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]](#footnote-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.
   1. Releve las características de los atributos.
   2. 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 al País.
   1. Ordene los resultados del MPI resultante y concluya al respecto. Help(order).
   2. 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.
   1. Realice diagramas de cajas y scatterplot’s. Documente las conclusiones.
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.**
   1. Verifique en primer lugar la distribución de los datos, utilice algún método gráfico para esto. A su criterio, ¿Cuál es la variable más “ruidosa”?
   2. 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.
   3. 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?
   4. 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>

1. Disponible en: <https://www.kaggle.com/ophi/mpi/data> [↑](#footnote-ref-1)