REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L’ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ DES SCIENCES ET DE LA TECHNOLOGIE HOUARI BOUMEDIENE

**FACULTÉ DE MATHÉMATIQUE**

**DEPARTEMENT** *PROBABILITÉS ET STATISTIQUE*



**RAPPORT**

**MODULE : PROGRAMMATION R**

**SPECIALITE :** *STATISTIQUES ET PROBABILITÉS APPLIQUES*

**Analyse discriminant pour le regroupement des données IRIS**

**NOM:** FERGUOUS

**PRÉNOM:**Wafa

**MATRICULE:**171732073558

[**Introduction**](#_64xdiuz8noxr) **3**

[**Description des données**](#_7ldomvg3oxf8) **4**

[**Description de la méthode utilisé :**](#_a9enj7yymwpo) **4**

[**Implémentation de l’analyse discriminant linéaire sur Iris**](#_9rz6k3byctp6) **6**

[Installation des packages](#_tb85dbv0b6cc) 6

[chargement des packages](#_4niwg5839orh) 6

[chargement de données IRIS](#_6uqynqjgaq7q) 6

[**Analyser les données IRIS :**](#_c31th3nohkcz) **8**

[**Visualisation des données**](#_hsd5cgsdm5x6) **9**

[scatter plot](#_cy8lkp8mq7i2) 9

[Boxplot](#_szbjerd4f4cw) 10

[Correlation](#_kejzp52tsckz) 13

[**Échantillon d’apprentissage et échantillon test**](#_5kyy9wdbx4lw) **14**

[**Validation du modèle**](#_aeyua385f2rs) **17**

[**Visualisation des résultats**](#_sr9m39iaycm9) **18**

# Introduction

Supposons qu'un botaniste amateur souhaite distinguer les espèces de certaines fleurs d'iris qu'il a trouvées. Elle a collecté quelques mesures associées à chaque iris, qui sont :

-la longueur et la largeur des pétales

-la longueur et la largeur des sépales, toutes mesurées en centimètres.

Elle possède également les mesures de certains iris qui ont été préalablement identifiés par un botaniste expert comme appartenant aux espèces setosa, versicolor et virginica. Pour ces mesures, elle peut être certaine de l'espèce à laquelle appartient chaque iris. Nous considérons que ce sont les seules espèces que notre botaniste rencontrera. Le but est de créer un modèle de classification qui puisse prendre des mesures de ces iris dont les espèces sont déjà connues, afin que nous puissions prédire l'espèce pour les nouveaux iris qu'elle a trouvés.

# Description des données

Pour cette étude, on utilisera le célèbre jeu de données **IRIS.** Ce dernier est une base de données regroupant les caractéristiques de trois espèces de fleurs d’Iris, à savoir *Setosa*, *Versicolour* et *Virginica*. Chaque ligne de ce jeu de données est une observation des caractéristiques d’une fleur d’Iris. Ce *dataset* décrit les espèces d’Iris par quatre propriétés :

- longueur de sépales en cm

- largeur de sépales en cm

- longueur de pétales en cm

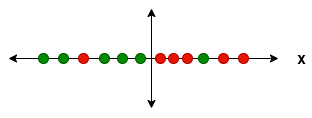
-largeur de pétales en cm

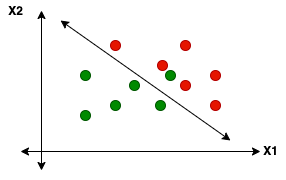
La base de données comporte 150 observations (50 observations par espèce) .

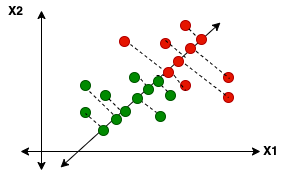
# Description de la méthode utilisé :

L'analyse discriminante linéaire est une méthode de classification supervisée et c'est l'un des outils les plus puissants dont dispose un analyste de données. similaire à ACP. Cependant, alors que ACP est un algorithme non supervisé qui se concentre sur la maximisation de la variance dans un ensemble de données, LDA est un algorithme supervisé qui maximise la séparabilité entre les classes.

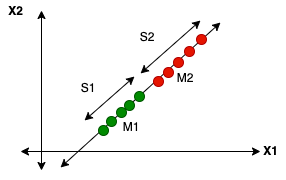
Supposons que nous ayons un ensemble de données avec deux colonnes - une variable explicative et une variable cible binaire (avec les valeurs 1 et 0). La distribution de la variable binaire est comme ci-dessous

Les points verts représentent 1 et les rouges représentent 0. Puisqu'il n'y a qu'une seule variable explicative, elle est notée par un axe (X). Cependant, si nous essayons de placer un diviseur linéaire pour délimiter les points de données, nous ne pourrons pas le faire avec succès car les points sont dispersés sur l'axe.

Il semble donc qu'une seule variable explicative ne soit pas suffisante pour prédire le résultat binaire. Nous allons donc apporter une autre caractéristique X2 et vérifier la distribution des points dans l'espace à 2 dimensions.

Il semble que dans l'espace à 2 dimensions la démarcation des sorties soit meilleure qu'avant. Cependant, augmenter les dimensions peut ne pas être une bonne idée dans un jeu de données qui a déjà plusieurs fonctionnalités. Dans ces situations, LDA vient à notre rescousse en minimisant les dimensions. Dans la figure ci-dessous, les classes cibles sont projetées sur un nouvel axe :

Les classes sont maintenant facilement délimitées. LDA transforme les caractéristiques d'origine en un nouvel axe, appelé discriminant linéaire (LD), réduisant ainsi les dimensions et assurant une séparabilité maximale des classes.



# 

# Implémentation de l’analyse discriminant linéaire sur Iris

## Installation des packages

install.packages("ggplot2",dependencies=TRUE, repos='http://cran.rstudio.com/')

install.packages("GGally",dependencies=TRUE, repos='http://cran.rstudio.com/')

install.packages("MASS",dependencies=TRUE, repos='http://cran.rstudio.com/')

install.packages("corrplot",dependencies=TRUE, repos='http://cran.rstudio.com/')

install.packages("caret",dependencies=TRUE, repos='http://cran.rstudio.com/')

install.packages("vegan",dependencies=TRUE, repos='<http://cran.rstudio.com/>')

chargement des packages

library(ggplot2)

library(GGally)

library(MASS)

library(corrplot)

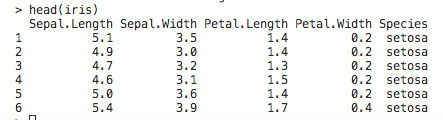
library(caret)

library(vegan)

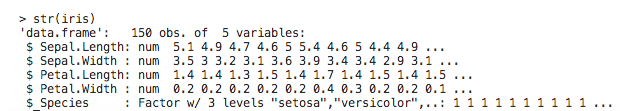
## chargement de données IRIS

data(iris)

head(iris)



str(iris)

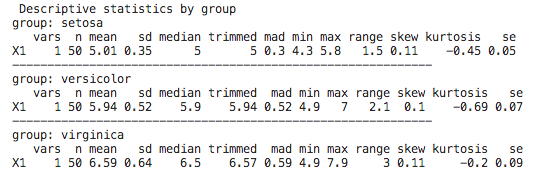


Nous pouvons voir que l'ensemble de données contient 5 variables et 150 observations au total.

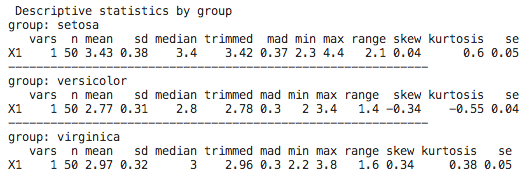
# Analyse descriptive pour les données IRIS :

Nous avons besoin de connaître plus d'informations sur les colonnes comme le nom de la colonne, le nombre de valeurs non nulles dans chaque colonne, le type de données des données, etc. Nous pouvons vérifier cela en utilisant la fonction describeBy .

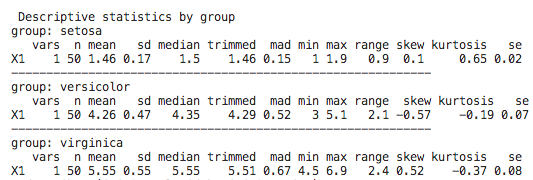
describeBy(iris$Sepal.Length, iris$Species)



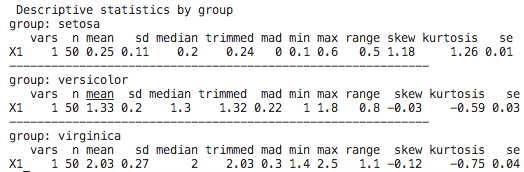
describeBy(iris$Sepal.Width, iris$Species)



describeBy(iris$Petal.Length, iris$Species)



describeBy(iris$Petal.Width, iris$Species)



À partir de cette description, nous pouvons voir toutes les descriptions des données, comme la longueur et la largeur moyennes, la valeur minimale, la valeur maximale, la valeur de distribution de 25 %, 50 % et 75 %, etc.

