UNIVERSITÉ NATIONALE DU VIETNAM (UNV)
Institut Francophone pour l'Innovation (IFI)

Option : Système Intelligent et Multimédia

Niveau : **Master I**

Cours : Fouille de Données

Professeur : NGUYEN Thà Minh Huyen

Rapport final des TPs

Préparé par le Binôme 13 :

Mongetro GOINT & Myderson SEMEURAND

Juillet 2018

Table des matières

Introduction

Dans le secteur financier, particulièrement au niveau du système bancaire, la bonne gestion de prêt aux clients pose souvent des difficultés. Cela, surtout au niveau de l'identification des clients qui pourraient être classés parmi ceux ayant un « **Bon** » ou un « **Mauvais** » dossier de prêt.

Ainsi, dans le cadre des TPs du cours de Fouille de Données au master I (en Système intelligent et multimédia) à l'IFI, nous avons choisi de travailler sur un jeu de données d'une institution financière relatives aux dossiers de ses clients.

Notre objectif est de prédire le risque de crédit des futurs emprunteurs : « *Ne pas octroyer un crédit qu'aux emprunteurs présentant un faible risque et un fort potentiel en ce qui concerne leurs dossiers de prêt* ».

Description et résumée des données

Comme c'était mentionné dans notre premier rapport, dans notre jeu de données, nous identifions un ensemble de variables. Parmi ces dernières nous comptons 17 « variables d'entrée » ou « variables explicatives » et une « variable de sortie » ou « variables d'intérêt ».

La variable qu'on souhaite prédire (Variable d'intérêt) : Dossier prêt

L'ensemble de données de la variable d'intérêt qui est un attribut nominal (à valeurs non ordonné) comporte **2 valeurs :**

a) Bon : état du dossier d'un client éligible (qui a bien rembourser l'argent emprunter à l'institution financière).

b) Mauvais : état du dossier d'un client non-éligible (qui avait des difficultés de rembourser à l'institution financière).

- Les variables explicatives de la variable d'intérêt sont au nombre de **17** :

1- Solde : Montant du compte du client au moment du prêt (en dollars), attribut ordinal.

2- Durée : Nombre de mois pour rembourser le prêt (3 à 52), attribut ordinal.

3- Remboursement antérieur : Etat des remboursements des prêts antérieur (Intégral, partiel, délicat), attribut nominal.

4- Objectif : Raison pour lequel le prêt est effectué (Logement, achat de véhicule, consommation), attribut nominal.

5- Montant : Somme d'argent empruntée par le client (300 à 30 000 dollars), attribut ordinal.

6- Epargne : La valeur de l'épargne du client (en dollars), attribut ordinal.

7- Ancienneté chez l'employeur actuel : Durée d'ancienneté chez l'employeur actuel (en années), attribut ordinal.
8- Endettement : Pourcentage du revenu disponible, attribut ordinal.
9- Statut Marital : Etat matrimonial du client (Marié, Célibataire, Divorcé), attribut nominal.
10- Sexe : Sexe de du client (Masculin /Féminin), attribut nominal.
11- Ancienneté dans le ménage : Durée d'ancienneté dans le ménage actuel (en années), attribut à valeur absolue.
12- Actif les plus importants : Possession du client au moment du prêt (Terrain, Véhicule, Logement), attribut nominal.
13- Age : Age du client (en année), attribut à valeur absolue.
14 : Autres prêts en cours : Prêts effectués par le client autre que le prêt à effectuer actuellement (Banques concurrentes, magasins, aucun), attribut nominal.
15- Type de logement : Type de logement actuel du client (Locataire, Propriétaire, A titre gracieux), attribut nominal.
16- Nombre de prêts antérieurs : Nombre de prêts antérieurs effectués par le client dans cette banque, attribut à valeur absolue.
17- Occupation : Activité du client (Cadre, Salarié expérimenté, Salarié sans expérience avec domicile, Salarié sans expérience sans domicile), attribut nominal.

Présentation de la méthode étudiée (LDA)

Dans le cadre de ce travail, nous avons choisi d'utiliser la méthode **LDA**.

En statistique, l'Analyse Discriminante Linéaire (aussi **LDA**, en anglais *: **l**inear **d**iscriminant **a**nalysis*) fait partie des techniques d'analyse discriminante prédictive.

Initiée par Fisher 1936, cette technique consiste à expliquer et à prédire l'appartenance d'un individu à une classe (groupe) prédéfinie à partir de ses caractéristiques mesurées à l'aide de variables prédictives.

En d'autres termes, l'analyse discriminante linéaire part de la connaissance de la partition en classes des individus d'une population et cherche les combinaisons linéaires des variables décrivant les individus qui conduisent à la meilleure discrimination entre les classes.

