PROBLEMAS PROPUESTOS DE AD
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

PROBLEMA 1

El conjunto de datos **iris** de R contiene las variables Sepal.Length (X1, longitud de los sépalos), Sepal.Width (X2, anchura de los sépalos), Petal.Length (X3, longitud de los pétalos), Petal.Width (X4, anchura de los pétalos). medidas en centímetros, de tres especies diferentes de flor de iris (Species: *setosa, versicolor* y *virginica*). Se desea realizar un análisis discriminante para determinar la especie de las flores de iris a partir de las magnitudes de sus pétalos y sépalos. Se pide:

1) Cargar el conjunto de datos y realizar un estudio descriptivo previo atendiendo a nuestro objetivo. En particular, debes dar respuesta a las cuestiones:
    a. ¿Existe alguna variable que permita discriminar entre las especies? Hacer gráficos caja-bigotes de cada variable distinguiendo por especie.
    b. ¿Se podrían separar bien los grupos solamente con dos variables? Realizar gráficos bidimensionales usando símbolos distintos para cada grupo
2) Realizar un análisis discriminante lineal (LDA) con todas las variables y probabilidades a priori iguales.
    a. ¿Cuál es la expresión de las funciones discriminantes en función de las variables observadas?
    b. ¿Cómo se clasificaría una flor con medidas z = (6, 3, 5, 2)? ¿Cuáles serían las probabilidades a posteriori de pertenencia a cada grupo? ¿Sería fiable la clasificación?
    c. Representar gráficamente las funciones discriminantes y la flor anterior a clasificar. ¿Dónde se sitúa esta flor? Según el gráfico, ¿en qué grupo la clasificaría?
3) Realizar un análisis discriminante cuadrático (QDA) incluyendo todas las variables y considerando probabilidades a priori iguales.

a. ¿Cómo se clasificaría a una flor con medidas z = (6, 3, 5, 2)? ¿Cuáles serían las probabilidades a posteriori de pertenencia a cada grupo? ¿Sería fiable la clasificación?
b. ¿Dónde se clasificaría z si las probabilidades a priori fueran 0.5, 0.25 y 0.25 para las especies *setosa*, *versicolor* y *virginica*, respectivamente? ¿Cuánto valdrán las probabilidades a posteriori? ¿Sería fiable la clasificación en este caso?

4) Estimar las probabilidades de clasificar correctamente a una flor desconocida usando LDA y QDA con validación cruzada y suponiendo las proporciones a priori iguales.
   a. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA?
   b. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA dentro de la especie *versicolor*?
   c. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA entre las flores clasificadas como *virginica*?
   d. ¿Cuáles son las flores mal clasificadas? ¿En qué grupo se encuentran y dónde se clasifican?
5) Obtener las matrices de covarianzas y realizar los test de normalidad en cada grupo. Con los resultados obtenidos, ¿qué procedimiento sería más adecuado, LDA o QDA? Razonar las respuestas.