MASTER EN DATA SCIENCE PARA FINANZAS

INFORME: ANÁLISIS DISCRIMINANTE

LUIS LLERA GARCÍA
TÉCNICAS DE CLASIFICACIÓN

1.EXECUTIVE SUMMARY

El objetivo del siguiente análisis será aplicar un Análisis discriminante sobre una base de datos, utilizaremos la base de datos de IRIS, que está predefinida en R-studio.

El Análisis Discriminante es una técnica estadística descriptiva que busca describir características específicas para la diferenciación entre los distintos grupos, los grupos se diferencian por medio de funciones lineales conocidas, como funciones discriminantes los que nos definirán la aportación de cada una, lo harán en función de sus pesos. Dichos pesos se denominan pesos discriminantes, estos serán los coeficientes. Dicha función es la combinación lineal de las p variables que maxifican la distancia entre las medias de los grupos.

En este caso realizaremos el análisis discriminante tanto lineal como cuadrático.

El objetivo es clasificar todas las observaciones en grupos predefinidos, en función de las similitudes entre ellos y los casos pertenecientes entre ellos mediante funciones lineales o cuadráticas, estas reciben el nombre de variables clasificadoras.

Pretende identificar la contribución de cada variable a la separación entre los grupos y encontrar un plano óptimo donde los puntos puedan ser representados y se puedan observar los diferentes grupos.

En dicho análisis se deben realizar un conjunto de pruebas previas que nos confirmen que es correcto realizar el Análisis discriminante, igualdad de medias por ejemplo, pero en nuestro caso no tiene sentido, ya que con el análisis exploratorio de datos observamos que las medias difieren(función summary en R).

Por tanto, analizaremos uno de los supuestos del Análisis Discriminante, que es el Test M-Box o Esfericidad de Bartlett, los cuales, son contrastes que analizan la homocedasticidad. Otro test que realizaremos más adelante es el de Shapiro que analiza la normalidad.

Concluyendo con la introducción el objetivo del siguiente informe será la división en función de las especies de las flores en grupos predefinidos que serán tres, virgínica, setosa y versicolor, que son las especies de dichas flores, la variable especies es la variable categórica.

Dichos análisis, se realizará con el programa R studio.

2.ANÁLISIS EXPLORATORIO DE DATOS

Previo al análisis discriminante debemos de realizar un análisis exploratorio de los datos, no será un análisis exhaustivo debido a que al ser una base de datos predefinida, no nos vamos a tener que enfrentar a la limpieza de la base de datos, como la gestión de valores faltantes (NAS), analizaremos las medias, para ver en cuanto difieren, su mediana, sus cuartiles…

Analizamos la estructura de los mismos, y se observa que tenemos cuatro variables numéricas y otra de tipo factor, que es la variable de tipo categórico la que posee la información para realizar los grupos.

Compararemos las medias, las cuales, nos arrojan los siguientes resultados, sepal.lenth con media 5.8, sepal.width con media 3.057, petal.lenth 4.3 y petal.width 1.199.

Ahora realizaremos el test M-Box que con un p-valor de 2.2e-16, se rechaza la hipótesis nula de que existe homocedasticidad, es un test muy sensible a la ausencia de normalidad y hay variables que no son normales.

Debido a todo eso podemos continuar con el análisis.

Ahora realizaremos un análisis de las correlaciones entre las variables.

	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

Los pétalos tienen una correlación muy alta con la longitud del sépalo, pero las demás correlaciones aparentemente son bastante bajas.

Utilizaremos una semilla(123).

3.ANÁLISIS DISCRIMINANTE

Para realizar el Análisis Discriminante, comenzaremos con la extracción de una parte de la base de datos, una parte será de entrenamiento y otra de testeo(0.6,0.4 respectivamente).

Ahora con la función *'lda'* de la librería MASS, vamos a realizar el Análisis discriminante propiamente dicho, la función anterior nos arroja los siguientes resultados.

Las probabilidades a priori de cada grupo que son:

Setosa	Versicolor	Virgínica
0.3370787	0.3370787	0.3258427

Las cuales, nos indican probabilidad de que el individuo descrito por el vector x pertenezca a la clase en cuestión.

Los diferentes pesos en las funciones discriminantes:

LD1 = SEPAL.LENGTH(0.3629)+SEPAL.WITH(2.22)+PETAL.LENGTH(-1.7854)+PETAL.WITH(-3.9745)

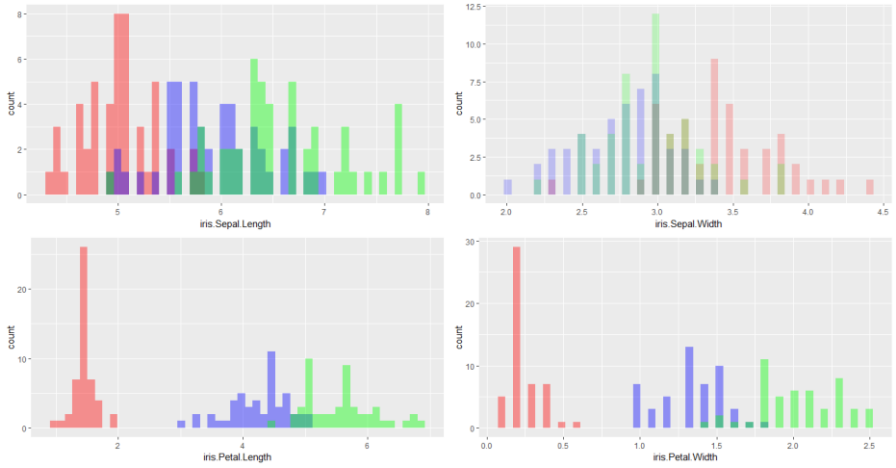
Aquí la variable que tiene un peso más grande y por tanto discrimina mejor es petal.with

LD2 = SEPAL.LENGTH(0.05) + SEPAL.WITH(1.47) + PETAL.LENGTH(-1.6) + PETAL.WITH(4.10)

Y en la segunda función la variable que mejor discrimina es la misma.

Los coeficientes se utilizan para definir a que clase pertenece cada ejemplar de la flor.

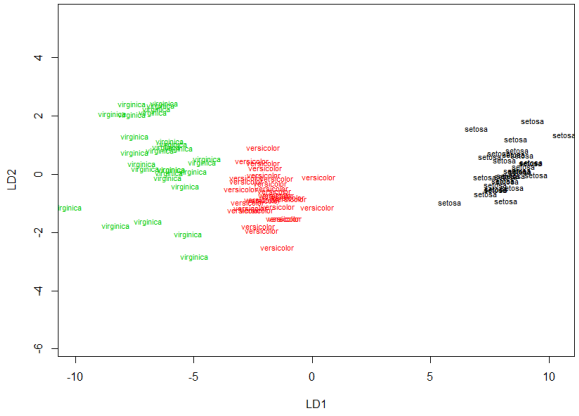
Como podemos observar en el siguiente gráfico, la variable que mejor discrimina es el pétalo, y concretamente, como hemos indicado anteriormente es el petal.with.



Fuente: elaboración propia

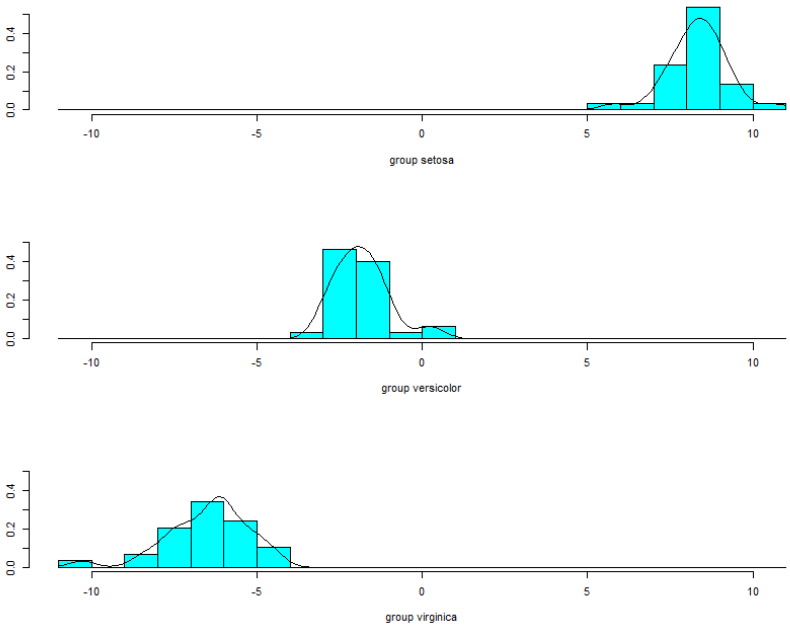
La proporción de la traza, al ser la proporción de LD1(0.9932) muy alta y prácticamente es uno, eso nos confirma que las flores las podemos clasificar prácticamente perfectamente utilizando un eje discriminante. La segunda proporción de la traza será LD2(0.0068).

El siguiente gráfico nos muestra los diferentes grupos que podemos representar en dos dimensiones con el eje ‘x’ que es el LD1 y eje ‘y’ LD2, como podemos ver están diferenciados perfectamente.



Fuente: elaboración propia

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

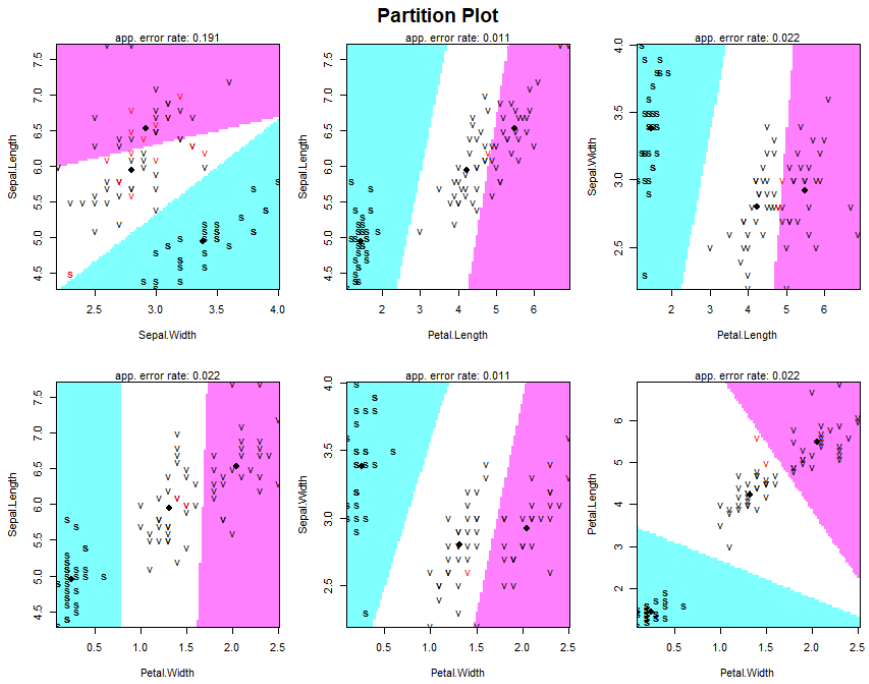


Fuente: elaboración propia

Como se puede observar los tres grupos se diferencian perfectamente y aparentemente son normales y la que mejor se diferencia es la setosa.

LUIS LLERA GARCÍA
TÉCNICAS DE CLASIFICACIÓN

El uso de la función partimat del paquete klaR proporciona una forma alternativa de trazar las funciones discriminantes lineales. Partimat emite una serie de gráficos para cada combinación de dos variables. Cada gráfico es una vista diferente de los mismos datos. Las regiones coloreadas delinean cada área de clasificación. Se predice que cualquier observación que caiga dentro de una región será de una clase específica. Cada gráfico también incluye la tasa de error aparente para esa vista de los datos.



