opération n'est pas possible en l'état. Il nous appartient de transformer de manière adéquate les variables prédictives en utilisant le composant 0_1_BINARIZE. Pour ce faire, nous insérons le composant DEFINE STATUS dans le diagramme, le plus simple est de passer par le raccourci dans la barre d'outil, et nous plaçons en INPUT les 6 variables prédictives candidates.



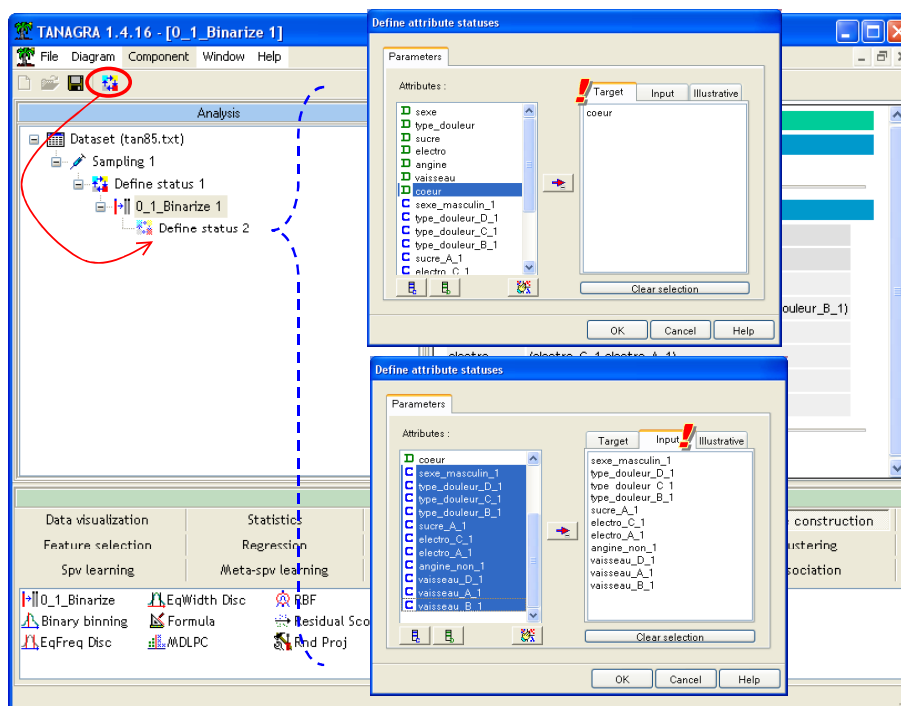
Nous insérons alors le composant 0_1_BINARIZE (onglet FEATURE CONSTRUCTION) puis nous activons le menu VIEW pour accéder aux résultats.



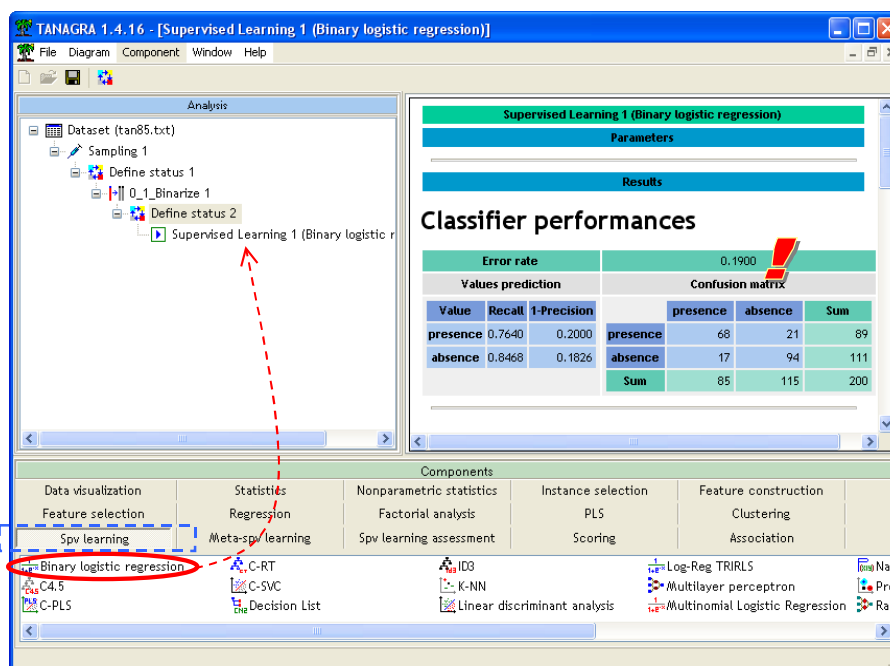
Nous constatons que chaque variable catégorielle à K modalités a été transformée en une série de (K-1) variables binaires. Il est possible, en modifiant le paramétrage du composant, de produire explicitement K variables. Cela peut s'avérer nécessaire pour certaines méthodes statistiques.

