**République Algérienne démocratique et populaire**

**Ministère de l'enseignement supérieur et de la recherche scientifique**

**Universite des sciences et de la technologie Houari Boumediene**

**Faculté d’informatique**

**Département des systèmes informatique**

**BIG DATA ANALYTICS**



**COMPTE RENDU TP ANADIS**

HADJI Mohamed Oussama

BERGHOUT Mohamed Wassim

Groupe 1

Décembre 2022

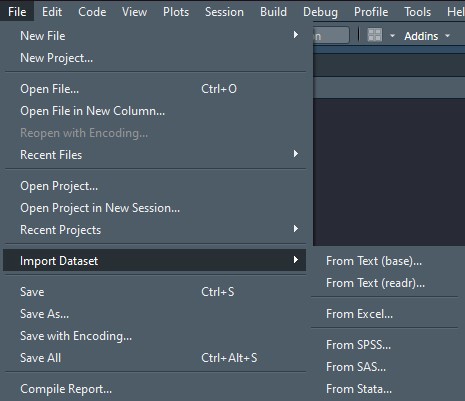
**Plan du compte rendu** **:**

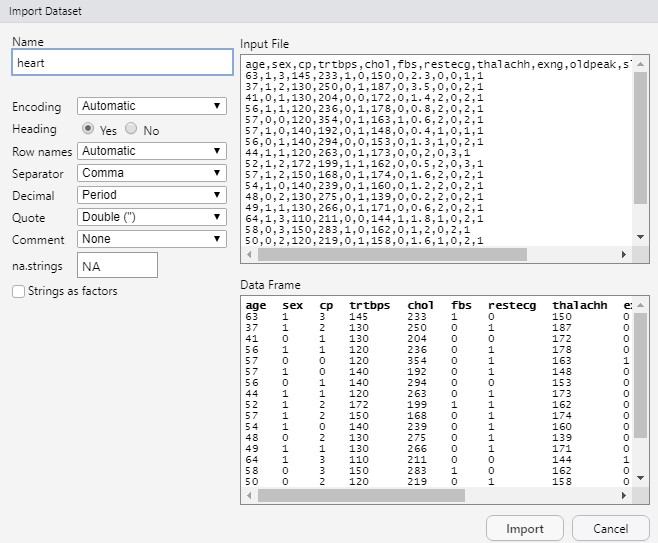
1. L’objectif du TP en donnant brièvement la description du tableau de données.
2. Faire quelques présentations graphiques.
3. Diviser le tableau de données en apprentissage/test.
4. Application de l'analyse discriminante sur l'échantillon d'apprentissage.
5. Prédiction des classes pour l'échantillon test.
6. Evaluation des résultats obtenus sur l'échantillon test.
7. Conclusion.

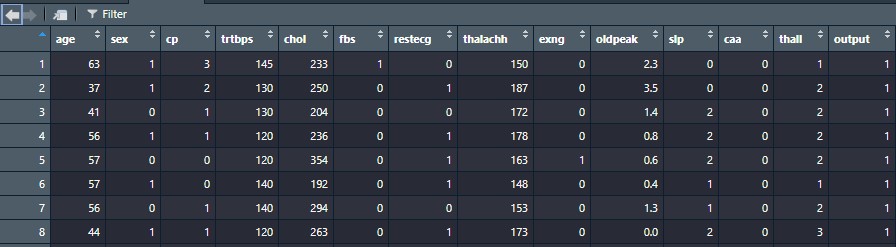
* **L’objectif du TP en donnant brièvement la description du tableau de données** :

Le TP a pour objectif de faire une prédiction concernant un contexte particulier, dont une personne donnée a problème cardio vasculaire ou pas, cette prédiction est faire sur un tableau de données « Heart » dont nous allons appliquer sur ce dernier une analyse factoriel discriminante ainsi qu’en exploitant un modèle d’apprentissage automatique de cette analyse dite LDA et QDA.

