CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

LABORATORIO I: Conceptos de Estadística Descriptiva & Preprocesamiento (V1)

Introducción:

Esta práctica inicial de laboratorio tiene como objetivo realizar una primera aproximación al Lenguaje R, utilizando el enfoque de análisis exploratorio de datos sobre un dataset, a efectos de repasar conceptos fundamentales de estadística descriptiva.

A su vez, se abordan técnicas correspondientes a la etapa de Preprocesamiento del Proceso de Descubrimiento de Conocimiento.

ESTADÍSTICA DESCRIPTIVA

A partir del dataset *MPI_subnational.csv*[1] (Multidimensional Poverty Measures), se solicita trabajar sobre las siguientes consignas:

1. **Exploración de datos.** Explore y explique en que consiste el dataset utilizando herramientas de exploración de datos.
    a. Releve las características de los atributos.
    b. Represente gráficamente la cantidad de ciudades agrupados por Región.
2. **Medidas de posición.** Calcule las medidas de posición para los atributos numéricos y agrupe los cálculos de acuerdo a la Región.
    a. Ordene los resultados del MPI resultante y concluya al respecto. Help(order).
    b. Grafique las variables y observe su comportamiento (graph : barplot, pie & hist).
3. **Medidas de dispersión.** Calcular el desvío estándar, la varianza y el rango para cada una de las variables.
    a. Realice diagramas de cajas y scatterplot's. Documente las conclusiones.
    b. ¿Qué variable es la que presenta mayor dispersión? Tenga en cuenta que cada variable puede estar expresada en diferentes unidades y magnitudes.

[1] Disponible en: https://www.kaggle.com/ophi/mpi/data

4. **Medidas de asociación.** Calcular el coeficiente de correlación de todas las variables y explique el resultado. ¿Qué tipo de gráficos describen mejor esta relación entre las variables?

PREPROCESAMIENTO:

1. **Integración de datos.** Analice e integre los datasets *MPI_subnational.csv* y *MPI_national.csv*. Tenga en cuenta las cuestiones trabajadas en clase como el método de integración, los nombres de las variables, granularidad, representación, etc.
2. **Atributos redundantes.** Verifique si existen atributos (categóricos o numéricos) redundantes en el dataset y actúe en consecuencia de acuerdo a las técnicas abordadas en clase.
3. **Manejo de Ruido.**
    a. Verifique en primer lugar la distribución de los datos, utilice algún método gráfico para esto. A su criterio, ¿Cuál es la variable más "ruidosa"?
    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 medias, calcular los bins con *anchos iguales* de 2 a 10 y compare los resultados gráficamente. ¿Qué ocurre conforme el bin aumenta?
    d. Compare los métodos de suavizado de los puntos *b.* y *c.*

Referencias sugeridas:

García, S., Luengo, J., & Herrera, F. (2016). Data preprocessing in data mining. Springer.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

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