LABORATORIO II: Preprocesamiento – Parte I

**Introducción:**

Esta práctica de laboratorio se abordan algunas técnicas correspondientes a la etapa de Preprocesamiento del Proceso de Descubrimiento de Conocimiento que tienen que ver con:

* Integración de datos,
* Selección de atributos y
* Manejo de Ruido.

**PREPROCESAMIENTO:**

1. **Integración de datos.** Analice e integre los datasets *MPI\_subnational.csv* y *MPI\_national.csv[[1]](#footnote-1)*. 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)