Pour se faire, nous devons d’abord importer notre tableau de donnees dans R studio afin d’exécuter le travail demandé. Pour se faire, il suffit de se placer dans l’onglet File ensuite Import Dataset et From Text (base)







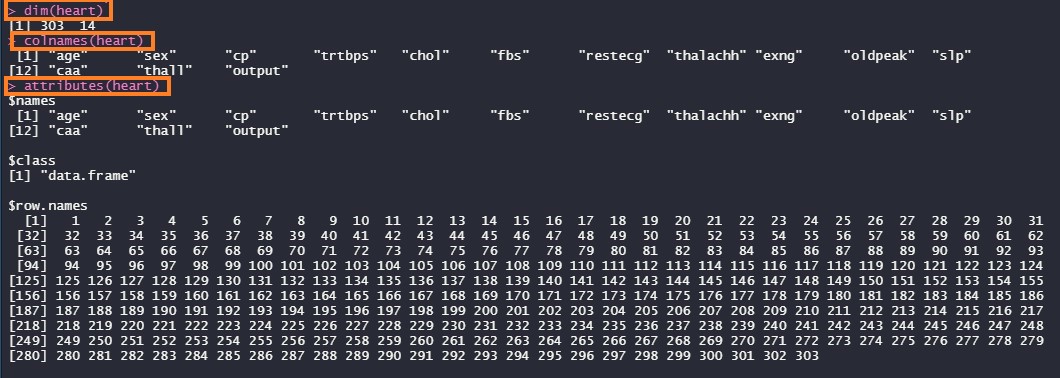
Afin de faire une description du tableau de données, nous allons utiliser les fonctions suivantes :

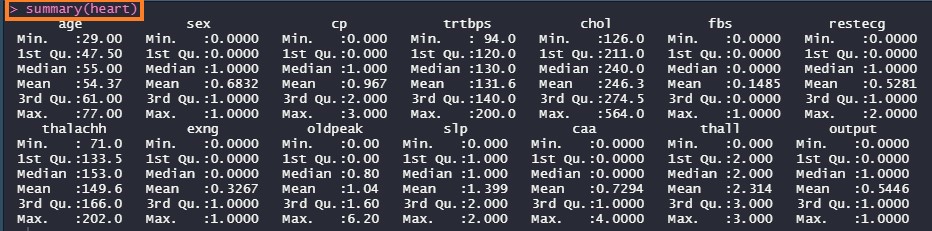
dim(heart)

colnames(heart)

attributes(heart)

summary(heart)



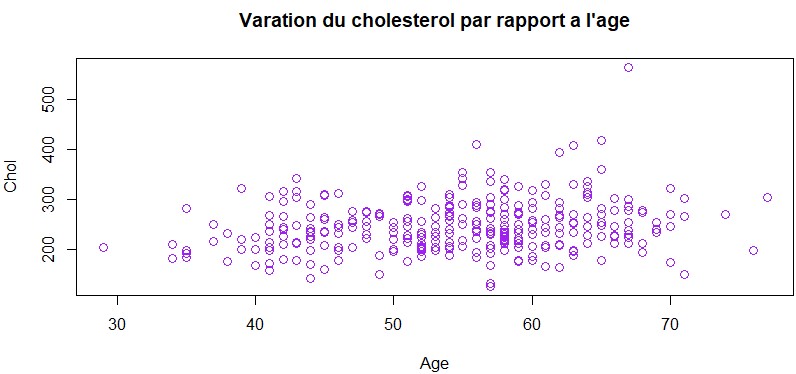


|  |  |
| --- | --- |
| **Fonction** | **Description** |
| Dim | La dimension du tableau de données |
| Colnames | Le nom des colonnes du tableau de donnees |
| Attributes | Les attributs du tableau de données |
| Summary | Une fonction qui retourne un tableau contenant des informations statistiques sur le jeu de données |
| Min | La valeur minimale de chaque variable |
| 1st QU | Le premier quartile |
| Median | La médiane de chaque variable |
| Mean | La moyenne de chaque variable |
| 3rd QU | Le 3eme quartile |
| Max | Le maximum de chaque variables |

* **Faire quelques présentations graphiques :**

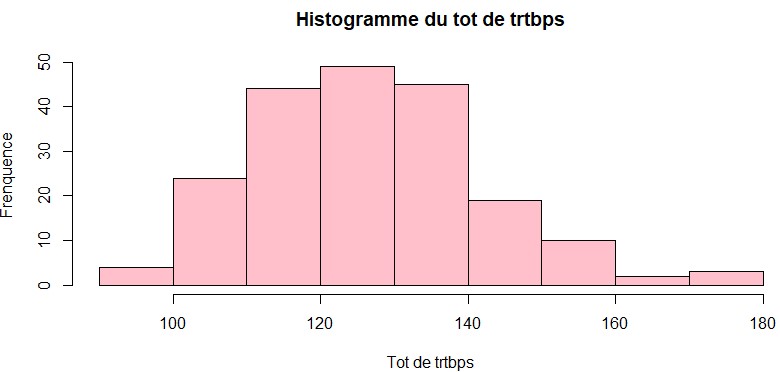
Dans cette partie, il nous a été demandé de faire des représentations graphiques (plots) de notre tableau de données

plot (heart$age, heart$chol, xlab = "Age", ylab = "Chol", col="purple", cex=1.2, main = "Varation du cholesterol par rapport a l'age")

**Représentation de la variation du tôt cholestérol par rapport à l’age**

Nous avons utilisé la fonction plot afin de représenter ce graphique qui décrit la variation du tôt de cholestérol selon l’age des individus de notre jeu de données. Le résultat s’affiche sous forme d’un nuage de points.

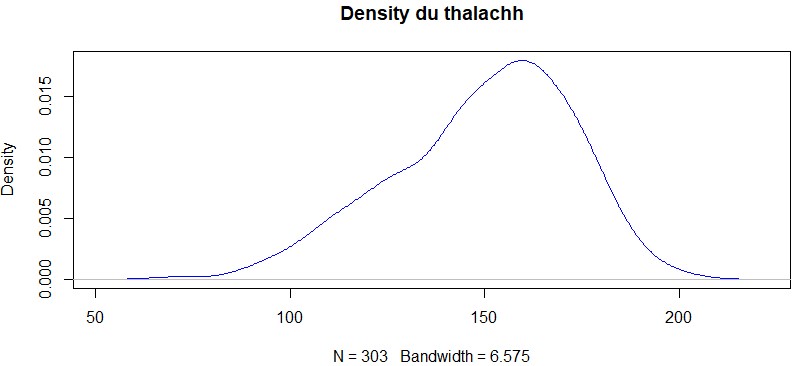
hist (heart$trtbps [1:200], xlab="Tot de trtbps”, ylab ="Frenquence», col="pink», main = "Histogramme du tot de trtbps")



**Représentation d’un histogramme du tôt de trtbps**

Nous avons utilisé la fonction hist afin de représenter ce graphique qui décrit la fréquence du tôt de trtbps de notre jeu de données. Le résultat s’affiche sous forme d’un histogramme.

plot(density(heart$thalachh), main = "Density du thalachh”, col="blue")



**Représentation de la densité du thalachh**

Nous avons utilisé la fonction plot dont nous luis avions passe la fonction Density comme paramètre afin d’avoir le plot qui représente la densité du thalachh dans notre dataset.

x=nrow(heart[heart$age>=29 & heart$age<=35,])

y=nrow(heart[heart$age>=36 & heart$age<=55,])

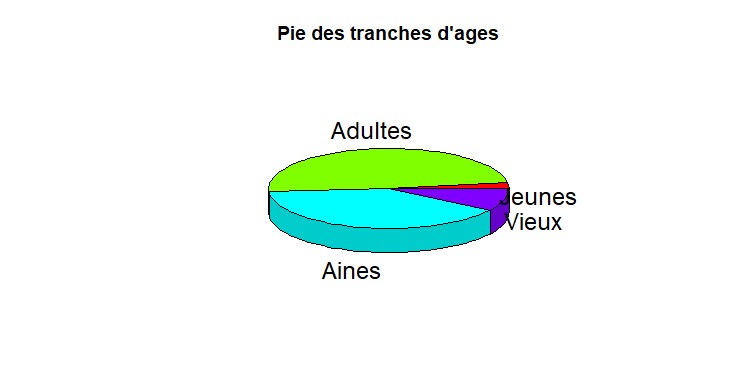
z=nrow(heart[heart$age>=56 & heart$age<=65,])

e=nrow(heart[heart$age>=66 & heart$age<=70,])

slices=c(x,y,z,e)

lbls <- c("Jeunes", "Adultes", "Aines", "Vieux")

pie3D(slices, labels = lbls, main="Pie des tranches d'ages")



**Représentation d’un diagramme en camembert des tranches d’ages**

Nous avons utilisé la fonction pie3D (disponible dans la package plotrix) qui représente un diagramme en camembert des tranches d’ages de notre jeu de données.

* **Diviser le tableau de données en apprentissage/test :**

Il nous a été demandé de diviser le tableau de données en sous sections notamment pour le modèle d’apprentissage et celui du test, pour se faire il faut procéder ainsi :

* **Données d’apprentissage** :

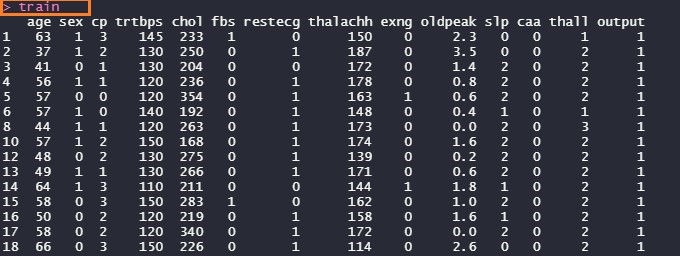
tr <- sample(nrow(heart), nrow(heart)\*0.8)

tr=sort(tr)

tr

train <-heart[tr,]

train

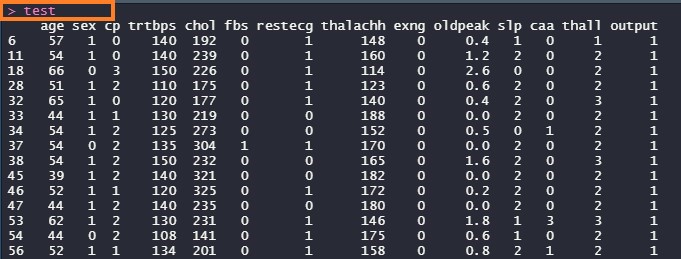


Nous remarquons à partir de la figure le résultat de train qui est un jeu de données de notre dataset initial, dont nous avons choisi d’utiliser 80% du Dataset pour le modèle d’entrainement.

* **Donnees test** :

test <-heart[-tr,]

test



Les données de test sont créées à partir du des données non affectées a train, autrement dit se sont les données qui existe dans notre Dataset mais pas dans train.

* **Application de l'analyse discriminante sur l'échantillon d'apprentissage :**

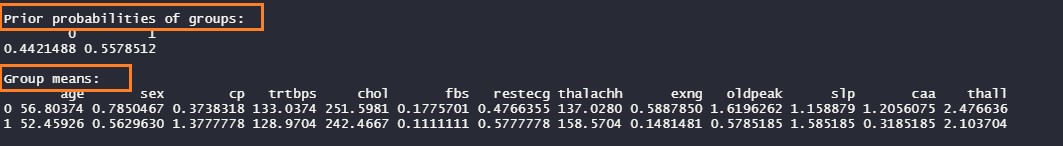
Il nous a été demandé d’appliquer une analyse discrimine sur l’échantillon d’apprentissage, dans R, nous avons deux fonctions qui nous permettent de faire cette analyse, LDA et QDA.

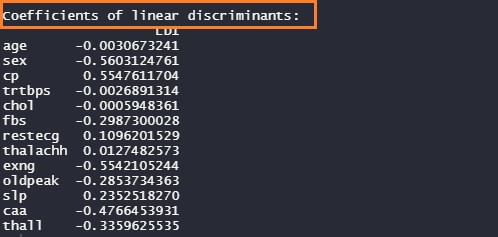
1. **LDA**:

Pour se faire, il suffit d’utiliser la fonction déjà prédéfinie LDA

res\_train\_lda=lda(output~age+sex+cp+trtbps+chol+fbs+restecg+thalachh+exng+oldpeak+slp+caa+thall,data=train)

res\_train\_lda





|  |  |
| --- | --- |
| **Sortie** | **Définition** |
| Prior | La probabilité d’appartenir à un groupe |
| Group means | La moyenne des variables dans chaque groupe |
| Coefficients of liner discriminants | Facteur discriminant |

Nous savons que l’AFD a pour objectif de trouver une combinaison libraire des p variables quantitatives qui permettent de mieux séparer les groupes. La première combinaison a pour objectif de maximiser la variance inter-groupe et donc minimiser la variance intra-groupe, ces combinaisons linaires définissent les variables discriminantes, nous savons également que le nombre de ces variables dépend du nombre de groupe de notre jeu de données, donc on ne peut créer que q-1 variables (q le nombre de groupe), dans notre cas on ne peut donc que créer 1 seules variables discriminantes.

Créons maintenant cette variable :

res\_train\_lda$scaling

u\_lda=res\_train\_lda$scaling

u\_lda=as.matrix(u\_lda)

dim(u\_lda)

g=colMeans(train[-14])

g=as.matrix(g)

x=as.matrix(train[-14])

t\_lda=x%\*%u\_lda

dim(t\_lda)

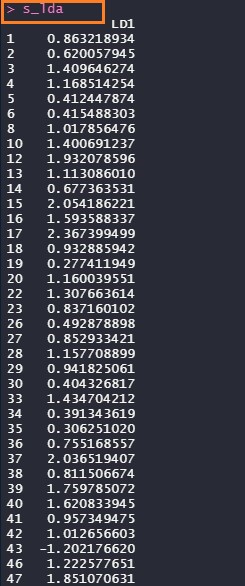
t2\_lda=t(u\_lda)%\*%g

v1\_lda=rep(t2\_lda,dim(train)[1])

v1\_lda=as.matrix(v1\_lda)

s\_lda=t\_lda-v1\_lda

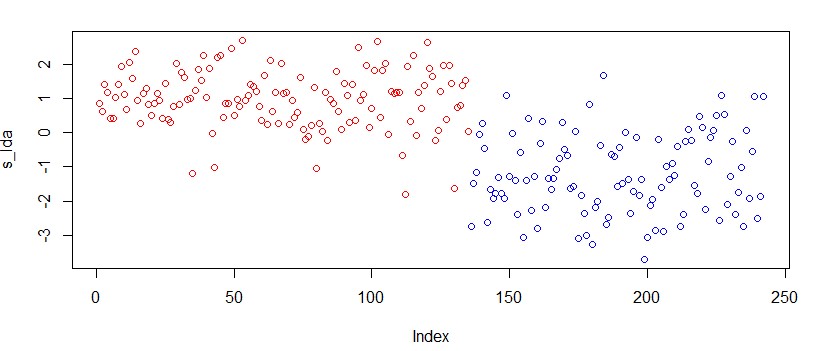
s\_lda



Nous pouvons également afficher la représentation du résultat de LDA sous forme graphique

couleur=c(rep('red',135),rep('blue',107))

plot(s\_lda,col=couleur)



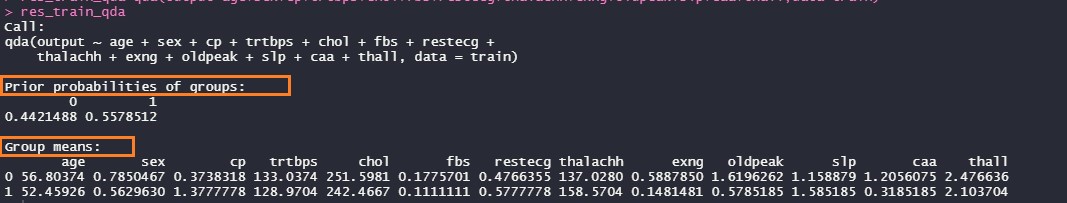
**Représentation du nuage des points des projections de la variable discriminante**

1. **QDA :**

Pour se faire, il suffit d’utiliser la fonction déjà prédéfinie QDA

res\_train\_qda=qda(output~age+sex+cp+trtbps+chol+fbs+restecg+thalachh+exng+oldpeak+slp+caa+thall,data=train)

res\_train\_qda



La 1ere case décrit la probabilité d’appartenance d’un nouvel individu a un des groupes, et le Group means décrit le centre de gravité des groupes dont la moyenne des variables.

