Ejercicios con el conjunto de datos Iris

Realice y reporte los siguientes ejercicios en Python. En su reporte incluya los resultados y su código. Entregue por correo:

a) Reporte pdf (formato IEEE).
b) Código (puede ser un programa o un *notebook*).
c) Datos usados por su código.

Introducción

El conjunto de datos de flores Iris es un conjunto de datos multivariados introducidos por el biólogo Ronald Fisher en su artículo de 1936 "*The use of multiple measurements in taxonomic problems*" como un ejemplo de análisis lineal discriminante.

Iris Versicolor **Iris Setosa** **Iris Virginica**

Los datos cuantifican la variación morfológica de las flores Iris de tres especies relacionadas. Consisten de 50 muestras de cada una de las tres especies de Iris: setosa, virginica, versicolor.

Cada muestra tiene cuatro características: longitud y ancho de los sépalos y pétalos, en centímetros.

Basándose en combinaciones de las cuatro características, Fisher desarrolló un modelo discriminante lineal para clasificar las especies.

EJERCICIOS EN PYTHON

A) Ejercicios Básicos

1. Cargue los datos iris en un *data frame (pandas)* e imprima la forma de los datos, tipo y las 10 primeras filas de los datos. Fuente de datos: https://archive.ics.uci.edu/ml/datasets/Iris.
2. Imprima las llaves y el número de filas y de columnas.
3. Obtenga el número de muestras faltantes o Nan.
4. Cree un arreglo 2-D de tamaño 5x5 con unos en la diagonal y ceros en el resto. Convierta el arreglo NumPy a una matriz dispersa de ScyPy en formato CRS. Nota: una matriz se considera dispersa cuando el porcentaje de ceros es mayor a 0.5.
5. Muestre estadísticas básicas como percentil, media, mínimo, máximo y desviación estándar de los datos. Use *describe* para ello. Imprima sólo la media y la desviación estándar.

6. Obtenga el número de muestras para cada clase.
7. Añada un encabezado a los datos usando los nombres en iris.names y repita el ejercicio anterior.
8. Imprima las diez primeras filas y las dos primeras columnas del *data frame* usando los índices de las columnas.

B) Ejercicios de visualización

Utilizando matplotlib y/o seaborn

1. Cree una gráfica de barras que muestre la media, mínimo y máximo de todos los datos.
2. Muestre la frecuencia de las tres especies como una gráfica de pastel.
3. Cree una gráfica que muestre la relación entre la longitud y ancho del sépalo de las tres especies conjuntamente.
4. Obtenga los histogramas de las variables SepalLength, SepalWidth, PetalLength y PetalWidth.
5. Cree gráficas de dispersión usando *pairplot* de *seaborn* y muestre con distintos colores las tres especies en las gráficas de dispersión.
6. Cree una gráfica usando *joinplot* de *seaborn* para mostrar la dispersión entre la longitud y ancho del sépalo y las distribuciones de estas dos variables.
7. Repita el ejercicio anterior, pero esta vez usando *joinplot* con *kind="hexbin"*.

C) Ejercicios de regresión logística

1. Muestre los percentiles, media y desviación estándar de cada especie ('Iris-setosa', 'Iris-versicolor' e 'Iris- virginica').
2. Cree una gráfica de dispersión de la longitud del sépalo y ancho del pétalo mostrando en la gráfica las tres especies con distintos colores.
3. En el modelado estadístico, el análisis de regresión es un proceso para estimar la relación entre variables. Investigue y describa la regresión logística.
4. Clasifique los datos mediante regresión logística y mida el desempeño de su modelo. Describa la medida usada para medir el desempeño de su modelo.