Le point de départ de l'ADL est une matrice **X** de données observées (individus x variables) dont les éléments sont identifiés dans une (et une seule) des *k* classes possibles. L'idée de Fisher a été de créer une méthode pour choisir entre les combinaisons linéaires des variables qui maximise l'homogénéité de chaque classe (*Fisher 1936*).

Préparation de données

Dans cette étape, nous tenons à préparer nos données afin de faire une comparaison entre nos deux méthodes (LDA et SVM), nous tenons à faire une séparation entre nos différents types de variables (discrètes et continues).
En effet, nous utilisons le composant **0_1_BINARIZE** après avoir sélectionné tous les attributs discrets, mis à part la variable à prédire dans notre qui est DOSSIER PRÊT.

Après l'application de ce processus nous obtenons le diagramme de traitement suivant :

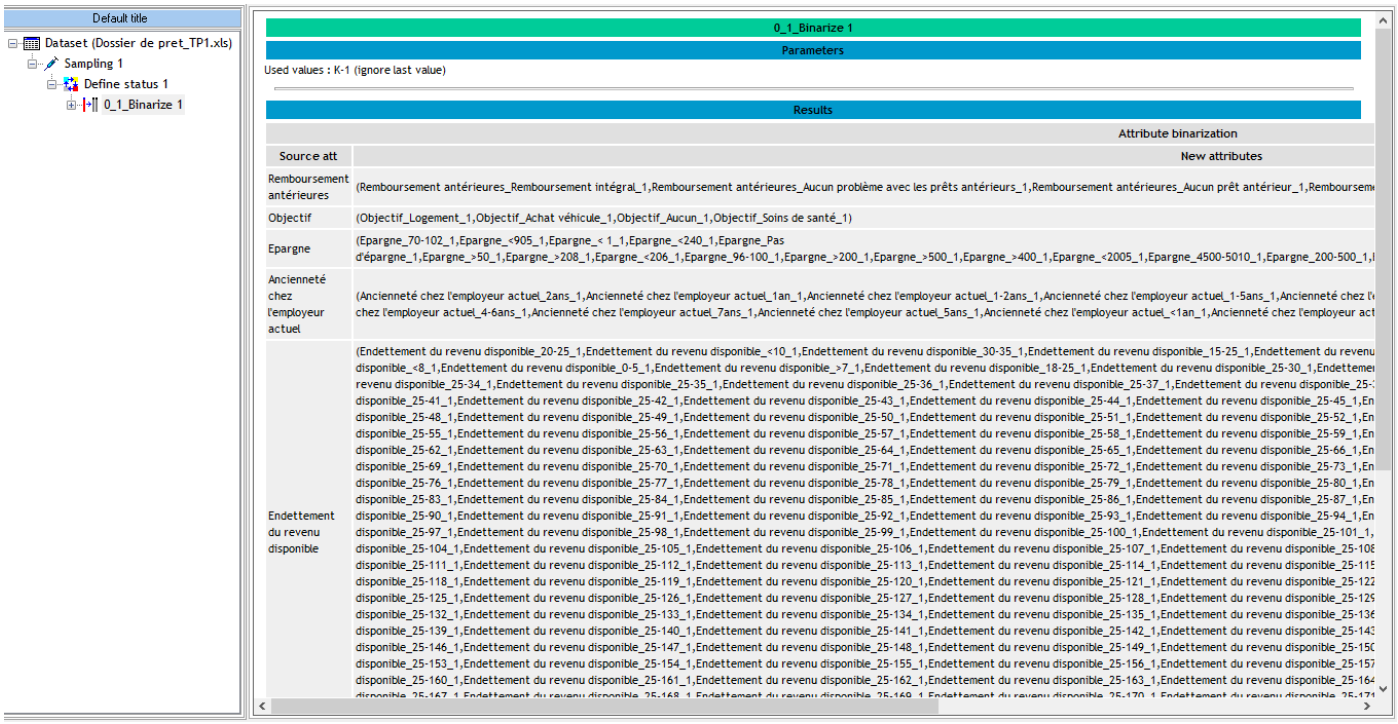


Figure1 : préparation des données

Choix des paramètres et application de la méthode étudiée au jeu de données et évaluation

Maintenant, après le lancement de l'apprentissage de la LDA, nous plaçons en INPUT tous les attributs continus et en TARGET l'attribut à prédire DOSSIER PRÊT. Alors, nous pouvons remarquer apparaitre les résultats dans la figure ci-dessous.

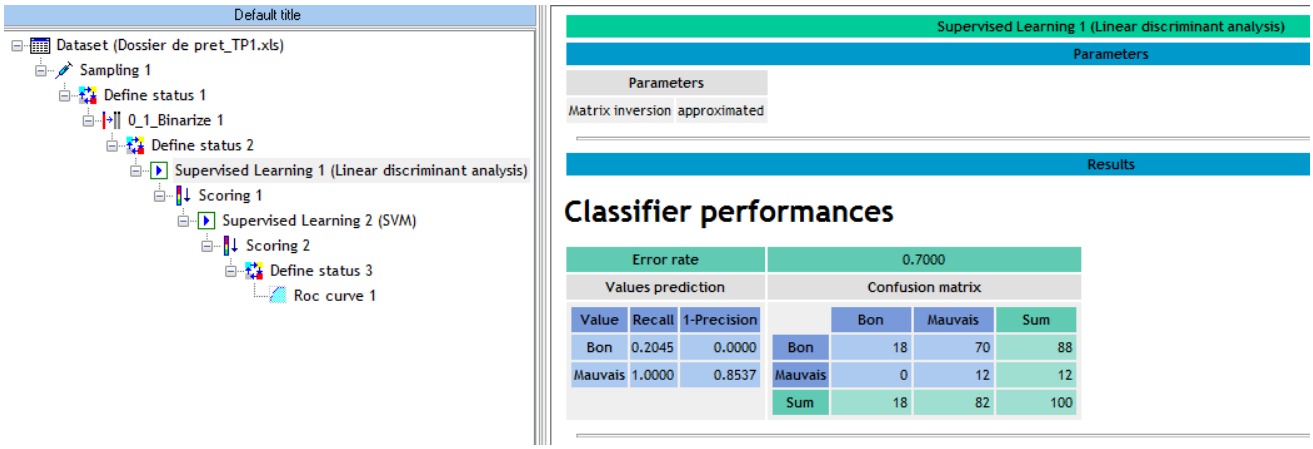


Figure2 : « Classifier performances » pour l'apprentissage LDA

Nous constatons bien que le modèle de prédiction a été construit à partir des données d'apprentissage (100 observations).

