TRABAJO PRÁCTICO: Nociones de Estadística descriptiva

Introducción:

Esta práctica inicial tiene como objetivo explorar y entender la información que puede proveer un dataset así como también repasar conceptos fundamentales de estadística descriptiva a través de la utilización de un software de análisis estadístico como R.

Consignas:

A partir del dataset *iris.csv*[1], se solicita trabajar sobre las siguientes consignas:

1. **Exploración de datos.** Describa las características de los datos que contiene el dataset: tipo de los atributos, cantidad de instancias, cantidad de instancias de cada clase, etc. Represente con gráficos de barra y de torta las proporciones de la muestra agrupada por cada variedad.
2. **Medidas de posición.** Calcular la media, la moda y la mediana para cada uno de los atributos y analice los resultados obtenidos. Calcule la media para cada una de las variedades y compare con la media de cada atributo. Documente los resultados y las conclusiones.
3. Grafique las variables y observe su comportamiento.
4. Lo observado en los gráficos, ¿se condice con los parámetros calculados en la consigna 1.?

[1] El dataset contiene una muestra de mediciones de tres variedades (Setosa, Versicolor y Virginica) de la flor Iris (lirio). El conjunto de datos o muestra tiene las dimensiones (ancho y largo) de sus pétalos y sépalos de cada variedad.

Pétalo: forma parte de la corola de una flor, la función de los pétalos o de la corola es la de atraer los polinizadores. Los pétalos están unidos en la base, formando un tubo floral.

Sépalo: son los que envuelven a las otras piezas florales en las primeras fases de desarrollo, evitan que los insectos accedan al néctar sin pasar por los estambres y estigmas, a menudo los sépalos son muy reducidos, apareciendo como dientes o crestas.

5. **Medidas de dispersión.** Calcular el desvío estándar y la varianza para cada una de las variables.
6. Calcule el rango de cada atributo y grafique el diagrama de cajas. Documente las gráficas y conclusiones.
7. Realice los gráficos de dispersión incluyendo todas las variables y comente que se observa a simple vista.
8. **Medidas de asociación.** Calcular el coeficiente de correlación de todas las variables y explique el resultado.
9. ¿Qué tipo de gráficos describen mejor la relación entre las variables? Realice los gráficos propuestos y documente las conclusiones.
10. Resuelva las consignas e integre en un único archivo las operaciones realizadas -se sugiere un archivo del tipo "R Markdown"- mostrando de manera sencilla el código, los resultados, las gráficas resultantes y ***sus conclusiones***. Utilice como nombre del archivo **tp00-<legajo>** y envíelo al equipo docente.

Referencias sugeridas:

Data Mining: Concepts and techniques. Jiawei Han and Micheline Kamber. Chapter 2.

An Introduction to R:

https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf