**Laboratorio 8 – Missing data y Feature Engineering**

# Parte 1

## Missing data por columna

Para comenzar, vale la pena mencionar que se eliminaron las columnas de “PassengerId” y la columna de “Name” porque no aportaban información para la futura creación del modelo.

Se debe aclarar que por datos faltantes entendemos un espacio en blanco, un NA o un valor que no haga sentido con la información solicitada. Por ejemplo, preguntar edad y tener como respuesta un carácter.

Las columnas con valores faltantes son las siguientes:

* **Sex**: Tiene 51 filas donde en lugar de tener el sexo del pasajero se tiene un signo de interogación “?”, posteriormente se necesitará hacer un ajuste para reconocerlo como NA.
* **Age**: La columna de edad tiene 25 valores faltantes, con NA.
* **SibSp**: La columna tiene tan solo 3 valores faltantes, con NA.
* **Parch**: En esta columna faltan 12 datos, con NA.
* **Fare**: A esta columna le faltan 8 datos, con NA.
* **Embarked**: Esta columna tiene 12 datos donde está vacío.

## Manejo del missing data por columna

* **Sex**: La forma en la que esta variable categórica se imputará es reemplazando los NA con el último valor no NA. Ej. 1, NA, NA, 3, 5, NA, 7 sería 1, 1, 1, 3, 5, 5, 7. Esto es solo un ejemplo, ya que se sabe que la variable “Sex” solo toma 2 valores o Masculino o Femenino. Esto se hace para tratar de no sesgar la distribución y al mismo tiempo es una de las formas más sencillas de imputar. Si esto no funciona y sesga la distribución entonces se utilizará una regresión logística para imputar.
* **Age**: Los valores faltantes para esta columna serán imputados por medio de la mediana, esto con la intención de no sesgar más la distribución de lo que ya se encuentra.
* **SibSp** (# de hermanos o pareja a bordo): Primero hay que mencionar que es una variable categórica, por lo que el método más directo de imputación es la moda. Solo hay 3 valores faltantes, estos representan un 1.6% del total de observaciones, algo mínimo.
* **Parch** (# de hijos o padres a bordo): Nuevamente, esta es una columna tipo categórica por lo que se imputará por medio de la moda ya que los datos faltantes representan tan solo el 6.5% del total; no se sesgará tanto por eso.
* **Fare**: Nuevamente se utilizará la imputación por la mediana, se tiene valores “muy extremos” que hacen que tenga una distribución con sesgo positivo, por lo que se quiere evitar que los datos que se imputen estén altamente afectados/sesgados por esos datos atípicos.
* **Embarked**: Esta es una variable categórica, en porcentaje, los 12 datos faltantes representan el 6.5% de las observaciones. Por lo mismo, se utilizará la moda, aunque se sabe que puede sesgar un poco la distribución, al no ser una cantidad considerable de datos es “aceptable”.

## Filas que están completas

* **Survived:** Si sobrevivió es 1; de lo contrario es 0
* **Pclass:** Clase en la que viajaba el pasajero.
* **Ticket:** Número del ticket.
* **Cabin:** Número de cabina donde se hospedaba.

## Imputando los datos y tratamiento de los outliers