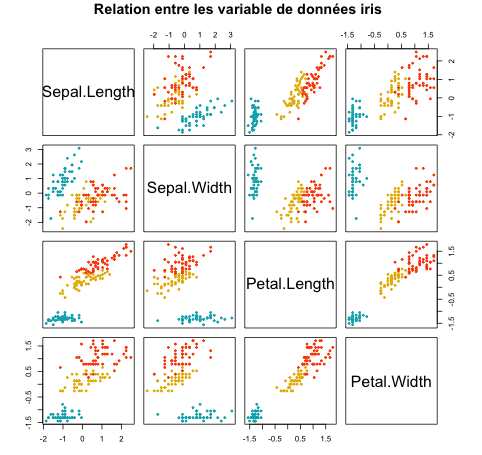
# Visualisation des données

## scatter plot

my\_cols <- c("#00AFBB", "#E7B800", "#FC4E07")

pairs(iris[,1:4], pch = 19, cex = 0.5, main = "Relation entre les variable de données iris ",

col = my\_cols[iris$Species])



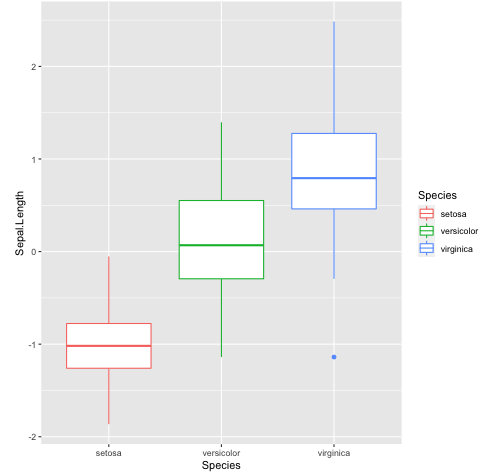
Nous pouvons voir que Setosa est très bien séparé que Versicolor et Virginica. En utilisant sepal.length et sepal.width, nous pouvons distinguer les fleurs de Setosa des autres. Séparer versicolor et virginica est un peu plus difficile.

## Boxplot

ggplot(data = iris) +

aes(x = Species, y = Sepal.Length, color = Species) +

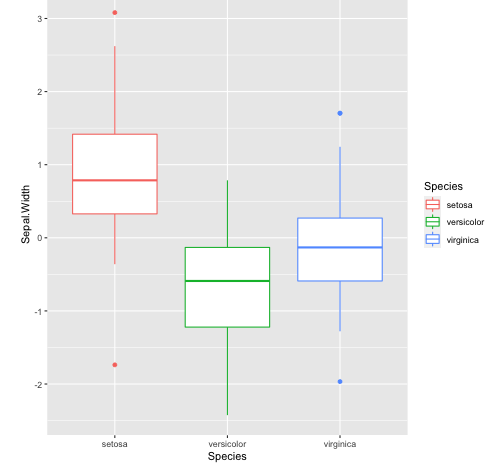
geom\_boxplot()



ggplot(data = iris) +

aes(x = Species, y = Sepal.Width, color = Species) +

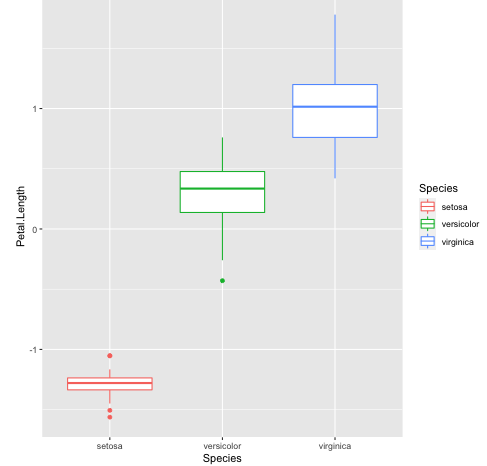
geom\_boxplot()



ggplot(data = iris) +

aes(x = Species, y = Petal.Length, color = Species) +

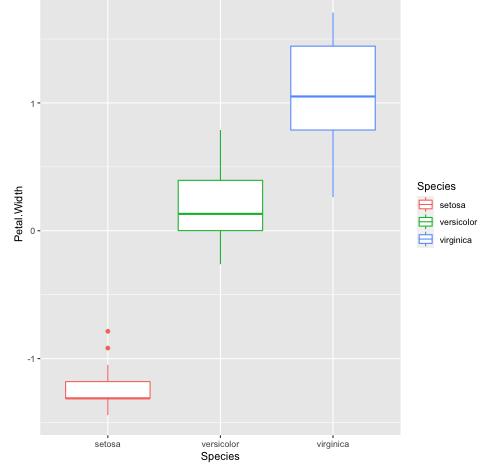
geom\_boxplot()



ggplot(data = iris) +

aes(x = Species, y = Petal.Width, color = Species) +

geom\_boxplot()



