Curso Introducción a la Minería de Datos
Universidad Nacional de Colombia
Semestre 2024-01

Introducción:
En la etapa de preprocesamiento se tiene como objetivo abordar técnicas que permitan obtener información organizada, evitar información que pueda ser redundante, identificar posibles problemas presentes en las bases de datos y hacer el respectivo tratamiento. Específicamente, en este trabajo se desarrollarán metodologías para el tratamiento de datos atípicos y datos faltantes.

1. A partir del dataset *ruidoso.txt* realice los siguientes análisis:
   a. Cargue y explore el dataset explicando en qué consiste y las características que posee el mismo.
   b. Realice un breve análisis exploratorio para identificar la distribución de las variables usadas en la base de datos ¿será que existe relación entre las variables?
   c. Verifique si existen problemas de datos atípicos en cada una de las variables usando las metodologías de detección a nivel univariado.
   d. ¿Se detectan valores atípicos a nivel multivariado?
   e. Para el caso univariado, escoja una variable y realice un análisis sobre las implicaciones que tiene realizar diferentes tratamientos a los datos atípicos en la distribución de la respectiva variable.

2. A partir del dataset *auto-mpg.data-original.txt* (https://archive.ics.uci.edu/dataset/9/auto+mpg) realice los siguientes análisis:
   a. Cargue y explore el dataset explicando en qué consiste y las características que posee el mismo.
   b. Realice un breve análisis exploratorio para identificar la distribución de las variables usadas en la base de datos ¿será que existe relación entre las variables?
   c. Verifique si existen datos faltantes en cada uno de las variables. ¿Cuál es la proporción de datos faltantes en la distribución de las variables?
   d. ¿Cuál cree que es el mecanismo inherente a esos datos faltantes?
   e. Aplique las técnicas de tratamiento de datos faltantes vistas en clase.
   f. Analice gráfica y analíticamente la variación en la distribución de los datos al aplicar las técnicas de imputación de datos. ¿Qué técnica afecta menos la distribución original?