1. EXECUTIVE SUMMARY

El objetivo de este trabajo es realizar un análisis discrimiante por cada tipología de planta en función de sus características, para ello trabajaremos con la anchura y longuitud del pétalo y del sépalo, y las clasificaremos según el análisis LDA y QDA en función de su grupo de pertencia, ya sea “setosa”, “versicolor” o “virginica”. Para ello realizaremos realizaremos una partición de las observaciones, trabajando con una muestra que contendrá el 60% del total y con el test que contendrá el 40% del total de las 150 observaciones. A partir de aquí realizaremos dos análisis, el LDA y el QDA.

En el primero de todos, evaluaremos la capacidad explicativa del LD1 y LD2 en donde veremos que el LD1, con un criterio del 0.9932 los grupos por especies quedan bien establecidos y veremos como la versicolor y la virginica están muy relacionadas entre sí, a diferencia de la setosa, realizando seguidamente un análisis de partición mediante el cual comprobaremos que la anchura del pétalo y la longuitud del sépalo establecen el mejor criterio de clasificación por especies, debido a que visualmente existe menor solapamiento. Finalmente a través de las observaciones iniciales realizaremos una predición tanto de la muestra como del test en donde asignaremos una clasificación en función del grupo al que mejor pertenzcan las plantas, de las cuales en setosa le corrsponderán 20 plantas, versicolor se le asignarán 19 y a la especie virginica 20, de la predcción para el test.

Por otra parte, a través del análisis QDA veremos que también existe menos solapamiento de clasificación por especies en función de la anchura del pétalo y la longuitud del sépalo y veremos como la predición para el test arroja el mismo resultado de clasificación para la setosa, para la especie versicolor se le asignarán 16 plantas y para virginica 19.

2. INTRODUCCIÓN

Las técnicas de análisis discrimiante tienen por objetivo la determinanción de un criterio que nos permita decidir a qué grupo pertence un cierto individuo, a partir de la información disponible. Los grupos ya están constituidos, siendo 3 su tipología (“setosa”, “virginica” y “versicolor”) y lo que buscamos es descubrir qué tiene de específico cada especie de planta para ser capaces de asignar de manera correcta las plantas a cada especie. Se trata, en definitiva, de discriminar a qué grupo pertenece cada planta.

3. ANÁLISIS EXPLORATORIO DE DATOS

Nos cargamos las librerías "tidyverse" que entre sus muchas funciones la utilizaremos para la manipulación de los datos y su posterior visualziación, "MASS" que es una librería utilizada para realizar el análisis discriminante LDA y QDA,"klaR" para poder trabajar con los datos y "ggpubr".

Nos cargamos el dataframe en donde tenemos 150 observaciones y 5 variables, que representan tanto la longuitud como el ancho del pétalo y del sépalo, así como el tipo de especie que podrá ser "setosa", "virgínica" o "versicolor".

Realizando un análisis de correlaciones entre las características de las plantas observamos una fuerte correlación entre la longuitud del pétalo así como su altura (0.96). así como una estrecha relación entre la longuitud del sépalo y la anchura del pétalo (0.87).

```
> cor(iris$Sepal.Length, iris$Petal.Width)
[1] 0.8179411
> cor(iris$Sepal.Length, iris$Sepal.Width)
[1] -0.1175698
> cor(iris$Petal.Length, iris$Sepal.Width)
[1] -0.4284401
> cor(iris$Petal.Length, iris$Petal.Width)
[1] 0.9628654
```

4. PREPARACIÓN DE LOS DATOS

Tendremos que crearnos una semilla (123) para que cada vez que se carguen los datos, estos no den diferentes resultados cada vez y clasificamos el conjunto de las 150 observaciones en dos subconjuntos que serán de entrenamiento o muestrales, las cuales las hemos llamado en R como "sample" y de los de test, este último coincide con la terminología empleada en R. El primero de todos contendrá el 60% , las cuales conforman un total de 89 observaciones y el último el 40%, las cuales contienen un total de 61 observaciones.

5. ANÁLISIS LINEAL DISCRIMINANTE (LDA)

El análisis discriminante es una técnica estádistica que identifica las variables que permiten diferenciar los grupos y cuántas de esas variables son necesarias para alcanzar la mejor clasificación posible. Otro de los objetivos de este análisis es encontrar la combinanción lineal de las variables independientes que mejor permitan diferenciar a los grupos. Se trata de un análisis multivariante que es capaz de aprovechar las relaciones existentes entre una gran cantidad de variables independientes con el objetivo de maximizar la capacidad de discriminación. A través del LDA podemos dividir el espacio

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

muestral en varios subgrupos mediante hiperplanos que permiten separar lo mejor posible los grupos objeto de estudio.

Los supuestos básicos para este análisis son normalidad multivariada e igualdad de matrices de covarianzas entre los grupos.

Realizamos un análisis lineal discriminante sobre las características de las plantas, tanto del sépalo como del pétalo, en su longuitud y anchura, respectivamente, para el "sample".

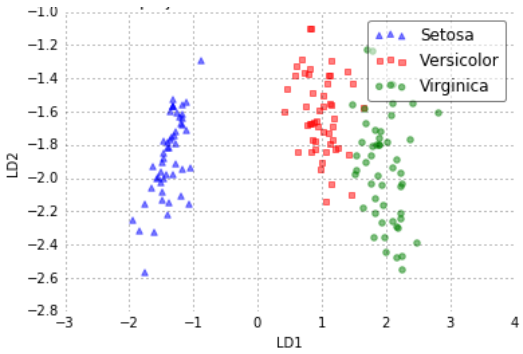
```
Prior probabilities of groups:
    setosa versicolor  virginica
 0.3370787  0.3370787  0.3258427
```

```
Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.946667    3.380000     1.443333    0.250000
versicolor     5.943333    2.803333     4.240000    1.316667
virginica      6.527586    2.920690     5.489655    2.048276

Coefficients of linear discriminants:
                    LD1         LD2
Sepal.Length  0.3629008  0.05215114
Sepal.Width   2.2276982  1.47580354
Petal.Length -1.7854533 -1.60918547
Petal.Width  -3.9745504  4.10534268

Proportion of trace:
   LD1    LD2
0.9932 0.0068
```

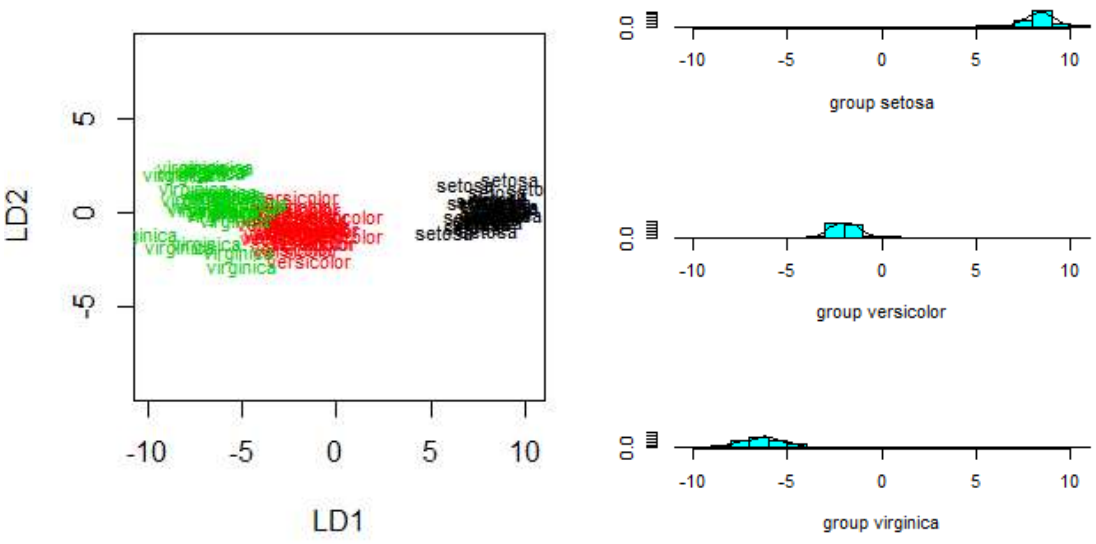
LD1 y LD2 son los coeficientes de la función de discrimiandción que permite diferenciar a las especies. Como hay 3 especies, aparecen dos funciones discriminantes



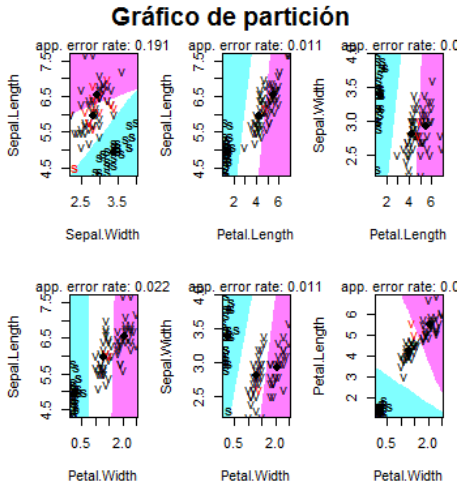
La proporción del LD1 es cuantitativamente muy superior al LD2 (0.9932 frente al 0.0068), esto quiere decir que a través del LD1 los grupos por especies están muy bien definidos, al contrario que la información representada a través del LD2.