D'après les boxplots ci-dessus, nous pouvons dire

-Les espèces de Setosa ont des valeurs aberrantes dans la longueur et la largeur des pétales (c'est-à-dire des points de données qui considèrent différent de la majorité des données extraites d'un échantillon ou d'une population)

-Les espèces de Setosa ont des caractéristiques (dimensions) plus petites et sont moins distribuées

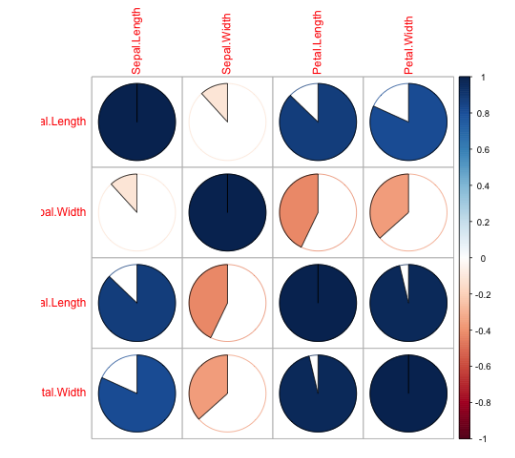
-Les espèces Versicolor sont réparties de manière moyenne et de dimensions moyennes.

-Les espèces de Virginica sont fortement distribuées avec un grand nombre de valeurs et de caractéristiques (dimensions).

## Correlation

cr <- cor(iris[1:4])

corrplot(cr,method="pie")



D'après le graph , nous voyons que la longueur et la largeur des pétales sont fortement corrélées, ainsi que la largeur des pétales et la longueur des sépales ont une bonne corrélation.

conclusion :

-L'ensemble de données est équilibré, c'est-à-dire que des enregistrements égaux sont présents pour les trois espèces.

- Nous avons quatre colonnes numériques tandis qu'une seule colonne catégorique qui à son tour est notre colonne cible.

-Une forte corrélation est présente entre la largeur des pétales et la longueur des pétales.

- L'espèce setosa est la plus facile à distinguer en raison de sa petite taille. -Les espèces Versicolor et Virginica sont généralement mélangées et sont parfois difficiles à séparer, tandis que Versicolor a généralement des tailles de caractéristiques moyennes et Virginica a des tailles de caractéristiques plus grandes .

L'une des hypothèses clés de l'analyse discriminante linéaire est que chacune des variables prédictives a la même variance. Un moyen simple de s'assurer que cette hypothèse est respectée consiste à mettre à l'échelle chaque variable de sorte qu'elle ait une moyenne de 0 et un écart type de 1.

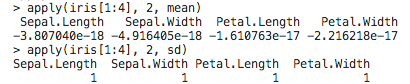
scale les données:

iris[1:4] <- scale(iris[1:4])

On utilise la fonction apply() pour vérifier que chaque variable prédictive a maintenant une moyenne de 0 et un écart type de 1

apply(iris[1:4], 2, mean)

apply(iris[1:4], 2, sd)



**Valeurs NA**

****

Il n'y a pas de valeur NA.

# Échantillon d’apprentissage et échantillon test

Par tirage aléatoire, Nous choisissons un échantillon pour appliquer l’AD, elle représente 90% de l'échantillon total ( sur lequel, on estimer ́e la fonction linéaire discriminante. Le reste 10% des observations est considéré comme échantillon test réserver pour la validation.

set.seed(1)

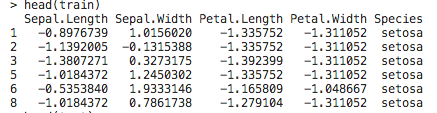
sample <- sample(c(TRUE, FALSE), nrow(iris), replace=TRUE, prob=c(0.9,0.1))

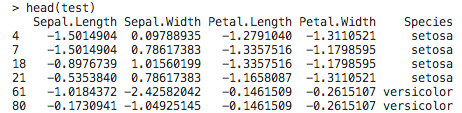
train <- iris[sample, ]

test <- iris[!sample, ]

head(train)

head(test)

