Técnicas de Clasificación

PRÁCTICA 1: Análisis discriminante. Iris.

V. Miguel Sempere Navarro

En el presente trabajo se va a analizar un famoso conjunto de datos de flores recogidos en el que se registran los datos de amplitud y longitud del pétalo y del sépalo de tres especies distintas. En la actualidad, este dataset está clasificado por especie. Sin embargo, se va a estudiar si, con los datos que tenemos, seriamos capaces de predecir de qué especie se trataría un iris que cogiésemos por ahí.

Nuestro trabajo tendrá un doble objetivo: un objetivo explicativo para analizar la contribución de cada variable a la función discriminante y un objetivo predictivo para determinar qué especie de flor se asigna a cada observación.

Para realizar el análisis discriminante es necesario que se cumplan dos condiciones:

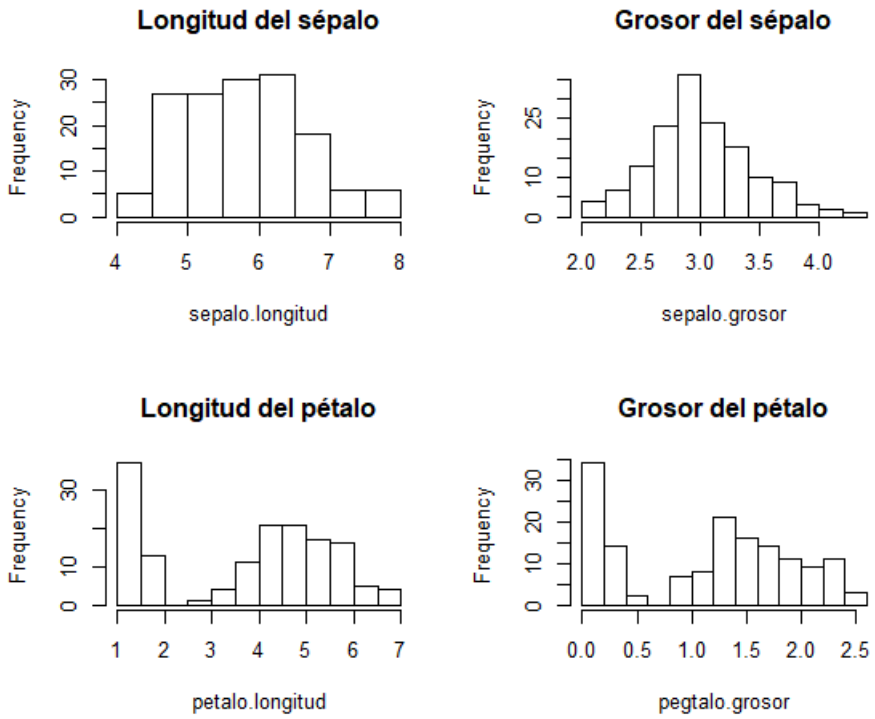
1. Que cada predictor del modelo se distribuya de forma normal en cada una de las clases de la variable respuesta. En el caso de múltiples predictores, las observaciones siguen una distribución normal multivariante en todas las clases.
2. Que la matriz de covarianza sea igual para todas las clases. En el caso de que esto no se cumple, se puede realizar un análisis discriminante cuadrático (QDA).

El dataset objeto de estudio está compuesto por 150 observaciones que reflejan el ancho y largo de los pétalos y sépalos de tres especies diferentes de la flor del iris. Como se ha expuesto anteriormente, las variables son las siguientes: “longitud de sépalo”, “grosor de sépalo”, “longitud de pétalo” y “grosor de pétalo”. Estas variables explicativas son las que contribuyen a definir la variable explicada y categórica ‘especie’ la cual está formada por tres factores -las tres posibles clases de iris-: setosa, versicolor, virginica.