A través de un gráfico podemos ver que la virginica y la versicolor están más cercanas la una con la otra. En cambio, la setosa está considerablemente más alejada del resto de las especies. Podemos sacar estas conclusiones a través de un gráfico conjunto de las especies o a través de un análisis individualizado, este último representado en la derecha.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu



Por otra parte, a través del comando partimat hacemos un gráfico de partición mediante el cual se relacionan cada una de las características propias de las plantas en función de su especie



Observamos que de los 6 diferentes gráficos, la relación más fuerte por criterio de clasificación es la relación existente entre la anchura del pétalo y la longitud del sépalo, en donde los grupos parecen están bien definidos y en donde existe menos solapamiento

El objetivo esencial es utilizar los valores previamente conocidos de las variables independientes para predecir en qué categoría de la variable dependiente corresponde. Es decir, asignar nuevos individuos al grupo que mejor corresponde a una clasificación ya establecida , construida a partir de las 3 clases de plantas que tenemos.

A través de la muestra se observa que tanto la setosa como la versicolor arrojan los misma clasificación, 30 plantas para cada grupo correspondiente. En cambio la virginica es el menor grupo de los 3.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

```
> table(train$lda,train$Species)

           setosa versicolor virginica
  setosa       30          0         0
  versicolor    0         30         0
  virginica     0          0        29
```

```
table(test$lda,test$Species)

           setosa versicolor virginica
setosa         20          0         0
versicolor      0         19         1
virginica       0          1        20
```

A través del test vemos que tanto la setosa como la virginica realizan la misma clasificación, en cambio con versicolor la clasificación es ligeramente inferior

6. ANÁLISIS CUADRÁTICO DISCRIMINANTE (QDA)

En términos generales LDA tiende a conseguir mejores clasificaciones que QDA cuando hay pocas observaciones con las que entrenar el modelo. Por el contrario, cuando trabajamos con una gran cantidad de observaciones de entrenamiento o si no es asumible que exista una matriz de covarianza común entre clases, QDA es más adecuado.

De nuestras observaciones de entrenamiento, las cuales son de 89 observaciones totales vamos a realizar el QDA para ver qué resultados son los mejores de cara a realizar la clasificación por especie de planta.

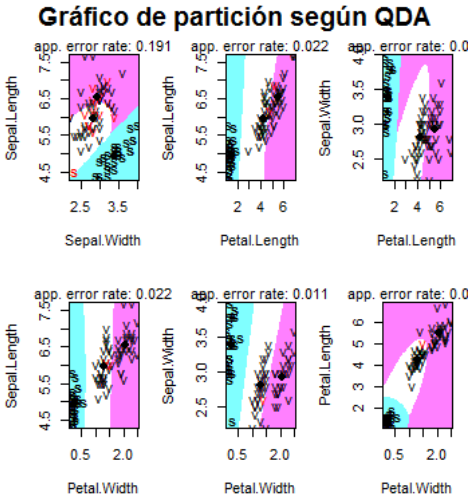
```
> qda.iris #show results
Call:
qda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
    data = train)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3370787  0.3370787  0.3258427

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa         4.946667    3.380000     1.443333    0.250000
versicolor     5.943333    2.803333     4.240000    1.316667
virginica      6.527586    2.920690     5.489655    2.048276
```


CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu

Observamos, al igual que en el gráfico de partición LDA, que de los 6 diferentes gráficos, la relación más fuerte por criterio de clasificación es la relación existente entre la anchura del pétalo y la longuitud del sépalo, en donde los grupos parecen están bien definidos y en donde existe menos solapamiento




```
table(train$qda,train$Species)

           setosa versicolor virginica
setosa         30          0         0
versicolor      0         30         0
virginica       0          0        29
```

```
table(test$qda,test$Species)

           setosa versicolor virginica
setosa         20          0         0
versicolor      0         16         2
virginica       0          4        19
```

Realizamos la predicción para nuestra muestra y para nuestro test y los resultados para la muestra son exactamente iguales que los resultados arrojados mediante el LDA. Sin embargo, realizando la predicción para el test los resultados son diferentes, en donde la mayor es la setosa por especie de planta, seguida de la virginica y finalmente la versicolor.

7. CONCLUSIONES

A través del análisis discrimiante hemos podido clasificar los diferentes tipos de observaciones en grupos. Las características de las plantas nos han proporcionado información relevante de cara a realizar la clasificación y la evaluación del poder discriminante de cada una de las especies.

Concluimos que a través del LDA como del QDA, con criterios de clasificación lineales y cuadráticos, respectivamente hemos tomado la decisión en cuanto al grupo en el que se clasifica cada planta.

CUNEF
Colegio Universitario de Estudios Financieros
Adscrito a la Universidad Complutense de Madrid
Leonardo Prieto 10
28040 Madrid
Tel. +34 91 448 08 92
Fax.+34 91 594 13 66
www.cunef.edu