Universidad Francisco Marroquín

Data Wrangling

# 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 qué tipo de modelo se utilizará (solo el nombre y el porqué) 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%)**

*No codear, solo explicar*

1. Reporte de qué filas están completas **(5%)**

***List wise, completas***

1. Utilizar los siguientes métodos para cada columna que contiene missing values: **(50%)**
   1. Imputación general (media, moda y mediana)
   2. Modelo de regresión lineal
   3. Outliers: Uno de los dos métodos vistos en clase (Standard deviation approach o Percentile approach)

Aplicar todo a numéricas, y las categóricas solo la moda.

1. 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%)**
2. Conclusiones **(10%)**

*Punto 4, 5 y 6.*

**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%)**
   1. Standarization *-> Promedio 0, std 1. De lo contrario está mal.*
   2. MinMaxScaling *-> scaling escalar [0,1]*
   3. MaxAbsScaler *-> Obviar*

*Métodos para variables numéricas.*

1. 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, 20 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.