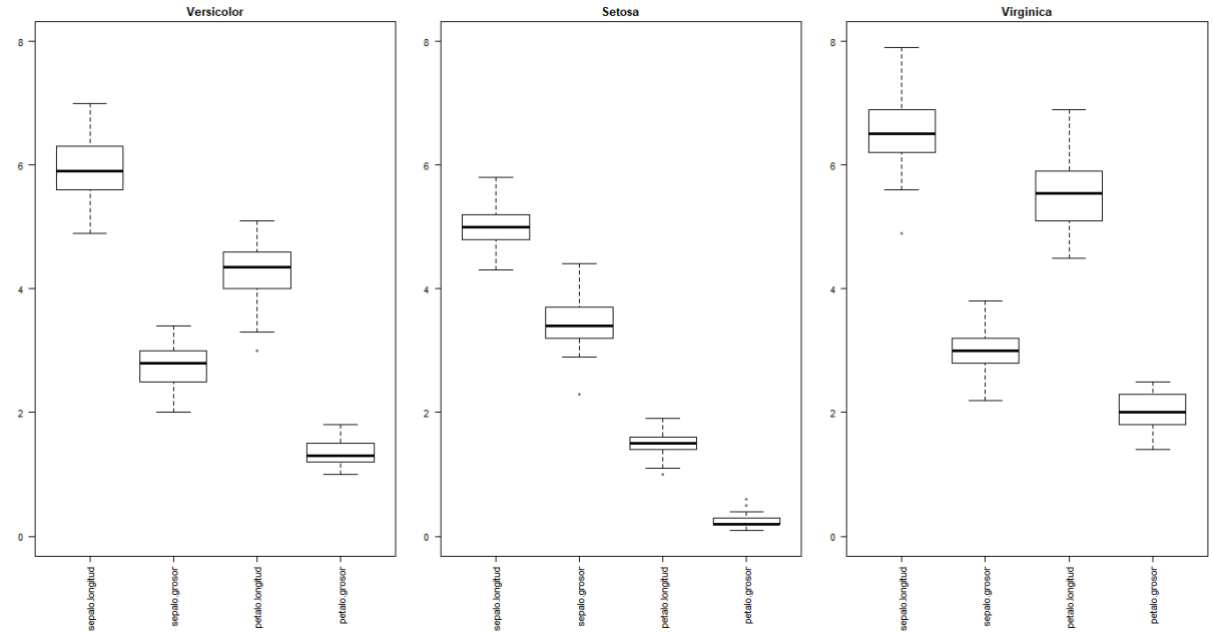
En una primera aproximación a los datos podemos ver cuál es la media, mediana y cuartiles de las variables discretas y también saber cuántas especies hay de cada flor:

```
sepalo.longitud sepalo.grosor   petalo.longitud petalo.grosor        especies
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Sin embargo, si atendemos a sus distribuciones, observamos que no existe una normalidad. Por ello, extraemos como primera conclusión que tiene más sentido clasificar los datos al principio por especie y no por sus características:

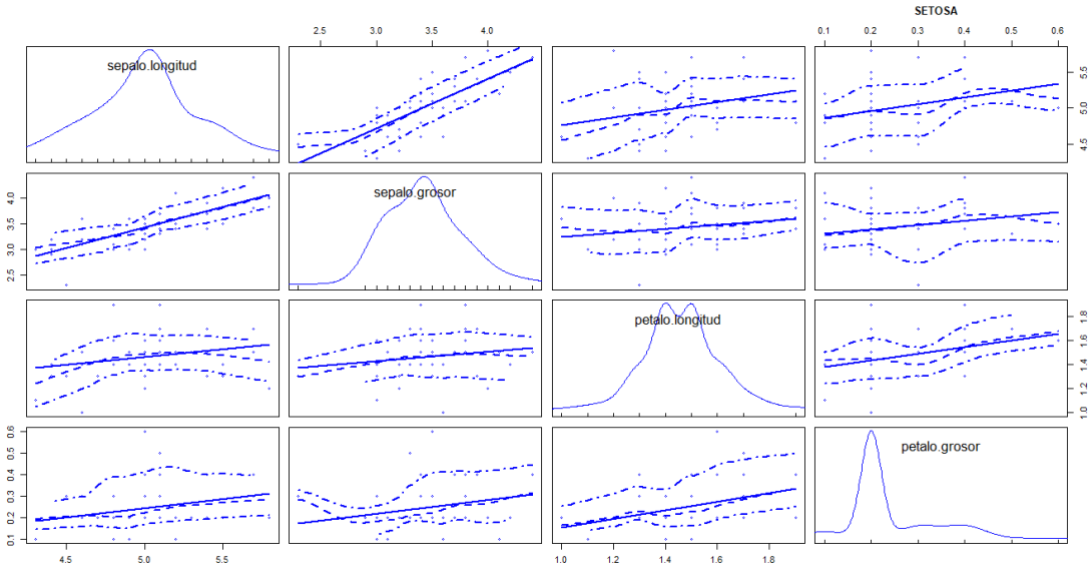


Con el siguiente gráfico vamos a ver cuáles son los valores de la media, mediana y outliers de las cuatro variables para cada especie, representados en un gráfico de cajas:



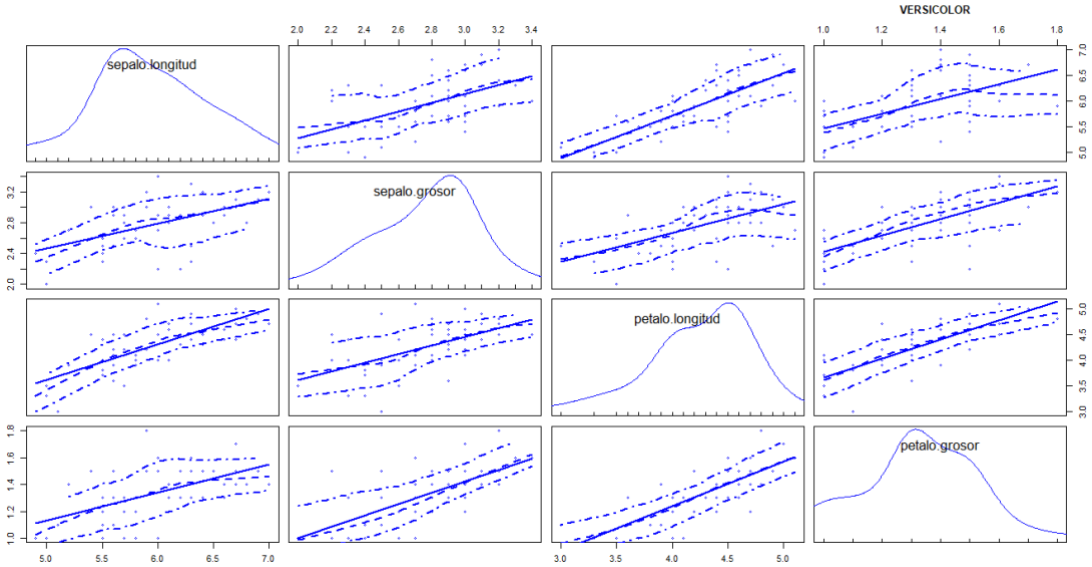
Observamos que, según la especie de flor, hay valores distintos de las mismas variables. Como era de esperar.

Ahora vamos a ver, con un gráfico, cuáles son todas las funciones de densidad de las variables discretas -en la diagonal principal- y las correlaciones de dichas variables en función de la especie de iris a la que pertenecen:

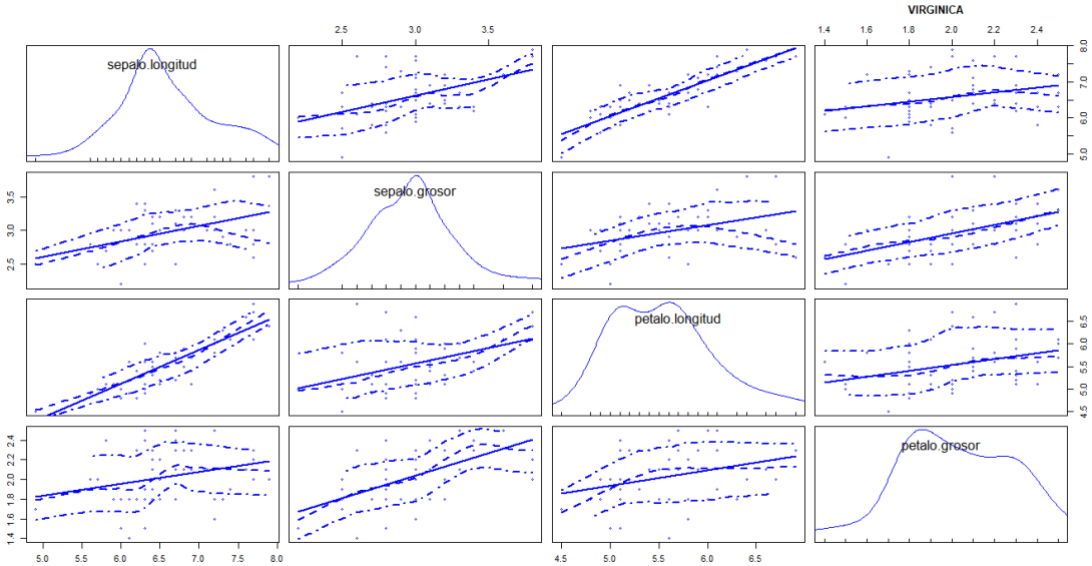