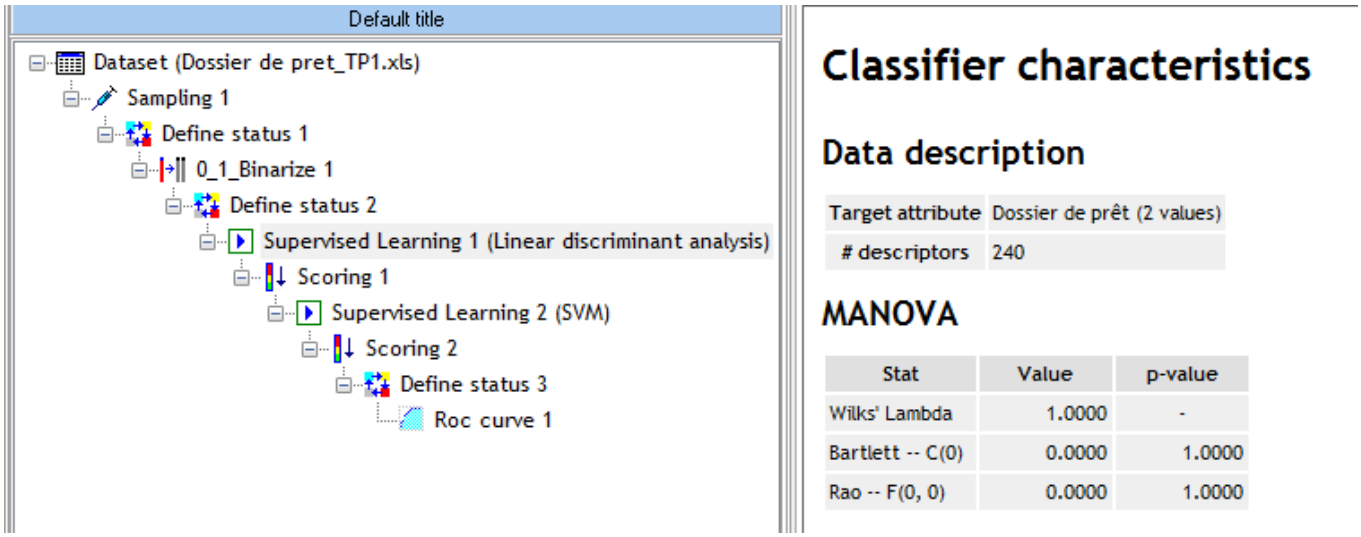


Figure3 : « Classifier characteristics » pour l'apprentissage LDA

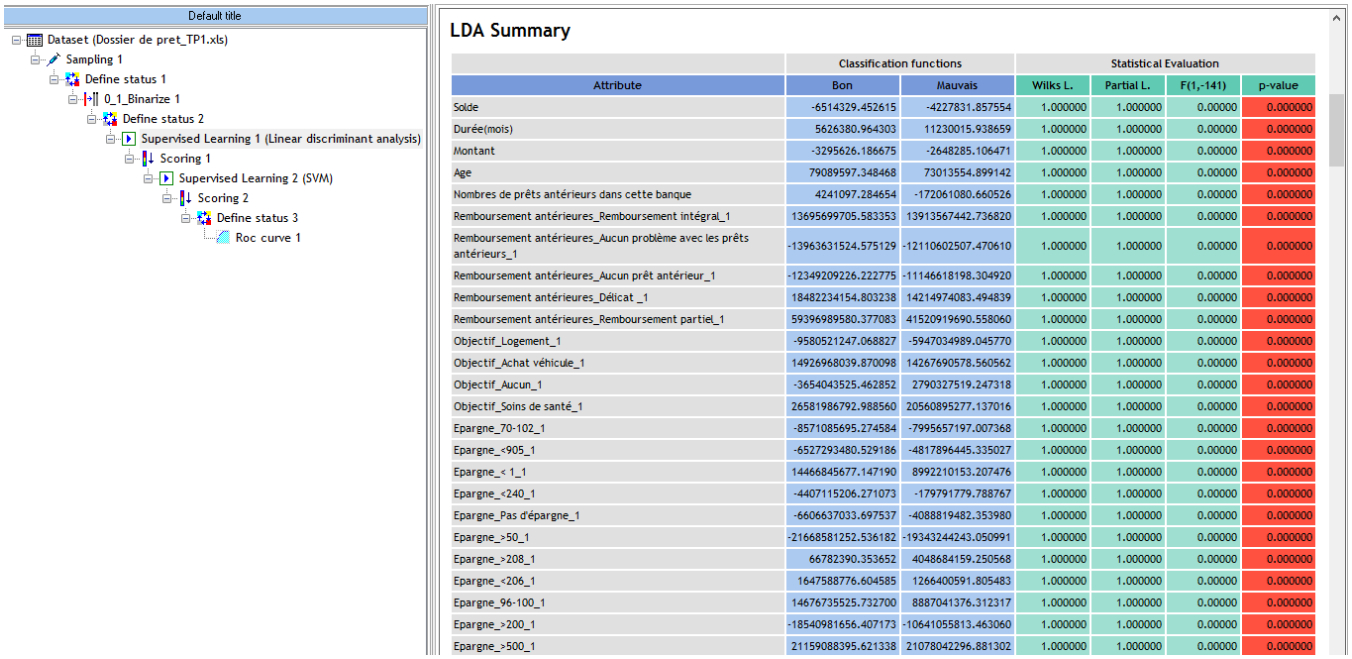


Figure4 : « LDA Summary »

SCORING pour la méthode LDA

Pour le *Scoring* de LDA, nous attribuons un score à tous les individus de l'ensemble de données en spécifier la modalité « **Bon** » comme modalité positive de la variable DOSSIER PRÊT.

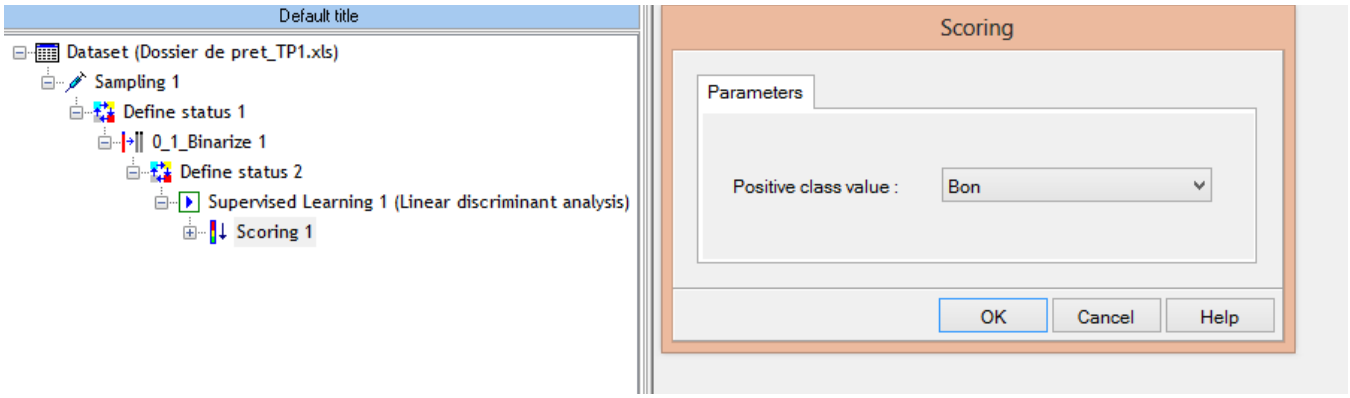


Figure5 : Paramétrage pour le *Scoring* LDA

Comparaison avec le résultat d'application d'une autre méthode d'apprentissage supervisé

Pour une comparaison avec notre méthode choisi (LDA), nous avons choisi la méthode SVM pour laquelle nous allons présenter les résultats.

Tout d'abord, faisons un tour d'horizon sur la méthode SVM :

Les SVM ont été introduites par Vladimir Vapnik. La popularité des méthodes SVM, pour la classification binaire en particulier, provient du fait qu'elles reposent sur l'application d'algorithmes de recherche de règles de décision linéaires ("hyperplan séparateur"), la recherche s'effectuant toutefois dans un espace ("feature space") de très grande dimension, lequel est l'image de l'espace d'entrée original par une transformation Φ non linéaire.

Les machines à vecteur support se situent sur l'axe de développement de la recherche humaine des techniques d'apprentissage.

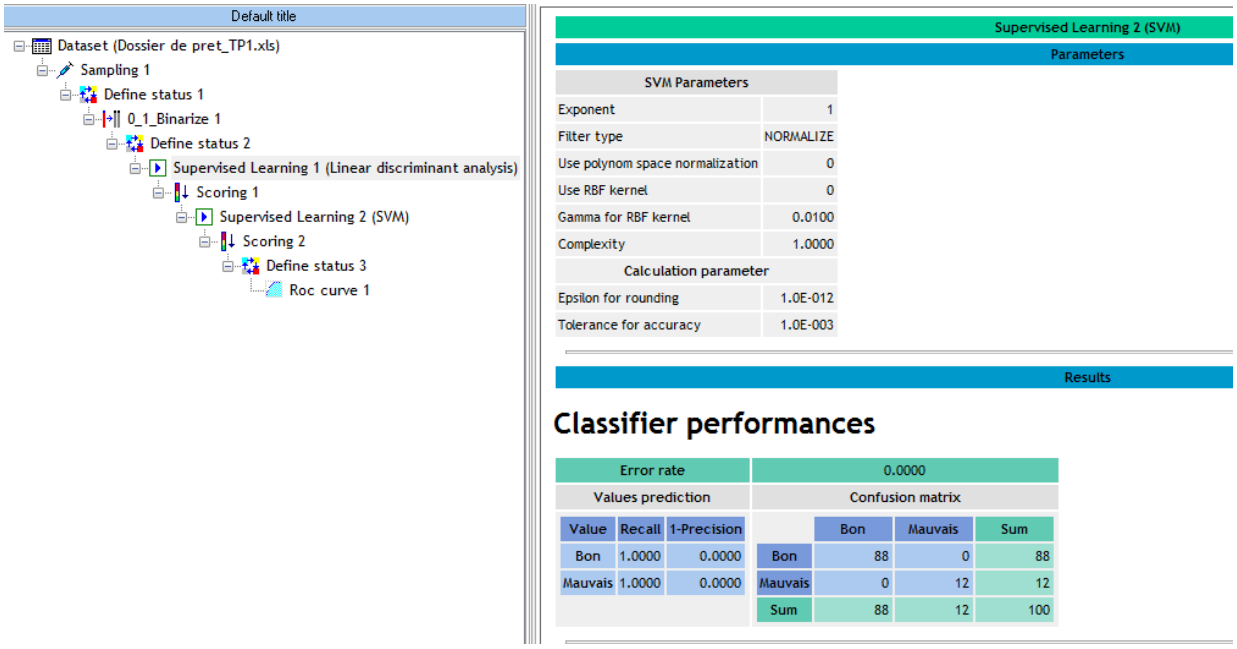


Figure6 : « Classifier performances » pour la méthode SVM

Comparativement à la méthode LAD, Il est à remarquer qu'il y a des différences dans les résultats obtenus pour la méthode SVM.
On peut constater des différences au niveau des résultats dans « *Values prediction* » et « *Confusion matrix* ».