Régression logistique

Les variables maintenant recodées peuvent être introduites dans une régression logistique. Nous insérons de nouveau le composant DEFINE STATUS dans le diagramme, nous plaçons en TARGET la variable CLASS à prédire, et en INPUT les variables binaires.



Puis il ne nous reste plus qu'à ajouter le composant BINARY LOGISTIC REGRESSION situé dans l'onglet SPV LEARNING. Nous activons le menu VIEW.



Le taux d'erreur en apprentissage est de 19.0%. Au niveau de signification de 5%, nous constatons que seules 4 variables présentes un coefficient significativement différent de 0 au sens de la statistique de WALD.

Adjustement quality

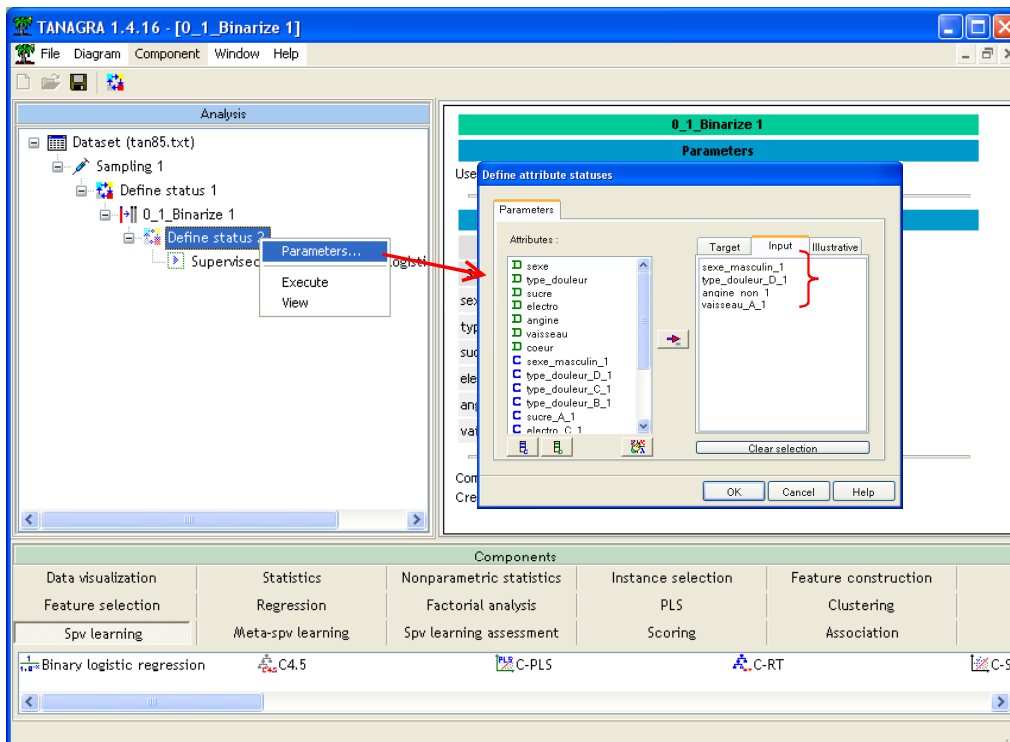
Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	159.4172
Chi-2	115.4167
P(>Chi-2)	0.0000

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	1.216147	-	-	-
sexe_masculin_1	1.812569	0.4704	14.8506	0.0001
type_douleur_D_1	2.452703	0.8692	7.9628	0.0048
type_douleur_C_1	1.057939	0.8852	1.4283	0.2320
type_douleur_B_1	0.663332	0.9715	0.4662	0.4947
sucre_A_1	0.384675	0.5761	0.4459	0.5043
electro_C_1	-1.776383	1.8446	0.9274	0.3355
electro_A_1	-2.006760	1.8413	1.1878	0.2758
engine_non_1	-1.389320	0.4613	9.0719	0.0026
vaisseau_D_1	-0.174881	1.1084	0.0249	0.8746
vaisseau_A_1	-2.661483	0.7256	13.4557	0.0002
vaisseau_B_1	-0.902098	0.7982	1.2771	0.2584