* **Prédiction des classes pour l'échantillon test** :

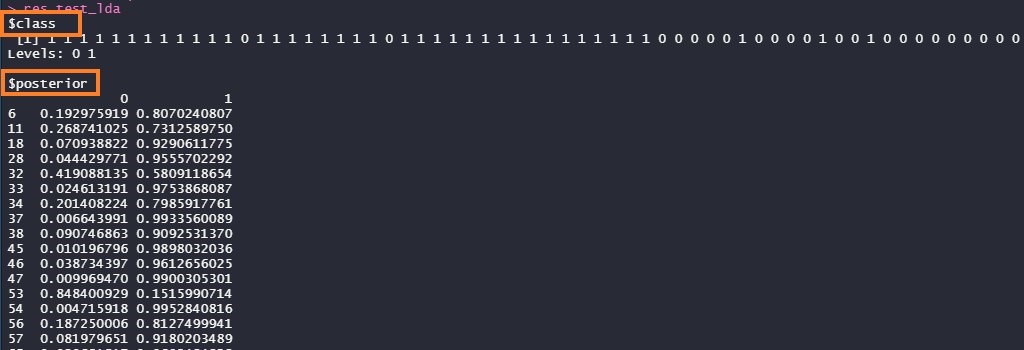
Une fois nous avons appliqué le modèle d’apprentissage pour notre échantillon d’apprentissage avec LDA et QDA, on peut maintenant appliquer une prédiction sur notre échantillon test afin de pouvoir prédire l’appartenance d’un nouvel individu a un des groupes déjà existants

* + **Prédiction pour le résultat de LDA** :

Pour se faire, nous allons utiliser la fonction predict qui va prendre comme paramètre le résultat de l’entrainement de LDA et le jeu de donnees test (déjà expliquée), et cela comme suit :

res\_test\_lda=predict(res\_train\_lda,newdata=test)

res\_test\_lda





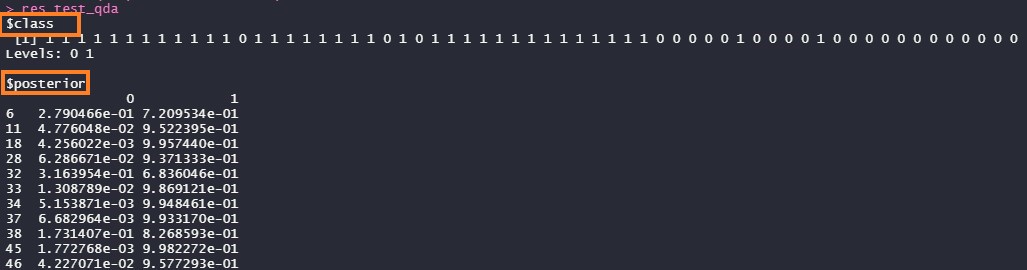
|  |  |
| --- | --- |
| **Attributs** | **Définition** |
| Class | L’appartenance d’un individu a une classe |
| Posterior | La probabilité d’appartenance a un groupe |
| X | La projection de l’individu sur l’axe discriminant |

* + **Prédiction pour le résultat de QDA** :

Pour se faire, nous allons utiliser la fonction predict qui va prendre comme paramètre le résultat de l’entrainement de QDA et le jeu de donnees test (déjà expliquée), et cela comme suit :

res\_test\_qda=predict(res\_train\_qda,newdata=test)

res\_test\_qda



|  |  |
| --- | --- |
| **Attributs** | **Définition** |
| Class | L’appartenance d’un individu a une classe |
| Posterior | La probabilité d’appartenance a un groupe |

* **Evaluation des résultats obtenus sur l'échantillon test** :

Une fois la prédiction a été faite sur l’échantillon test, nous devons évaluer cette prédiction, pour se faire, nous devons :

1. Calculer la matrice de confusion
2. Calculer le taux de précision
3. Calculer le taux d’erreur

* **Pour LDA** :

1. Matrice de confusion :

matrice\_confusion\_test\_lda=table (test [,14], res\_test\_lda$class)

matrice\_confusion\_test\_lda

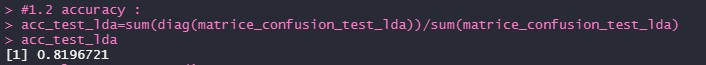


Cela signifie que pour le groupe 0, le modèle avait prédit 15 individus dans le groupe 0 (groupe réelle) et 6 dans le 1er groupe, idem pour le groupe 1, le modèle avait prédit 35 individus dans le 1er (groupe réelle) groupe mais 5 dans le groupe 0.

1. Taux de précision :

acc\_test\_lda=sum(diag(matrice\_confusion\_test\_lda))/sum(matrice\_confusion\_test\_lda)

acc\_test\_lda

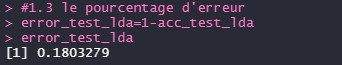


Cela signifie que le modèle a un taux de précision égale à 82%

1. Taux d’erreur :

error\_test\_lda=1-acc\_test\_lda

error\_test\_lda



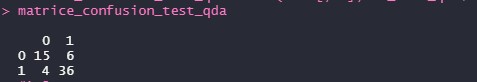
Cela signifie que le modèle a un taux d’erreur égale à 18%

* **Pour QDA** :

1. Matrice de confusion :

matrice\_confusion\_test\_qda=table (test [,14], res\_test\_qda$class)

matrice\_confusion\_test\_qda

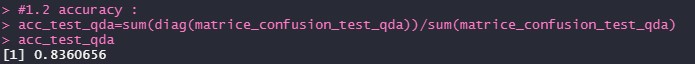


Cela signifie que pour le groupe 0, le modèle avait prédit 15 individus dans le groupe 0 (groupe réelle) et 6 dans le 1er groupe, idem pour le groupe 1, le modèle avait prédit 36 individus dans le 1er (groupe réelle) groupe mais 4 dans le groupe 0.

1. Taux de précision :

acc\_test\_qda=sum(diag(matrice\_confusion\_test\_qda))/sum(matrice\_confusion\_test\_qda)

acc\_test\_qda

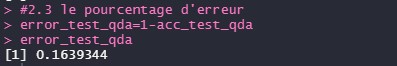


Cela signifie que le modèle a un taux de précision égale à 84%

3. Taux d’erreur :

error\_test\_qda=1-acc\_test\_qda

error\_test\_qda



Cela signifie que le modèle a un taux d’erreur égale à 16%

* **Conclusions :**

A l’issu de l’application d’une analyse factorielle discriminante sur notre jeu de données en exploitant les deux fonctions LDA et QDA., nous constatons les résultats suivants :

|  |  |  |
| --- | --- | --- |
|  | **Taux de précision** | **Taux d’erreur** |
| **LDA** | 82 % | 18 % |
| **QDA** | 84 % | 16 % |

Nous pouvons constater que le taux de précisions de QDA est supérieur a celui de LDA (84% >= 82%), d’une manière équivalente, le taux d’erreur de QDA sera donc inferieur a celui de LDA (16% <= 18%) et cela est dû probablement aux matrices de variance-covariance intra-groupe.

Nous avons le résultat affirmant que l’utilisation de la fonction LDA dans le cas ou les matrices de variance-covariance intra-groupe sont identiques donne de meilleur résultat que l’utilisation de la règle QDA, et donc dans le cas contraire, QDA donne de meilleurs résultats si les matrices ne sont pas identiques. Ce qui nous mène à admettre que les matrices de variance-covariance intra-groupe de notre jeu de données ne sont donc pas égales.

Nous pouvons donc aboutir au résultat suivant :

* Si les matrices de variance-covariance intra-groupe d’un je données sont identiques, alors il est préférable d’utiliser la règle LDA pour une meilleure précision.
* Nous utilisons la règle QDA dans le cas contraire, ou les matrices de variance-covariance ne sont pas identiques.

A l’issu de ce TP, nous avons manipulé un tableau de données en décrivant ce dernier ainsi que des présentations graphiques intéressantes, nous avons par la suite appliqué deux approches différentes d’analyse discriminante dont la règle LDA et QDA sur un échantillon de notre tableau de données. Les résultats de la comparaison entre ces deux approches nous en permis d’affirmer un résultat pertinent concernant le rapport entre les matrices de variance-covariance intra-groupe et le choix de l’approche a utiliser. (Cette conclusion a déjà été démontrée dans le cours).