Ingeniería de Features (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 la discretización, normalización y transformación de las variables.

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**

En este caso trabajaremos con un set de datos con estadísticas a nivel país de casos de COVID-19 ([Link de descarga](https://github.com/dm-uba/dm-uba.github.io/blob/master/2021/laboratorios/LAB07/Covid_Worldmeters_May_3_2021.csv)). Este dataset fue generado con los datos publicados en el sitio <https://www.worldometers.info/coronavirus/> al 3 de Mayo de 2021.

1. **ANALISIS DEL SESGO**
   1. Calcule el sesgo de la variable Casos Totales y observe la distribución en un histograma.
   2. Modifique este sesgo utilizando una transformación logaritmica y observe las diferencias analítica y gráficamente
2. **DISCRETIZACIÓN DE VARIABLES**
   1. Calcule qué porcentaje representan los casos totales de cada país sobre la población y convierta estos porcentajes en 5 categorías utilizando una discretización por igual frecuencia.
   2. Visualice[[1]](#footnote-1) qué países pertenecen al intervalo de mayor porcentaje de casos.

¿Se encuentra Argentina entre ellos?

1. Se recomienda utilizar la librería *highcharter* y la codificación de países de 3 dígitos ([dataset de conversión](https://github.com/dm-uba/dm-uba.github.io/blob/master/2021/laboratorios/LAB07/CountryCodes.csv)). [↑](#footnote-ref-1)