**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

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)

\*\*Todo el tratamiento de los que se encuentra en el archivo Script de R, donde se realizó todo el código.

## ¿Qué método se acerca más a la realidad y por qué?

Como se mencionó en el código, solo se compararon los datos en las variables de Fare y Age, porque todas las demás variables como missing values eran categóricas y se tenían que imputar por la moda y se sabe por definición que eso obviamente sesgará totalmente como se distribuyen las categorías porque solo se asignan a los NA la categoría que es más frecuente.

En lo que respecta a las variables de Fare y Age, luego de comparar la distribución de la data real y la imputación por medio de la media y de la mediana, se puede concluir que cualquiera de los dos métodos se ajusta muy bien a los datos originales; simplemente hay unas diferencias de unos decimales y por la naturaleza de los datos que estamos tomando es prácticamente irrelevante esa diferencia.

## Conclusión

Trata “Missing Values” podría decirse que es un “arte”, en este caso se puede concluir que en base a las variables numéricas que imputamos se comportan de acuerdo con cómo se comportan los datos de el dataset original, por lo que no es una preocupación. Por otro lado, las variables categóricas al ser imputadas por la moda agregarán sesgo a la distribución que antes se tenía, por lo que será conveniente siempre revisar que tanto afecta ese sesgo a los modelos de machine learning que se harán. Cabe mencionar, que se sabe que la imputación por la moda agrega sesgo a la distribución, pero se “está a favor” se aceptar esto ya que la cantidad de datos que se imputan de esta manera no son considerables (excepción de la variable Sex y por eso se imputó con un modelo) y entonces las “ganancias” de información que aportan al modelo son mayores a los costos de esta imputación.

# Parte 2

## Luego del pre-procesamiento de la data con Missing Values, normalice las columnas numéricas por los métodos:

1. Standarization
2. MinMaxScaling
3. MaxAbsScaler

\*\*En el código se podrá observar la normalización de las columnas numéricas.

Se utilizó las variables de Age y Fare que habían sido imputadas por la mediana por las razones dichas en la sección anterior.

## Compare los estadísticos que considere más importantes para su conclusión y compare contra la data completa de “titanic.csv”

Nuevamente, gracias a la función “summary” de R, podemos tener datos de los cuantiles y otros datos propios de una gráfica de caja y bigotes (solo los datos, no la gráfica). De esa forma se pudo comparar entre los resultados de las normalizaciones/estandarizaciones de la data original y la que se obtuvo de imputaciones previas. Observando, se obtuvo un resultado satisfactorio ya que las variaciones entre la original y la completa es enormemente parecida, únicamente varían por solo unos decimales, en algunos casos solo por centésimas o milésimas. Por lo que a grandes rasgos se podría decir que se comportan prácticamente igual y que el dataset que se preparó con imputaciones y otras transformaciones es un dataset bien preparado para utilizar en un futuro modelo.