Universidad Francisco Marroquín
Data Wrangling - 2020

Laboratorio #8 - Missing Data and Feature Engineering

Parte 1: (70%)
La tabla “titanic_MD.csv” contiene missing values en varias columnas. Utilizando R o Python, realice lo siguiente:

1. Reporte detallado de missing data para todas las columnas. **(5%)**
2. Para cada columna especificar que tipo de modelo se utilizará y qué valores se le darán a todos los missing values. (Ej. Imputación sectorizada por la moda, bins, y cualquier otro método visto anteriormente). **(10%)**
3. Reporte de qué filas están completas **(5%)**
4. Utilizar los siguientes métodos para cada columna que contiene missing values: **(50%)**
    a. Pairwise deletion
    b. Imputación general (media, moda y mediana)
    c. Imputación sectorizada
    d. Modelo de regresión lineal simple
    e. Eliminación de outliers: Standard deviation approach
    f. Eliminación de outliers: Percentile approach
5. Al comparar los métodos del inciso 4 contra “titanic.csv”, ¿Qué método (para cada columna) se acerca más a la realidad y por qué? **(20%)**
6. Conclusiones **(10%)**

Parte 2: (30%)
Utilizando la misma tabla de “titanic_MD.csv” en R o en Python realice lo siguiente:

1. Luego del pre-procesamiento de la data con Missing Values, normalice las columnas numéricas por los métodos: **(50%)**
    a. Standarization
    b. MinMaxScaling
    c. MaxAbsScaler
2. Compare los estadísticos que considere más importantes para su conclusión y compare contra la data completa de “titanic.csv” (deberán de normalizar también). **(50%)**

El laboratorio deberá de ser entregado por medio de MiU a más tardar el Domingo, 15 de Noviembre a las 11:59pm. No estaremos aceptando entregas tarde ni por correo electrónico. La entrega será el link al documento en GitHub, en formato markdown o PDF, estén trabajando en R o en Python.