CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

LABORATORIO III: Detección de Outliers (Preprocesamiento)

Introducción:

Esta práctica de laboratorio tiene como objetivo abordar nuevas técnicas correspondientes a la etapa de Preprocesamiento del Proceso de Descubrimiento de Conocimiento, puntualmente el análisis, detección y tratamiento de valores atípicos (en adelante, outliers).

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R, a efectos de ejercitar los conceptos abordados en las clases teóricas.

CONSIGNAS

A partir del dataset *MPI_national.csv*, se solicita trabajar sobre las siguientes consignas:

1. **SOBRE LOS DATOS**
    a. Cargue[1] y explore el dataset: explique en qué consiste el mismo y qué características posee. La descripción de las variables puede encontrarse en https://www.kaggle.com/jamesmuniu/mpi-nationalcsv
    b. Con las técnicas abordadas en la práctica de laboratorio anterior, realice un breve análisis exploratorio para identificar cual es la distribución de sus variables

2. **TRATAMIENTO DE OUTLIERS**
    a. Verifique la existencia de *outliers* en cada uno de los atributos. ¿Existen atributos que poseen valores atípicos?
    b. Seleccione uno de los *features* del dataset que a su entender posea *outiers* y aplique las técnicas de análisis y detección vistas en clase (IRQ, SD, Z-SCORE, LOF y Mahalanobis).

[1] Explore la instrucción *read.csv().*

c. Realice un análisis en torno a la diferencia de utilizar las diferentes técnicas, que implicancias tienen en la nueva distribución del dato (en caso que se opte por eliminar los valores atípicos) e indague sobre los valores categorizados como *outliers* por cada una de las técnicas. Concluya al respecto.

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.