Executive summary

El objetivo de este informe es exponer un análisis discriminante sobre el dataset Iris el cual tiene por objetivo:

- Describir características específicas para la distinción de los grupos, por medio de funciones lineales conocidas como funciones discriminantes. (Análisis discriminante descriptivo)
- Clasificar casos -individuos- en grupos preexistentes según las similitudes entre el caso y los casos pertenecientes los grupos, mediante funciones lineales o cuadráticas que reciben el nombre de funciones clasificadoras (Análisis discriminantes predictivo)

El análisis descriptivo incluye identificar la contribución relativa de p variables a la separación de los grupos y encontrar un plano óptimo donde los puntos puedan ser proyectados ilustrando de la mejor manera la separación de los grupos.

Vamos a trabajar con una muestra que contendrá el 60% del total y una muestra test que tendrá el 40% del total de las 150 observaciones. A partir de aquí se proseguirá con un análisis LDA y QDA. En el primero se evaluará la capacidad explicativa del LD1 y LD2 donde veremos que el primero tiene un criterio del 0.9932 donde los grupos por especies quedan bien establecidos y veremos que la versicolor y virgínica están muy relacionadas entre sí a diferencia de la setosa. A través de un análisis de partición se establece que la anchura del pétalo y la longitud del sépalo establecen el mejor criterio de clasificación por especies dado el menor solapamiento.

Finalmente, la idea es hacer una predicción sobre el training y el test set donde se clasifican las variables según su mejor pertenencia a cada grupo.

Por otra parte, a través del análisis QDA veremos que también existe menos solapamiento en la clasificación por especies y veremos que la predicción para la setosa es la misma, pero para la versicolor cambiará.

Análisis Exploratorio de Datos

Antes de realizar el análisis discriminante debemos hacer un análisis exploratorio de los datos. Como ya hemos comentado usaremos el dataset original que viene precargado en R, donde las variables estarán en inglés. A partir de ahí haremos carga de las librerías que nos ayudarán con el análisis discriminante como *klaR* y *ggpubr*.

Nuestro dataset original tiene 150 observaciones de 5 variables que representan la longitud y el ancho del pétalo y sépalo de las 3 especies.

Procediendo a realizar la matriz de correlaciones tenemos:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Sepal.Length	1.0000000	-0.1175698	0.8717538	0.8179411
Sepal.Width	-0.1175698	1.0000000	-0.4284401	-0.3661259
Petal.Length	0.8717538	-0.4284401	1.0000000	0.9628654
Petal.Width	0.8179411	-0.3661259	0.9628654	1.0000000

Fuente: Elaboración propia

La mayor correlación puede observarse entre el ancho y largo del pétalo seguida por el largo del pétalo junto con el del sépalo.

Análisis discriminante

Se procede a la extracción de una muestra de datos sobre el conjunto completo para hacer los dataframe de entrenamiento y prueba con unos porcentajes de 60% y 40% respectivamente.

Haciendo el análisis LDA de la librería *MASS* vamos a proceder a realizar el análisis discriminante, el cuál nos arroja los siguientes resultados:

- Las probabilidades de que el individuo descrito por el vector pertenezca a la clase en cuestión.

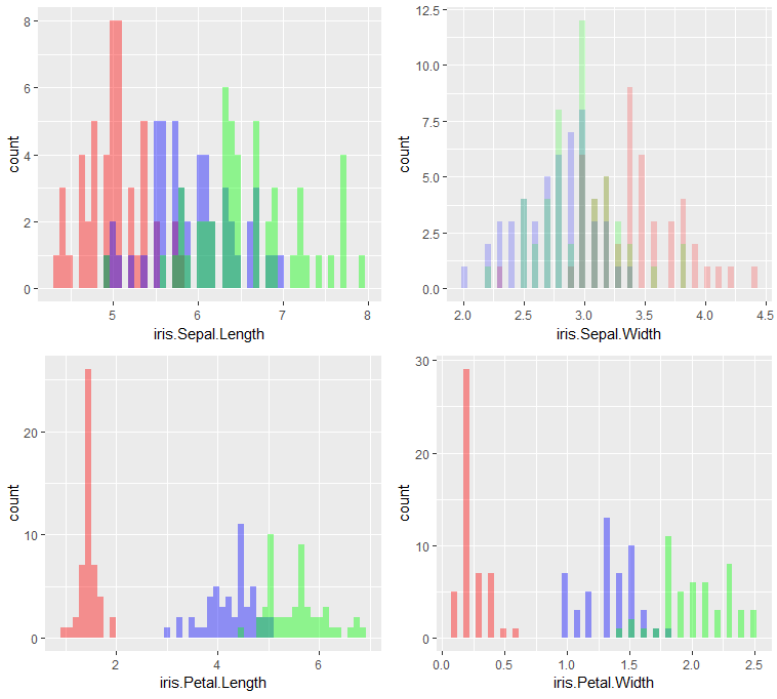
Setosa	Versicolor	Virginica
0.3370787	0.3370787	0.3258427

- Las medias del grupo
- Pesos discriminantes para las variables
    - LD1: 0.36 + 2.22 -1.78 -3.97
    - LD2: 0.05 + 1.47 -1.60 + 4.10

La variable que tiene un peso más grande y por tanto discrimina mejor es el ancho del pétalo en LD1 y LD2. Los coeficientes sirven para decidir a que clase pertenece cada ejemplar de la flor. Al ser LD1 mucho mayor que LD2 y prácticamente uno, esto confirma que las flores las podemos clasificar muy bien utilizando solamente un eje discriminante.

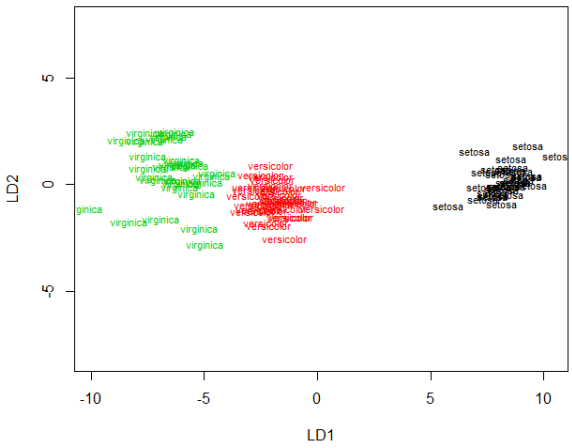
- Proporción de traza LD1: 0.9932
- Proporción de traza LD2: 0.0068

En las siguientes gráficas se puede apreciar la variable que mejor discrimina como hemos dicho anteriormente:

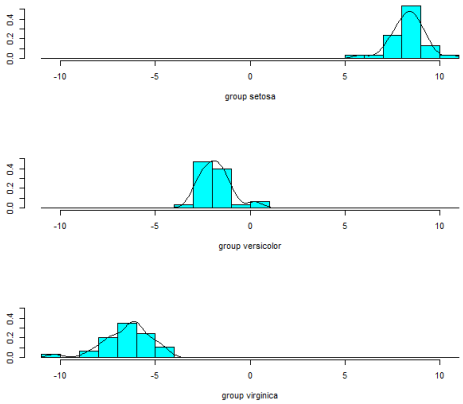


Fuente: Elaboración propia

El siguiente gráfico nos muestra como hacer la clasificación grupal para dos dimensiones donde el eje X es LD1 y el eje Y es LD2 y como podemos ver la diferenciación es perfecta.



Fuente: Elaboración propia

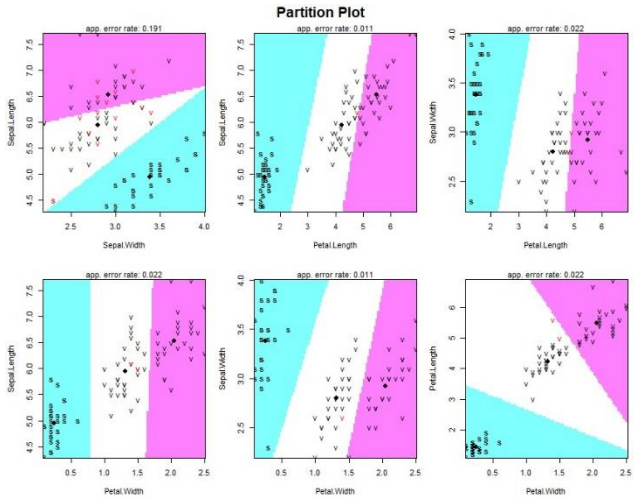


Como con LD1 podemos explicar toda la información debido a la proporción de traza vamos a representar las especies en una sola dimensión la que corresponde a LD1 a través de un histograma.

Fuente: Elaboración propia

Como se puede observar, los tres grupos se diferencian perfecta y aparentemente se distribuyen con normalidad y la que mejor se diferencia es la setosa.

La función *partimat* proporciona una forma alternativa de trazar las funciones discriminantes lineales. Muestra una serie de gráficos para cada combinación de dos variables. Cada gráfico es una vista diferente de los distintos datos, las partes coloreadas denotan cada área de clasificación. Se dice que prácticamente cualquier observación que este dentro de una región será de una clase concreta. De la misma manera cada grafico incluye la tasa de error para esa vista de los datos.



Como podemos observar en este gráfico el grupo mejor diferenciado es la setosa. Observando la función LDA las medias de los grupos, los valores que mas difieren unos con otros es el pétalo de la setosa tanto la longitud como el ancho como se ha especificado anteriormente.

Ahora realizaremos la predicción con la parte de entrenamiento, la suma de la diagonal principal es la suma de las observaciones por lo tanto la predicción es buena:

	Setosa	Versicolor	Virginica
Setosa	30	0	0
Versicolor	0	30	0
Virginica	0	0	29

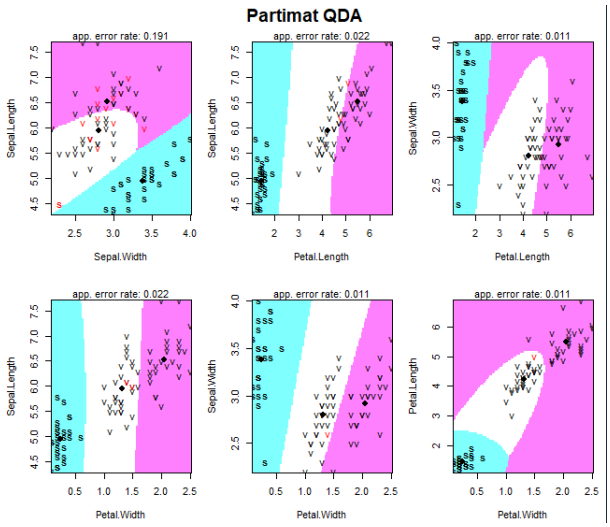
Haciendo lo mismo para la parte de test obtenemos los siguientes datos:

	Setosa	Versicolor	Virginica
Setosa	20	0	0
Versicolor	0	19	1
Virginica	0	1	20

En este caso la clasificación no es lo más acertada posible ya que tenemos errores fuera de la diagonal principal.

Ahora vamos a proceder a realizar la función cuadrática discriminante la cual vamos a comparar con la LDA ya realizada.

Como observamos las medias de la función QDA nos indican que el pétalo será lo que diferencie unos grupos de otros, como está representado en el siguiente gráfico.



Fuente: Elaboración propia

El grupo que mejor está diferenciado es *setosa* como pasaba en el anterior, por tanto, vamos a analizar las predicciones y con ello elegiremos el mejor modelo que se adapte a nuestros datos.

De la misma manera podemos obtener las predicciones con la muestra de entrenamiento y de test para QDA:

	Setosa	Versicolor	Virginica
Setosa	30	0	0
Versicolor	0	30	0
Virginica	0	0	29

En este caso los resultados son iguales a los anteriores, pero para la muestra de test obtenemos:

	Setosa	Versicolor	Virginica
Setosa	20	0	0
Versicolor	0	16	2
Virginica	0	4	19

Conclusiones

Tras haber realizado el análisis exploratorio hemos procedido a realizar el análisis discriminante tanto para el modelo lineal como el cuadrático, hemos podido diferenciar las observaciones en tres grupos claros que han sido los que nos indicaban la variable categórica que es especies y los grupos son en función de estos: Setosa, Versicolor y Virginica.

Finalmente concluimos que el modelo linear es el que mejor se ajusta a nuestra base de datos, como hemos comprobado gráficamente no existe mucha diferencia entre ambos modelos, pero una vez realizada la predicción nuestra parte de entrenamiento nos arroja los mismos resultados en ambos modelos, pero observando la predicción para la muestra de testeo podemos concluir que el mejor análisis para nuestros datos es el lineal debido a que existen menos flores mal clasificadas, dos frente a seis.