Il existe plusieurs stratégies de sélection de variables, plus ou moins avantageuses. Il faut surtout les voir comme autant de techniques pour produire des scénarios de variables pertinentes. Dans notre cas, nous mettons en œuvre une stratégie simple : nous supprimons de la régression les

variables non significatives à 5%. Pour cela, nous revenons donc au composant DEFINE STATUS 2 qui précède la régression logistique et nous modifions le paramétrage (menu PARAMETERS) de manière à ne conserver en INPUT que les variables SEXE_MASCULIN, TYPE_DOULEUR_D, ANGINE_NON, VAISSEAU_A. Le mieux est de vider la sélection courante en cliquant sur le bouton CLEAR SELECTION, puis de procéder à la sélection.



Nous relançons de nouveau la régression logistique avec le menu VIEW.

Classifieur performances

Error rate			0.185			
Values prediction			Confusion matrix			
Value	Recall	1-Precision		presence	absence	Sum
presence	0.7416	0.175	presence	66	23	89
absence	0.8739	0.1917	absence	14	97	111
			Sum	80	120	200

Classifieur characteristics

Data description

Target attribute	values)
# descriptors	4

Adjustement quality

Predicted attribute	coeur
Number of examples	200
-2 Log Likelihood	164.1061
Chi-2	110.7279
P(>Chi-2)	0

Attributes in the equation

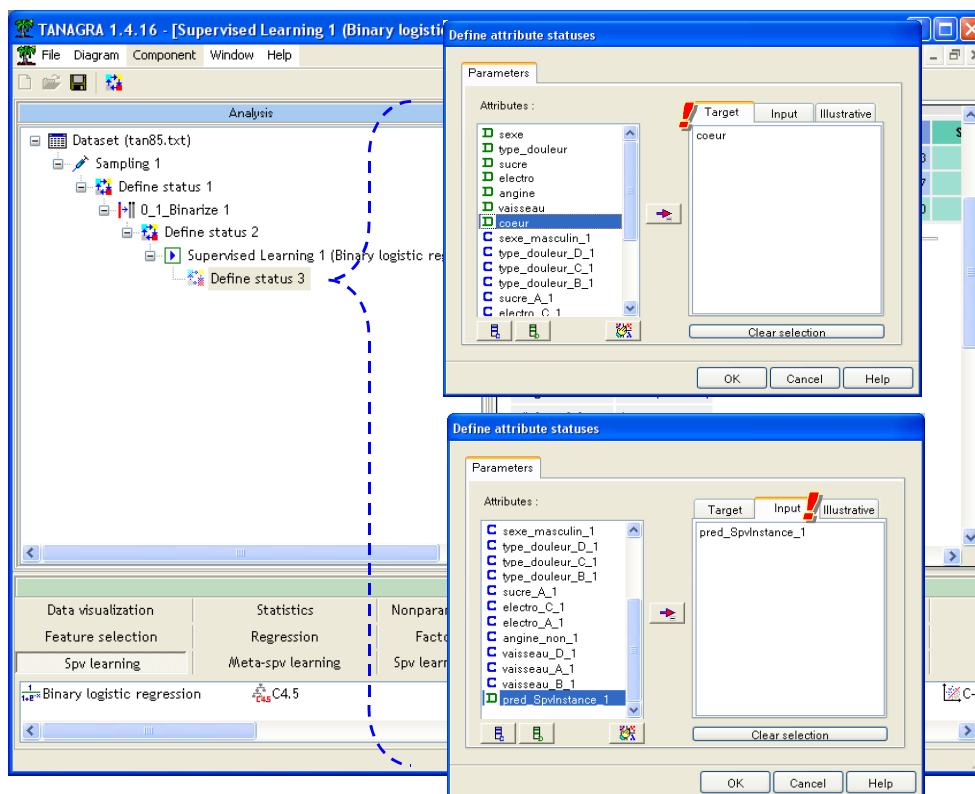
Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.127911	-	-	-
sexe_masculin_1	1.659938	0.4354	14.5359	0.0001
type_douleur_D_1	1.772698	0.4165	18.1156	0
angine_non_1	-1.317853	0.4465	8.7096	0.0032
vaisseau_A_1	-2.024704	0.4146	23.8492	0

Le taux d'erreur en apprentissage est de 18.5%. Toutes les variables sont maintenant significatives à 5%.

