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**
   1. Cargue[[1]](#footnote-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>
   2. Elija algún método abordado en el material visto hasta ahora y realice un breve análisis sobre la distribución de las variables numéricas.
2. **TRATAMIENTO DE OUTLIERS**
   1. Verifique la existencia de *outliers* en cada uno de los atributos. ¿Existen atributos que poseen valores atípicos?
   2. Seleccione uno de los *features* del dataset que a su entender posea *outiers* y aplique las técnicas de análisis univariadas vistas en clase (IRQ, SD, y Z-SCORE) y compare los resultados.
   3. Observe qué ocurre con la distribución de la *feature* elegida en caso de eliminar los outliers. Grafique un boxplot de con la nueva distribución. Concluya al respecto.
   4. Extienda el análisis a 3 variables y analice si existen valores atípicos utilizando algún método multivariado.

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.

1. Explore la instrucción *read.csv().* [↑](#footnote-ref-1)