Fuente: elaboración propia

Como podemos observar en el gráfico el grupo mejor diferenciado es la setosa, si observamos la función lda, las medias de los grupos concretamente, los valores que mas difieren, unos con otros es el pétalo, como hemos definido anteriormente, es el de la setosa tanto la longitud como la anchura.

Ahora realizamos la predicción, con la parte tanto del test como la de entrenamiento.

	Setosa	Veriscolor	Virgínica
Setosa	30	0	0
Veriscolor	0	30	0
Virgínica	0	0	29

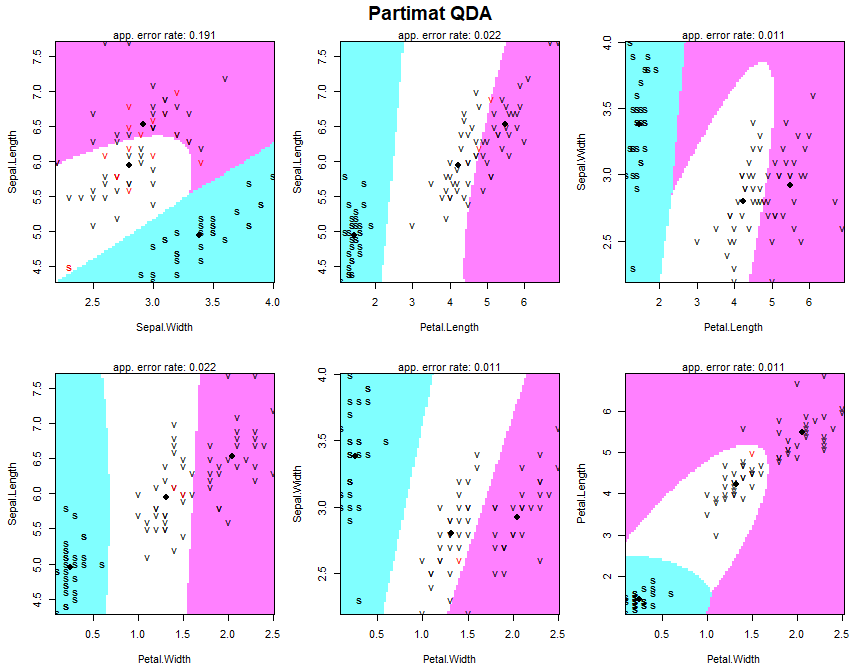
Como podemos observar la parte de entrenamiento, la suma de la diagonal principal es la suma de las observaciones, estamos realizando una buena predicción, ahora con la parte del test.

	Setosa	Veriscolor	Virgínica
Setosa	20	0	0
Veriscolor	0	19	1
Virgínica	0	1	20

Observamos que existen varias malas clasificaciones erróneas, existen dos flores que se han clasificado de manera errónea.

Para saber si hemos realizado un buen análisis vamos a realizar el modelo discriminante cuadrático, y finalmente compararemos los resultados.

Como podemos observar, las medias de la función QDA, nos indican que el pétalo va a ser lo que diferencie unos grupos de otros, como podemos observar en el siguiente gráfico.



Fuente: elaboración propia

Como podemos observar el grupo que mejor está diferenciado es setosa otra vez, por tanto vamos a analizar las predicciones y con ello elegiremos el mejor modelo para nuestros datos.

QDA, con la parte de entrenamiento:

	Setosa	Veriscolor	Virgínica
Setosa	30	0	0
Veriscolor	0	30	0
Virgínica	0	0	29

QDA, con la parte de test:

	Setosa	Veriscolor	Virgínica
Setosa	20	0	0
Veriscolor	0	16	2
Virgínica	0	4	19

4.CONCLUSIONES

Una vez realizado el Análisis discriminante tanto por el método lineal como por 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 los mismos, setosa, versicolor y virgínica.

Finalmente podemos concluir con que el modelo lineal 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 train nos arroja los mismos resultados en ambos modelos.

Pero observando 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 flores mal clasificadas.