TRABAJO PRÁCTICO: Introducción a EDA

Nociones de estadística descriptiva e introducción al análisis de datos con Pandas

Introducción:

Esta práctica inicial tiene como objetivo explorar y entender la información que puede ser obtenida desde un conjunto de datos, así como también repasar conceptos fundamentales de estadística descriptiva a través de la utilización de un software de análisis es el módulo Pandas del lenguaje Python.

Para el desarrollo del práctico utilizar la Guía de Clase 01 publicada en el repositorio GitHub del curso como referencia inicial y ampliar de ser necesario.

Consignas:

A partir del dataset *Metabolic Syndrome*[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, si el mismo está completo o presenta NA, identifique cómo se agrupan los datos en tiempo y espacio, etc. Utilice herramientas gráficas y de resumen para hacer una primera aproximación.
2. **Medidas de posición.**
    a. Calcular la media, la moda y la mediana para cada uno de los atributos y analice los resultados obtenidos.
    b. Calcule la media por estado civil y por raza. ¿Cómo son esos resultados con respecto a la media general de esos atributos?
    c. Documente los resultados y las conclusiones.

[1] https://www.kaggle.com/datasets/antimoni/metabolic-syndrome

3. Grafique las variables cuantitativas utilizando gráficos de dispersión.
    a. ¿Qué puede decir a partir de esos gráficos sobre la relación entre las variables?
    b. Seleccione una relación de interés e indique cómo es esa relación. ¿Tiene modificaciones según la edad de los individuos? ¿Cómo se relaciona con el sexo de los individuos?
    c. Documente los resultados y las conclusiones.

4. **Medidas de dispersión.**
    a. Calcular el desvío estándar y la varianza para cada una de las variables.
    b. Grafique el diagrama de cajas del índice de masa corporal agrupado por rangos de edad. ¿Qué puede comentar? ¿Existen diferencias si se analiza por razas?
    c. Documente las gráficas y conclusiones.
5. **Medidas de asociación.**
    a. Calcular el coeficiente de correlación de todas las variables y explique el resultado.
    b. ¿Qué relación encuentra con los resultados del punto 3?
6. Guarde los archivos resultantes de las actividades prácticas en una carpeta denominada tp0-<legajo> que a su vez tenga un directorio por cada uno de los puntos de este trabajo, comprima la carpeta y envíelo al equipo docente.

Referencias sugeridas:

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

pandas documentation Date: Apr 10, 2024 Version: 2.2.2
https://pandas.pydata.org/docs/index.html

Mukhiya, S. K., & Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd.