CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

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). 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**
    a. Calcule el sesgo de la variable Casos Totales y observe la distribución en un histograma.
    b. Modifique este sesgo utilizando una transformación logaritmica y observe las diferencias analítica y gráficamente
2. **DISCRETIZACIÓN DE VARIABLES**
    a. 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.
    b. Visualice[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).