Para la especie *setosa* las funciones de densidad corresponden a distribuciones normales y las correlaciones son todas positivas, siendo la más alta la del 'sepalo.longitud' y 'sepalo.grosor' con un valor de 0.74, lo que supone que ante aumentos o disminuciones en la longitud del sépalo se producen también aumentos o disminuciones en el grosor del sépalo, en la misma dirección y en una proporción muy grande.

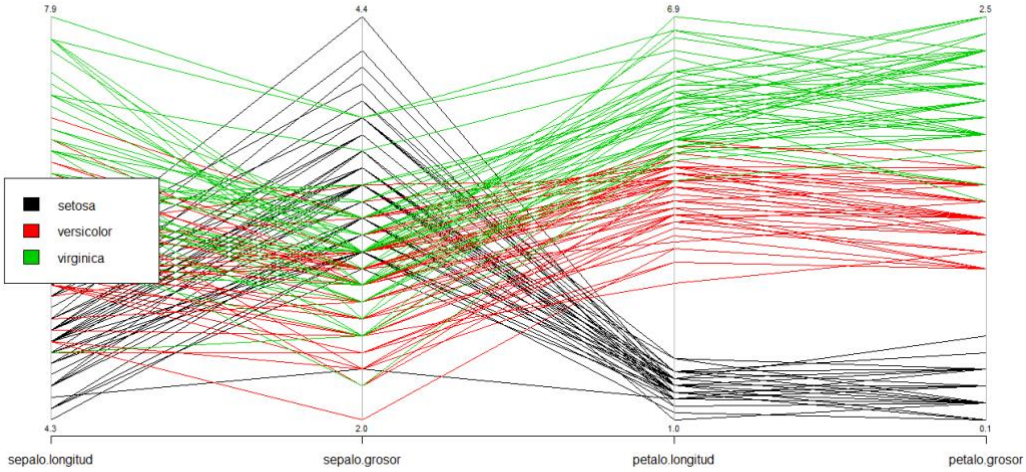
Se observa como la función de densidad para la especie *versicolor* también sigue una distribución normal y las correlaciones entre las variables son más altas que para la *setosa*. Las correlaciones más altas se dan entre 'petalo.longitud' y 'petalo.grosor' siendo de 0.78 y entre 'sepalo.longitud' y 'petalo.longitud' del 0.75.



Por último se observa la densidad con distribuciones normales para la especie *virginica,* presentando una correlación muy elevada entre 'sepalo.longitud' y 'petalo.longitud' del 0.86.



Para terminar con la aproximación a los datos, en este gráfico se puede ver cuál es la dispersión de las observaciones de cada flor, donde el color de las líneas muestra la especie:



Se puede observar que la *setosa* tiene sépalos más cortos, pero más anchos y pétalos más cortos y estrechos, la *versicolor* muestra tamaños medios de todas las variables y la *virginica* presenta sépalos más largos, pero con anchura media y pétalos más largos y anchos.

Ya con un análisis exploratorio de datos realizado, vamos a aplicar distintos test que nos van a permitir discriminar y/o clasificar según otras características.

El primer test será el denominado Saphiro-Wilk. Como hemos visto, aparentemente, las variables independientes del tamaño de los pétalos y sépalos siguen una distribución normal en todas sus variantes, por lo que está justificada la aplicación del LDA, pero vamos a comprobarlo mediante un contraste de normalidad Saphiro- Wilk.

especies	variable	p_value_Shapiro.test
setosa	sepalo.longitud	0.45951
setosa	sepalo.grosor	0.27153
setosa	petalo.longitud	0.05481
setosa	petalo.grosor	0.00000
versicolor	sepalo.longitud	0.46474
versicolor	sepalo.grosor	0.33800
versicolor	petalo.longitud	0.15848
versicolor	petalo.grosor	0.02728
virginica	sepalo.longitud	0.25831
virginica	sepalo.grosor	0.18090
virginica	petalo.longitud	0.10978
virginica	petalo.grosor	0.08695

Podemos observar que p-value < 0.05 en 'petalo.grosor' de setosa y versicolor, por lo que rechazamos la hipótesis nula en esas observaciones de que cumplan una distribución normal.

Pese a ello vemos que en el resto de variables sí que se cumple dicha normalidad, por lo que el LDA sigue siendo robusto y seguiremos con su aplicación.

El otro requisito para que sea de aplicación el LDA es que la matriz de covarianza sea igual en todas las clases. Si no lo es sería de aplicación el Análisis Discriminante Cuadrático (QDA).

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  iris[, -5]
Chi-Sq (approx.) = 140.94, df = 20, p-value < 2.2e-16
```

Según el test se rechazaría la hipótesis nula de que las matrices de covarianzas son homogéneas, pero este test es muy sensible a la ausencia de normalidad en la distribución de todas las variables, por lo que vamos a asumir que sí que son homogéneas y continuar con el LDA.

Con este análisis (LDA) vamos a reconocer patrones y generar un modelo que encuentre una combinación lineal de rasgos que puedan caracterizar a las diferentes especies de iris.

```
Call:
lda(iris$especies ~ ., data = iris)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           sepalo.longitud sepalo.grosor petalo.longitud petalo.grosor
setosa               5.006         3.428           1.462         0.246
versicolor           5.936         2.770           4.260         1.326
virginica            6.588         2.974           5.552         2.026

Coefficients of linear discriminants:
                     LD1         LD2
sepalo.longitud  0.8293776  0.02410215
sepalo.grosor    1.5344731  2.16452123
petalo.longitud -2.2012117 -0.93192121
petalo.grosor   -2.8104603  2.83918785

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

En el modelo se muestra que la asignación de probabilidades a priori a cada grupo es equitativa (1/3) y se asigna la misma a cada especie. También se recogen los valores medios de las observaciones en función de la especie.

La proporción de la traza para cada LD es 0.9912 y 0.0088, lejos de unos valores proporcionados.

Tras realizar una predicción sobre el modelo LDA, obtenemos las probabilidades de cada observación de pertenecer a cada una de las tres especies. Observamos que todas las probabilidades tienen un alto grado de firmeza, al superar en todos los casos el 75%.

Y después el modelo asigna a cada observación la especie más probabilidad tenga de que le corresponda:

```
$posterior
          setosa   versicolor    virginica
1   1.000000e+00 3.896358e-22 2.611168e-42
2   1.000000e+00 7.217970e-18 5.042143e-37
3   1.000000e+00 1.463849e-19 4.675932e-39
4   1.000000e+00 1.268536e-16 3.566610e-35
5   1.000000e+00 1.637387e-22 1.082605e-42
6   1.000000e+00 3.883282e-21 4.566540e-40
7   1.000000e+00 1.113469e-18 2.302608e-37
8   1.000000e+00 3.877586e-20 1.074496e-39
9   1.000000e+00 1.902813e-15 9.482936e-34
10  1.000000e+00 1.111803e-18 2.724060e-38
11  1.000000e+00 1.185277e-23 3.237084e-44
12  1.000000e+00 1.621649e-18 1.833201e-37
13  1.000000e+00 1.459225e-18 3.262506e-38
14  1.000000e+00 1.117219e-19 1.316642e-39
15  1.000000e+00 5.487399e-30 1.531265e-52
```

```
  [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [12] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [23] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [34] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa
 [45] setosa     setosa     setosa     setosa     setosa     setosa     versicolor versicolor versicolor versicolor versicolor
 [56] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
 [67] versicolor versicolor versicolor versicolor virginica  versicolor versicolor versicolor versicolor versicolor versicolor
 [78] versicolor versicolor versicolor versicolor versicolor versicolor virginica  versicolor versicolor versicolor versicolor
 [89] versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor versicolor
[100] versicolor virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[111] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[122] virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[133] virginica  versicolor virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[144] virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

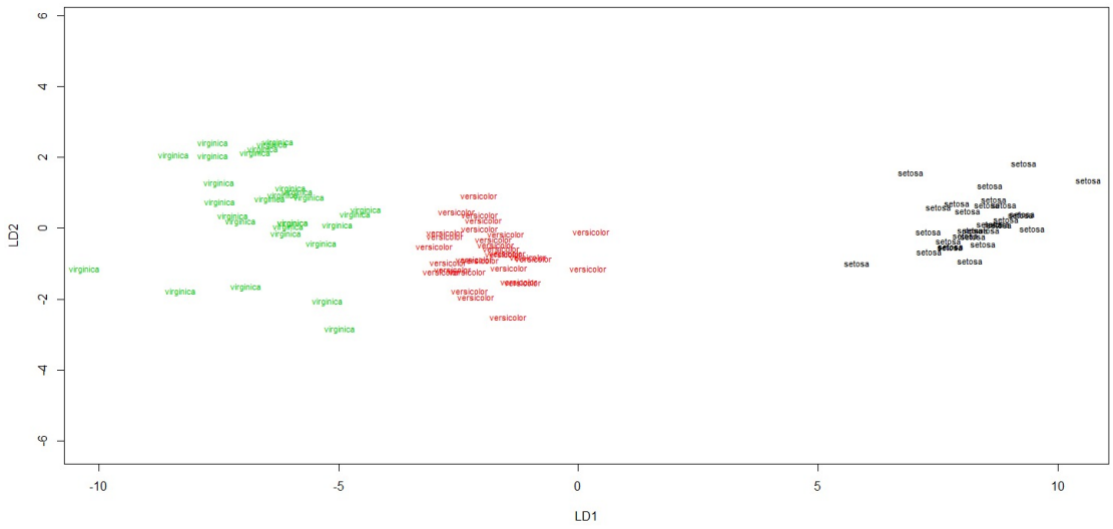
Una vez asignada cada observación a su clase, vamos a comprobar cual ha sido la eficacia de nuestro modelo LDA con una tabla de contingencia:

Observamos que el modelo ha predicho bien las 50 observaciones de la especie *setosa*, ha clasificado 48 *versicolor* como tal y 2 que lo eran como *virginica,* y 49 *virginica* correctamente, asignando una que era *virginica* a la especie *versicolor.*

El error de nuestro modelo es ínfimo, (3/150) de tan solo 2%, lo que indica que el modelo es bueno.

Hemos visto que, aunque no se cumpla la condición de tener una distribución de normalidad multivariante, el modelo LDA se ajusta mucho a la realidad y ha hecho una predicción muy buena.

La distribución de las especies en función del análisis discriminante lineal quedaría de la siguiente manera:



Hemos realizado un Análisis Discriminante Lineal y hemos podido comprobar que es un modelo muy bueno para predecir una clasificación de este dataset, por lo que no sería necesario realizar un Análisis Discriminante Cuadrático.