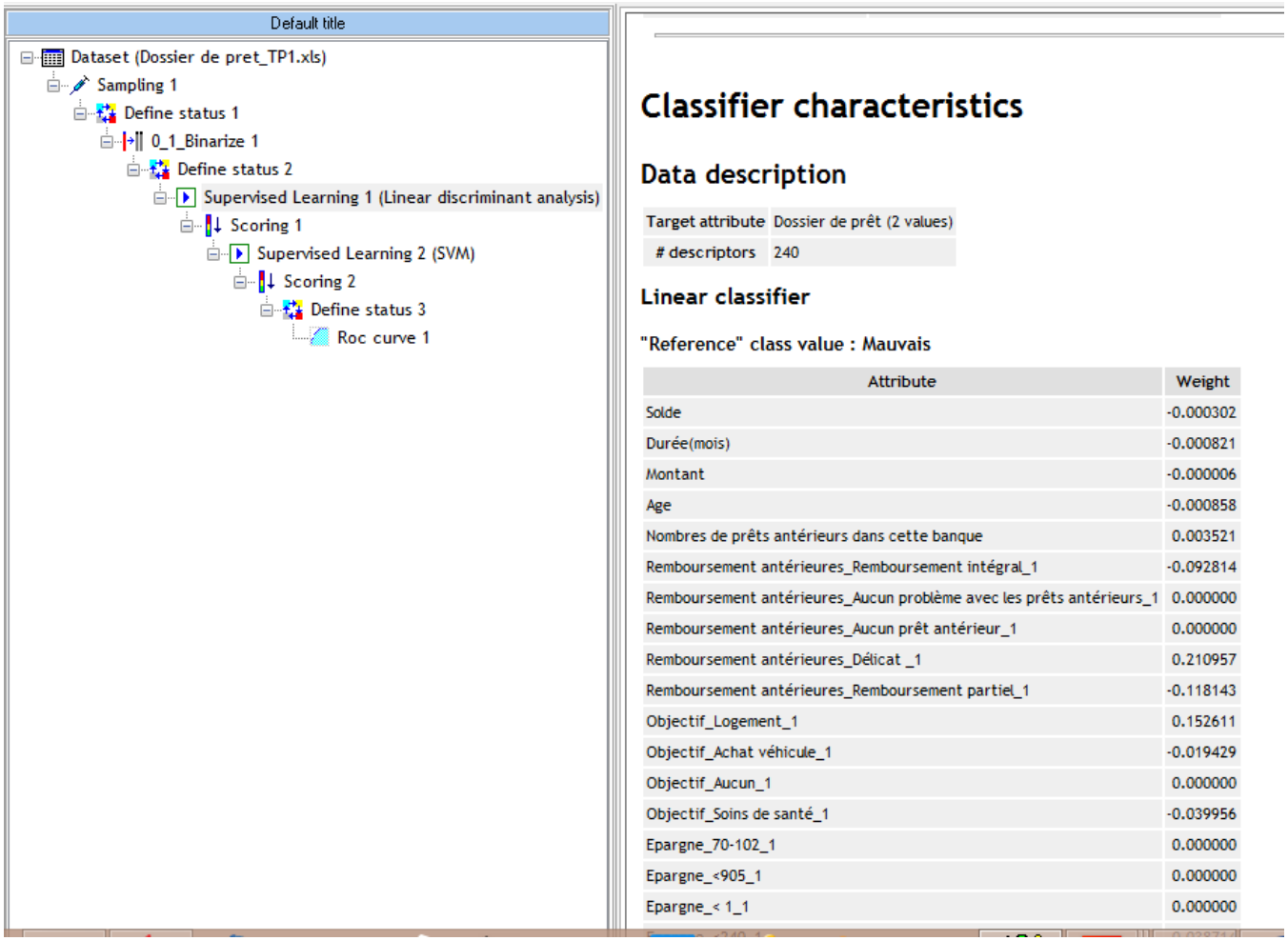


Figure7 : « Classifier characteristics » pour l'apprentissage SVM

SCORING pour la méthode SVM

Similaire à la méthode LDA, nous attribuons pour le *Scoring* de SVM un score à tous les individus de l'ensemble de données en spécifier la modalité « **Bon** » comme modalité positive de la variable DOSSIER PRÊT.

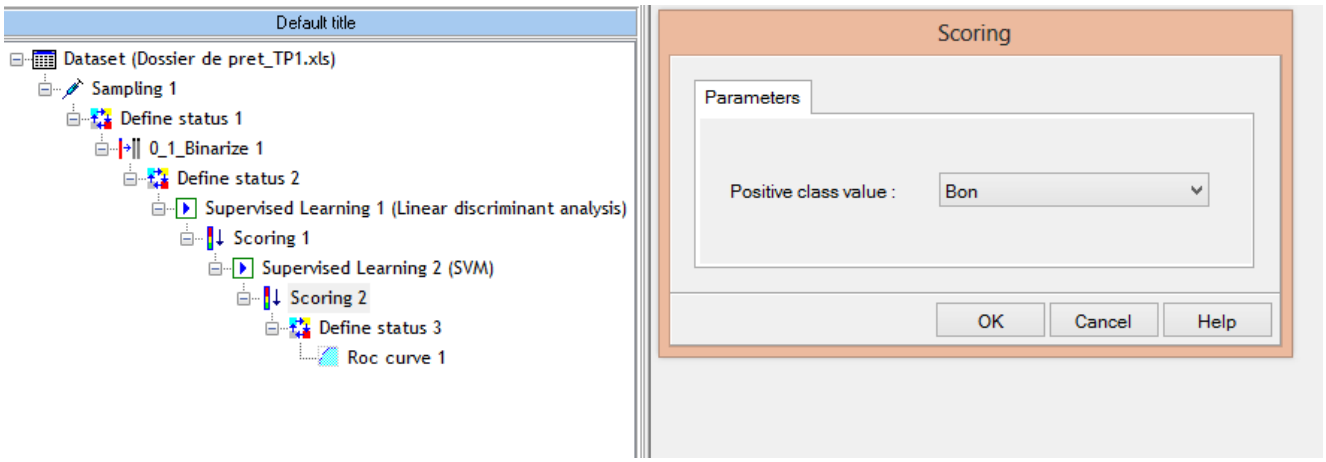


Figure8 : Paramétrage pour le *Scoring* SVM

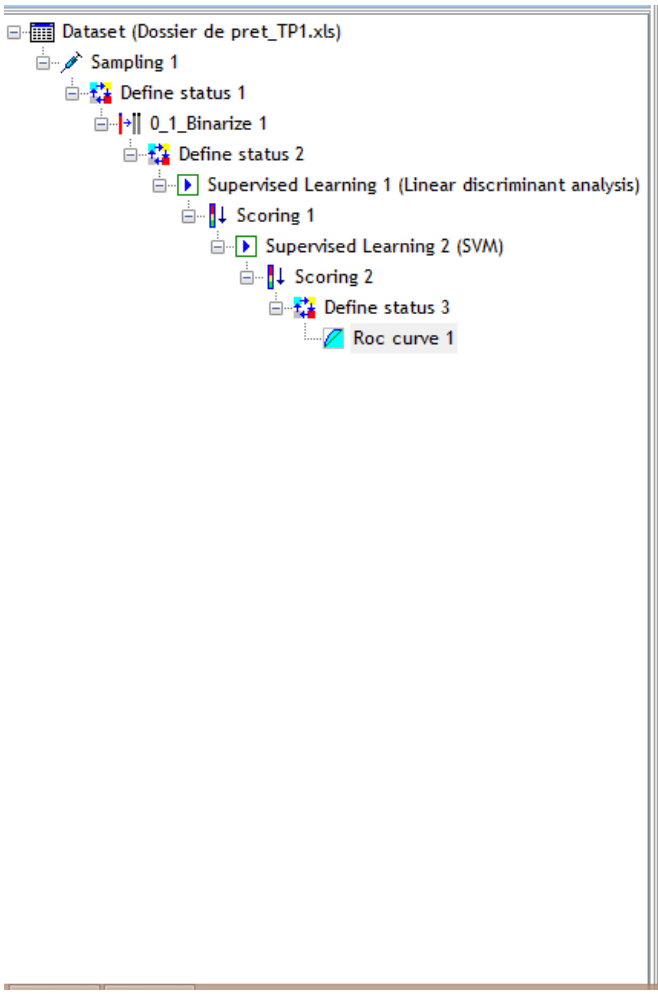
Construction de la courbes ROC

Afin de bien comparer nos deux méthodes choisies, nous tenons à construire la courbe ROC.

Pour construire la courbe ROC, nous tenons d'abord à spécifions l'attribut à prédire DOSSIER PRÊT en *TARGET* et les variables de Classement des individus (les scores) en *INPUT*.

Ensuite, nous plaçons le composant *ROC CURVE* dans le diagramme de traitements, en réglant les paramètres : la modalité de la variable à prédire qui représentera la modalité « positive » ; l'ensemble de données qui servira à construire la courbe.

Après tout, nous obtenons le résultat ci-dessous :



ROC Curve

Sample size : 104
Positive examples : 86
Negative examples : 18

Score Attribute	Score_1			Score_2		
AUC	0.7523			0.9780		
Target size (%)	Score	FP-Rate	TP-Rate	Score	FP-Rate	TP-Rate
0	1.0000	0.0000	0.0000	0.8937	0.0000	0.0000
5	1.0000	0.0556	0.0465	0.8839	0.0000	0.0581
10	1.0000	0.0556	0.1047	0.8728	0.0000	0.1163
15	1.0000	0.0556	0.1628	0.8684	0.0000	0.1744
20	1.0000	0.0556	0.2209	0.8576	0.0000	0.2326
25	1.0000	0.0556	0.2907	0.8484	0.0000	0.3023
30	1.0000	0.0556	0.3488	0.8446	0.0000	0.3605
35	1.0000	0.0556	0.4070	0.8342	0.0000	0.4186
40	1.0000	0.0556	0.4651	0.8250	0.0000	0.4767
45	1.0000	0.1667	0.5000	0.8194	0.0000	0.5349
50	1.0000	0.1667	0.5698	0.8142	0.0000	0.6047
55	1.0000	0.1667	0.6279	0.7962	0.0000	0.6628
60	0.0000	0.2222	0.6744	0.7761	0.0000	0.7209
65	0.0000	0.3889	0.6977	0.7452	0.0556	0.7674
70	0.0000	0.4444	0.7442	0.7329	0.0556	0.8256
75	0.0000	0.5000	0.8023	0.7278	0.0556	0.8953
80	0.0000	0.5000	0.8605	0.6735	0.1111	0.9419
85	0.0000	0.5000	0.9186	0.6226	0.1667	0.9884
90	0.0000	0.7222	0.9302	0.3149	0.3889	1.0000
95	0.0000	1.0000	0.9302	0.2758	0.6667	1.0000
100	0.0000	1.0000	1.0000	0.2661	1.0000	1.0000

Figure9 : Courbe Roc Curve pour les deux méthodes LDA et SVM

Dans la courbe *Roc Curve*, nous pouvons constater que les deux méthodes LDA et SVM proposent des performances pas toutes similaires (Score_1 : 0,752 ; Score_2 : 0,978), TP-Rate = 0.0000 des deux côtés pour la première ligne, mais avec de légère différence des deux côtés dans la plupart des autres lignes.

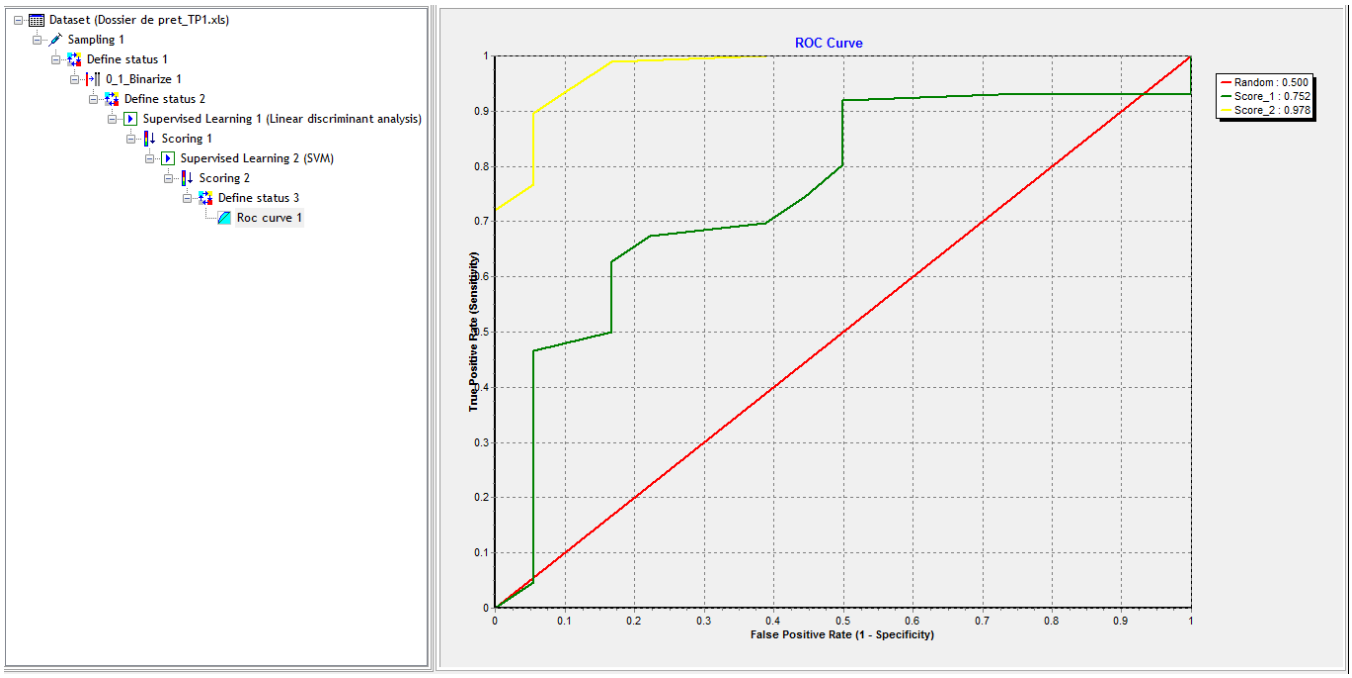


Figure10 : Graphe Roc Curve pour les deux méthodes LDA et SVM

Conclusion

Pour finir avec ce rapport, nous voulons porter quelques petites précisions sur la différence qui existe entre nos deux méthodes utilisées ci-dessus : LDA et SVM.

Autant que nous sachions, à propos de la LDA, cette dernière suppose que les données sont normalement distribuées, Tous les groupes sont distribués de manière identique, dans le cas où les groupes ont des matrices de covariance différentes, LDA devient Quadratic Discriminant Analysis. Tandis que SVM elle, suppose que tous les groupes sont totalement séparables, SVM utilise une « variable de marge » qui permet un certain chevauchement entre les groupes.

Aussi vrai que SVM se concentre uniquement sur les points qui sont difficiles à classer, tandis que LDA se concentre sur tous les points de données. Ces points difficiles sont proches de la limite de décision et sont appelés *vecteurs de support*.

Et LDA suppose que les points de données ont la même covariance et que la densité de probabilité est supposée être distribuée normalement. Mais SVM n'a pas une telle hypothèse. Donc, LDA est générative tandis que SVM est discriminante.

À partir de notre courbe ROC curve, nous observons que les performances des deux méthodes sont similaires dans certains cas mais pas totalement. Une remarque importante lorsqu'on observe bien notre graphe présenté ci-dessus, c'est que les résultats de performance des deux méthodes ne sont pas vraiment similaires.