Práctica 2: Limpieza y análisis de datos

Hugo Martín Méndez

Marc Subira Zurita

Tipología y ciclo de vida de los datos

09 de Junio de 2020

# Descripción del dataset

El conjunto de datos pertenece a los datos de los pasajeros del titanic. Los datos contienen datos de 887 pasajeros. Cada fila representa una persona. Las columnas describen diferentes atributos acerca de la persona, si sobrevivió o no, su edad, la clase en la que viajaba, el genero y el coste del billete.

Los datos han sido dividido en dos:

* training set (train.csv)
* test set (test.csv)

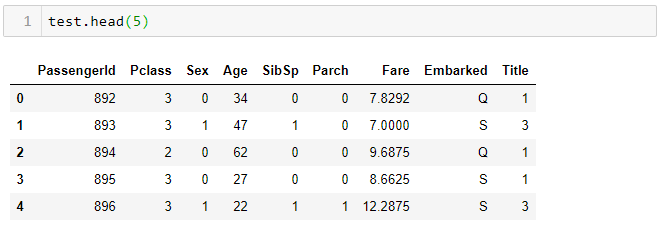
El training set se utiliza en el caso de querer construir un modelo de machine learning. El test set, se utiliza para comprobar que bien trabaja el modelo que se ha realizado.

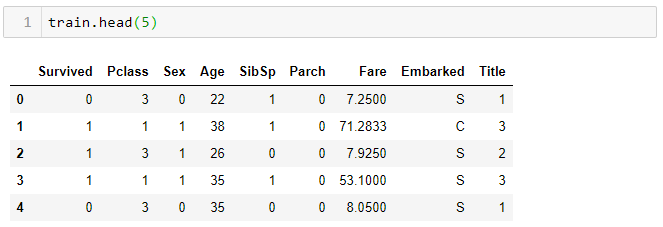
Las columnas para el set test:

* **PassengerId.** Id del pasajero
* **Pclass.** Clase del billete. 1, primera clase, 2, segunda clase, 3, tercera clase.
* **Sex.** Genero del pasajero, hombre o mujer.
* **Age.** Edad.
* **SibSp.** Numero de gemelos.
* **Parch.** Numero de padres con hijos a bordo.
* **Fare.**
* **Embarked.** Puerto de embarque. C=Cherbourg, Q=Queenstown, S=Southampton
* **Title.** Titulo de la persona.

**Columnas del set train:**

* **Survived.** 1 si sobrevivió, 0 si murió.
* **Pclass.** Clase del billete. 1, primera clase, 2, segunda clase, 3, tercera clase.
* **Sex.** Genero del pasajero, hombre o mujer.
* **Age.** Edad.
* **SibSp.** Numero de gemelos.
* **Parch.** Numero de padres con hijos a bordo.
* **Fare.**
* **Embarked.** Puerto de embarque. C=Cherbourg, Q=Queenstown, S=Southampton
* **Title.** Titulo de la persona.





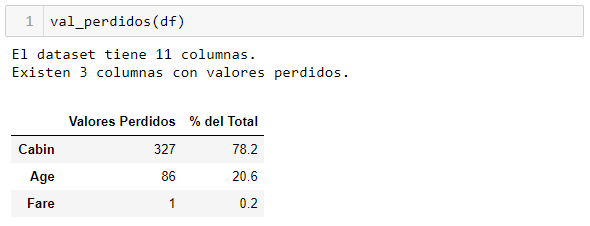
# Integracion y selección de los datos de interés a analizar

Explicar porque son importantes los datos seleccionados, y decir cuales no tenemos encuenta

# Limpieza de los datos

## Los datos contienen ceros o elementos vacíos?

Analizando en nuestro dataframe encontramos que existen tres columnas con valores perdidos.



Los datos correspondiente a cabina representan un 78.2% del total de los datos, sin embargo, no todos los pasajeros disponían de un camarote, por lo tanto, tiene sentido esa gran cantidad de datos perdidos.

Comenzando con la edad, mostramos con un histograma el numero de supervivientes por rango de edad.

## 

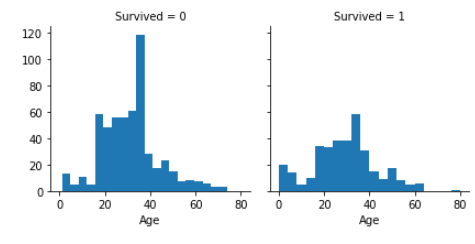
Con este resultado debemos considerar la columna edad para llegar a algun tipo de conclusion. Además, dado la importancia de esta columna, tenemos que buscar por datos perdidos y completar, en el caso que falten. Podríamos considerar tres métodos para completar una característica numérica continua como lo es la edad.

Una de las posibles formas de completar los datos que faltan sería generar números aleatorios entre la media y la desviación estándar de los datos disponibles.

Una forma más precisa de terminar de completar los valores perdidos es utilizar otras funciones correlacionadas. En nuestro caso, para la columna genero, se ha observado una diferencia estadística significativa entre las dos poblaciones. Además, en nuestro estudio estadístico hemos que el gráfico muestra que la mayor correlación la encontramos entre las variables Pclass y fare, como es intutivo pensar que a mayor precio del bilete, mejor la clase. En nuestro estudio estadístico, también hemos observado que en las correlaciones con la clase Survived, vemos que los factores mas correlacionados son el genero y el título de la persona, que entre otros cosas refleja datos del género y la edad.

Entonces para completar los datos de la edad, usaremos números aleatorios entre la media y la desviación estándar, basados en la clase y el genero.

Una vez completada la columna edad, volvemos a realizar el histograma.



Vemos en el grafico anterior, resultado de haber completado la columna edad, como se mantiene la distribución de los datos para los diferentes rangos de edad.

## Identificación y tratamiento de valores extremos

Text texttext

# Análisis de los datos

## Selección de los grupos de datos que se quieren analizar

Text texttext

## Comprobacion de la normalidad y homogeneidad de la

## varianza.

Text texttext

## Aplicacion de pruebas estadisticas para comparar los grupos de datos.

Text texttext

# Representación de los resultados a partir de tablas y gráficas.+

Text text text

# Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones?

# ¿Los resultados permiten responder al problema?