TRABAJO PRÁCTICO 01: Preprocesamiento y transformación de datos

Introducción

En este trabajo práctico se abordan las cuestiones relacionadas con la selección de variables y reducción de dimensionalidad de un dataset a efectos de reconocer aquellos atributos que mejor lo representan.

Se plantean ejercicios y datasets cuyas resoluciones serán realizadas utilizando Pandas en Python.

Parte 1: Manejo de ruido, outliers y transformación de datos.

Consignas:

1. **Manejo de Ruido.** Para el siguiente dataset *signals.csv* realice las siguientes operaciones:
    a. ¿Qué características tienen las variables? ¿Cómo se distribuyen las variables? Verifique gráficamente utilizando un gráfico QQ. ¿Qué puede decir sobre esas distribuciones?
    b. Realice un suavizado utilizando *binning* por *frecuencias iguales* y estime el valor del Bin por el cálculo de medias. Grafique las dos series resultantes y comente los resultados observados.
    c. Utilizando suavizado por extremos calcular los bins con *anchos iguales* de 2 a 10 y comparar los resultados gráficamente. ¿Qué ocurre conforme el bin aumenta?

d. Encuentre un caso de aplicación de binning, consiga datos para probar un suavizado por binning con alguna de las variantes presentadas. Justifique la elección del conjunto de datos y comente en qué contribuye la técnica.

2. **Detección de outliers.** Utilizando el conjunto de datos *weather.csv*[1] analice los siguiente requerimientos:
   a. A partir de una primera exploración qué variables poseen observaciones que pueden etiquetarse como outliers.
   b. Analice las variables de manera gráfica utilizando boxplots.
   c. Verifique a cuantos desvíos de la media se encuentran las observaciones que se identificaron como outliers. ¿Hay relación con los criterios de identificación utilizados en los boxplots? Compare estos valores con los valores (bigotes) del boxplot.
   d. Analice si existen filas con outliers en más de una variable. Agregue una columna "CANT_OUT" al dataset que contabilice la cantidad de variables observadas como outliers en el dataset y realice al menos dos gráficos para ver dónde se ubican esos valores. ¿Qué puede comentar al respecto?

3. **Discretización.** A partir del dataset *weather.csv*, opere sobre el atributo *humedad* de la siguiente manera:
   a. Transforme el atributo a discreto, definiendo 5 rangos de acuerdo al análisis de frecuencia de los valores encontrados para el atributo. ¿Qué tipo de variable se obtiene?
   b. Transforme el atributo a discreto, definiendo 5 rangos utilizando intervalos de clases.

[1] https://www.kaggle.com/datasets/nikhil7280/weather-type-classification

c. Transforme el atributo a discreto, utilizando percentiles cada 20%. Es decir, [0% a 20%), [20 a 30%), ... etc
d. Analice los resultados encontrados. Compare los mismos realizando gráficos de frecuencia sobre los intervalos resultantes en cada caso. ¿Qué conclusiones se pueden obtener en términos del balanceo de las mismas de acuerdo a la técnica utilizada? ¿Son consistentes todos los abordajes?

4. **Normalización.**

Entrada en calor: [Papel y lápiz] Para el siguiente dominio:

$$E = \{1, 5, 9, 17, 18, 34, 45, 89, 99\}$$

¿Cuál sería el valor de *escalado decimal* para 65? ¿Y si utilizo *mínimo-máximo*?

A partir del dataset *weather.csv*, opere sobre los atributos ***Temperature*** *y* ***Wind Speed*** de la siguiente manera:

a. Normalice el atributo utilizando la técnica de mínimo-máximo.
b. Ahora, normalice el atributo mediante la técnica de z-score.
c. Por último, utilice la técnica de escalado decimal para llevar adelante la tarea de normalización.
d. Analice los resultados encontrados. Compare los mismos realizando gráficos sobre los valores originales y además los atributos resultantes en cada caso.
e. Repita el análisis utilizando boxplots y además el agrupamiento por la variable *season.* Compare los resultados obtenidos e indique qué ventajas tiene la comparación con las variables normalizadas.

Parte 2: Manejo de datos faltantes y reducción de dimensionalidad

5. **Datos faltantes.** Identifique cuáles de las variables del dataset *car_ad.csv* poseen datos faltantes. Aplique los siguientes métodos a efectos de reemplazar esos valores:
    a. Analice el tipo de faltante. ¿Qué puede decir sobre eso?
    b. Sustituya los valores faltantes por una medida de tendencia central según corresponda..
    c. Sustituya los valores faltantes de acuerdo al método de "*hot deck imputation*".
    d. Implemente una función *hot_deck* en python siguiendo el criterio explicado en clase de hot deck y compare los resultados obtenidos con los del módulo KNNImputer de Sklearn. Defina el criterio de similitud que crea adecuado para el dataset car_ad.csv
    e. Sustituya los valores faltantes realizando una imputación por regresión.
    f. Analice los resultados encontrados a partir de la aplicación de los métodos anteriores. Compare los mismos realizando gráficos sobre los valores resultantes en cada caso.

6. **Análisis de Componentes Principales.** Cargue el dataset *automobile.csv*[2] y conteste las siguientes consignas:
    a. Calcule la matriz de covarianzas. ¿Qué nos indica la misma sobre los atributos del dataset?

[2] https://www.kaggle.com/datasets/tawfikelmetwally/automobile-dataset

b. Realice ahora el análisis de componentes principales. ¿Cuánto explica de la variación total del dataset la primera componente? ¿Y si se incorpora la segunda? ¿Y el primer auto-valor?
c. Grafique el perfil de variación de las componentes en un gráfico de dispersión donde las X es la varianza y la Y el componente.
d. Analice la matriz de loading. ¿Qué información provee? ¿Qué variables están más correlacionadas con la primera componente?
e. Genere un gráfico de biplot y explique brevemente qué información le provee el mismo.
f. En función de los análisis realizados en los puntos anteriores. ¿Cuántas componentes principales elegiría para explicar el comportamiento del dataset? Justifique esa cantidad.

Referencias sugeridas:

Principal component analysis. Hervé Adbi & otros. 2010.

Data mining and the impact of missing data. Marvin L. Brown & otros. 2003.

Data Mining. Concepts & Techniques. Jiawei Han and Micheline Kamber. 2006.