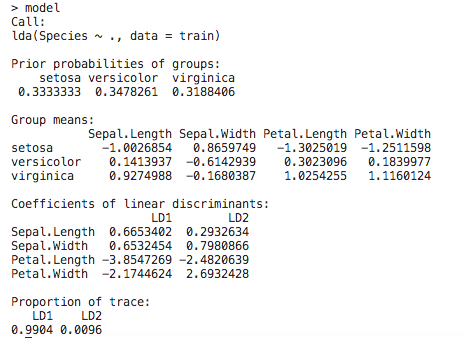




Maintenant, nous effectuons LDA sur les données d’iris:

model <- lda(Species~., data=train)

model



**interprétation des résultats**

**Prior probabilities of groups:** elles représentent les proportions de chaque espèce dans l'échantillon d'apprentissage. Par exemple, 35,8 % de toutes les observations de l'échantillon d'apprentissage concernent l'espèce virginica.

**Coefficients of linear discriminants :** ils affichent la combinaison linéaire des variables prédictives utilisées pour former la règle de décision du modèle LDA.

Par exemple:

* **LD1:** 0.792\*Sepal.Length + 0.571\*Sepal.Width – 4.076\*Petal.Length – 2.06\*Petal.Width
* **LD2:** 0.529\*Sepal.Length + 0.713\*Sepal.Width – 2.731\*Petal.Length + 2.63\*Petal.Width

**Proportion of trace:** le pourcentage de séparation atteint par chaque fonction discriminante linéaire.

# Prediction

predicted <- predict(model, test)

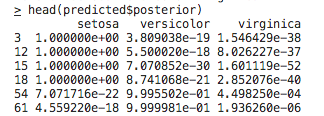
head(predicted$class)

head(predicted$posterior)

head(predicted$x)



la class prévue pour chaque observation



matrice dont les colonnes sont les groupes, les lignes sont les individus et les valeurs sont la probabilité a posteriori que l'observation correspondante appartient aux groupes.

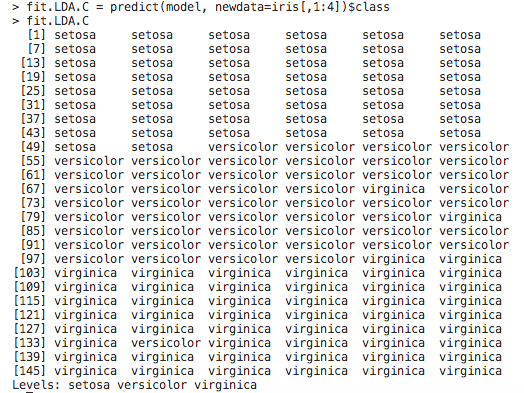
# Validation du modèle

L’une des méthodes pour valider un classificateur c’est l'étude de la table de confusion , consiste à comparer les valeurs observées de la variable dépendante ( classe originale ) avec la valeur prédite .

Les affectations des individus sont les suivantes :

fit.LDA.C = predict(model, newdata=iris[,1:4])$class

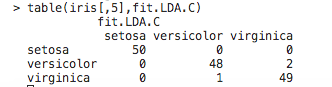
fit.LDA.C



Pour s’assurer que la fonction discriminante classifie bien les individus en sous-groupes, on analyse la matrice de confusion qui regroupe les individus bien classés et les mal classés .

La matrice de confusion de notre fonction score se présente comme suit :

table(iris[,5],fit.LDA.C)



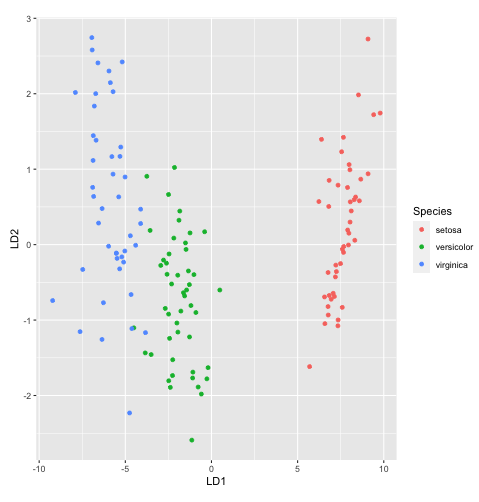
Seules 3 mesures ont été mal classées

# Visualisation des résultats

lda\_plot <- cbind(train, predict(model)$x)

ggplot(lda\_plot, aes(LD1, LD2)) +

geom\_point(aes(color = Species))



On voit bien comment les axes discriminants séparent les trois espèces différentes dans notre ensemble.

le 1er axe discriminant est le meilleur axe pour faire la discrimination.