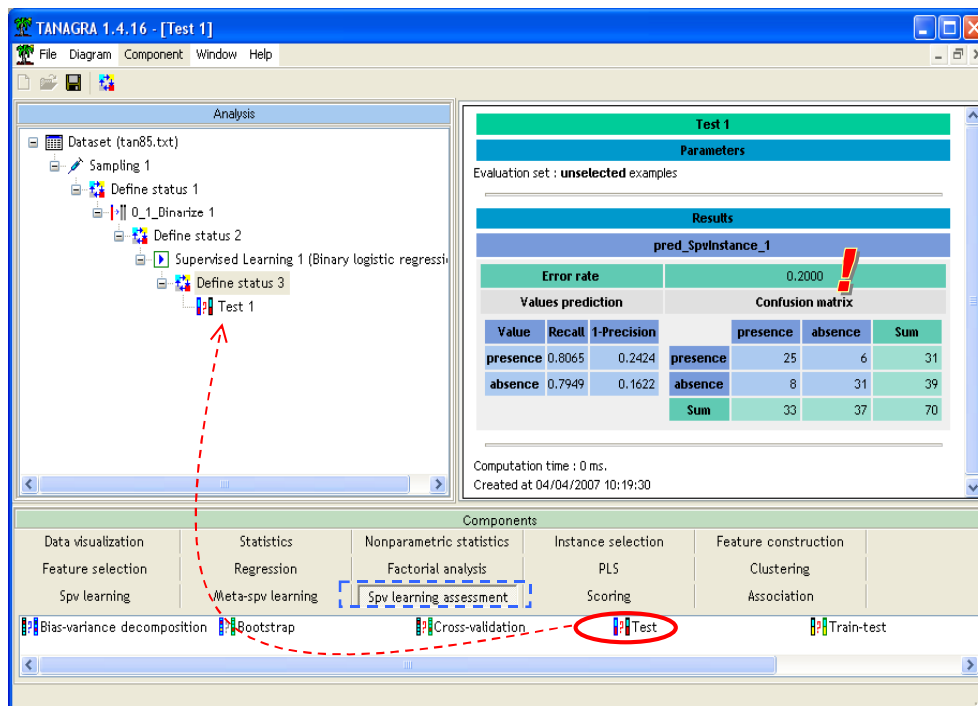
Evaluation sur l'ensemble test

Reste alors à évaluer ce modèle de prédiction sur la seconde partie du fichier, les 70 observations que nous avons laissées de côté.

Nous insérons le composant DEFINE STATUS. Nous plaçons en TARGET la variable à prédire CŒUR et en INPUT la variable produite par la régression logistique : PRED_SPV_INSTANCE_1. Cette variable correspond à la prédiction du modèle effectué sur la totalité du fichier, c.-à-d. sur les observations sélectionnées et non-sélectionnées.



Pour mesurer le taux d'erreur en test, il faut ajouter le composant TEST (onglet SPV LEARNING ASSESSMENT) dans le diagramme. Par défaut, il confronte les vraies valeurs de la variable à prédire avec la prédiction du modèle sur les données non sélectionnées, donc celles qui n'ont pas participé à l'apprentissage. C'est exactement ce qui nous convient ici. Nous activons le menu VIEW pour accéder aux résultats.



Le taux d'erreur en test est de 20%. *Grosso modo*, si nous appliquons ce modèle dans la population, nous avons approximativement 20% de « chances » d'effectuer une mauvaise prédiction.

Conclusion

Dans la littérature, lorsque nous utilisons des variables prédictives binaires, il est souvent conseillé d'utiliser la régression logistique plutôt que l'analyse discriminante. Tout simplement parce que cette dernière repose sur une hypothèse probabiliste, la distribution conditionnelle gaussienne, qui est plus restrictive, *a priori* mal adaptée pour les variables 0/1.

Dans la pratique, nous constatons rarement des différences flagrantes de performances. Notamment parce que géométriquement ces deux techniques produisent une frontière linéaire dans l'espace de représentation. Si le biais de représentation est donc le même, le biais d'apprentissage est effectivement différent, la première maximise la vraisemblance, la seconde correspond à une méthode des moindres carrés. C'est en sens que peuvent survenir des solutions différentes, sur des cas pathologiques assez rares néanmoins.

Dans notre exemple, l'analyse discriminante appliquée sur les mêmes variables produit exactement les mêmes performances : le taux d'erreur en test est de 20%.