*Practica 2 – Tipología y Ciclo de Vida de los Datos*

# 1. Descripción del Dataset

## Introducción

Para el desarrollo de esta práctica se ha optado por la elección del dataset: "Titanic: Machine Learning from Disaster" que se encuentra en el link: "https://www.kaggle.com/c/titanic". Este dataset contiene información relacionada con uno de los naufragios más conocidos de la historia, donde se tienen datos relativos a sus pasajeros, como edad, sexo, clase en que viajaban y, finalmente, si han consiguieron sobrevivir o no. Es un dataset cuyo uso es muy extendido para el entrenamiento de algoritmos supervisados o para árboles de decisión donde la variable objetivo es precisamente si lograron sobrevivir o no en función de las características propias del viajero.

## Descripción de Columnas

Dicho conjunto de datos con 891 instancias no se puede considerar de un gran tamaño, sin embargo sí que está constituido por 12 columnas que hacen que la descripción de cada uno de los individuos sea razonablemente completa:

* PassengerId: id que contiene cada pasajero dentro del dataset;
* Survived: variable que nos dice si el pasajero ha sobrevivido, valor 1, o si finalmente ha muerto, valor 0;
* Pclass: clase en la que viajaba el pasajero;
* Name: nombre del pasajero;
* Sex: sexo del pasajero;
* Age: edad del pasajero;
* SibSp: número de hermanos y/o conyugues del pasajero a bordo;
* Parch: número de parientes y/o hijos/hijas del pasajero a bordo;
* Ticket: número del ticket del pasajero;
* Fare: precio pagado por el pasajero;
* Cabin: cabina en la que se encontraba el pasajero;
* Embarked: puerto de embarcación.

## Importancia y problemas para responder

A partir del análisis de este conjunto de datos, se pretende dar respuesta a una serie de preguntas que envuelven el accidente del titanic. En este caso, pretendemos determinar si, efectivamente, podemos decir que las mujeres y niños tenían una mayor probabilidad de haber sobrevivido o si el hecho de viajar en primera clase aportaba mayores posibilidades de supervivencia.

Este análisis nos permitirá entender de qué forma afectaba la clase social o género del pasajero a la hora de decidir las condiciones sobre las que escapaban del conocido hundimiento de Titanic.

# 2. Integración y selección de los datos de interés a analizar

## Lectura Archivo

El primer paso antes de poder realizar cualquier análisis consiste en la lectura del archivo o archivos de estudio. En nuestro caso, tenemos el archivo "train.csv" y leeremos sus datos a través de la función read.csv().

Una vez obtenidos los datos, observamos su estructura con str() y obtenemos un resumen de los valores con summary():

## Selección de datos

El siguiente paso, consistirá en eliminar aquellas columnas que no contengan información útil para el desarrollo de esta práctica. Entre estas columnas tenemos:

* Ticket: ya que no contiene información que pueda diferenciar los pasajeros;
* PassangerId: al ser simplemente un identificador del pasajero en nuestro conjunto de datos;
* Cabin: ya que consiste en una variable con un gran número de valores incompletos.
* Name

Ahora, convertiremos las variables categóricas de forma a facilitar el posterior análisis. Entre las variables que convertiremos a categóricas tenemos: Survived, Pclass, Sex y Embarked.

# 3. Limpieza de los datos

### 3.1.1. ¿Contienen los datos ceros o elementos vacíos?

Lo primero de lo que tenemos que hablar es de los ceros en nuestro dataset. Hay varias columnas en las que las que dicho valor tiene mucho sentido, luego no todo cero en nuestro enemigo.

* Comenzando por “Survived”, un 0 es un valor FALSE indicando que no sobrevivió, luego en esta columna no queremos mirar.
* En “PClass” y en “sex” sabemos que ninguno de los valores es cero porque los hemos categorizado y podemos ver que en ninguno de los niveles aparece dicho valor, en cambio en “Embarked” sabemos que esto sí que sucede precisamente por el mismo motivo.
* Un valor 0 en “SibSp” y en “Parch” es un valor absolutamente razonable (que el pasajero no tenga hermanos o parientes a bordo), luego tampoco nos interesa buscar.
* En cambio que un ticket haya sido gratis o que el pasajero tenga 0 años sería muy sospechoso, así que nos centraremos en estas dos columnas para la búsqueda de valores 0.

En cambio sí que queremos buscar elementos nulos en todas aquellas variables que no son categóricas, es decir en “Age”, “SibSp”, “Parch” y “Fare”. Luego para esta segunda cuestión nos centraremos en esas.

Comenzando por los ceros, podemos ver gracias a la función creada “where.ceros”, que no hay ninguna edad con dicho valor, en cambio sí que vemos que muchos tickets han sido gratuitos y trataremos con ellos en el siguiente epígrafe.

En cuanto a los valores NA, vemos que estos sólo se encuentran en la variable Age y además son extremadamente habituales, suponiendo 177 de los 891 registros de los que disponemos, algo que sin duda afectará a la decisión que decidamos tomar en el siguiente epígrafe.

Tan sólo nos queda por hablar de la variable “Embarked”, que como mencionábamos presenta un nivel vacío en el que imputaremos NAs que posteriormente habrá que tratar.

### 3.1.2 ¿Cómo gestionar los casos?

Comenzando por los ceros, nos gustaría saber cómo se distribuyen los registros según la clase antes de tomar una decisión, así que antes de tomar una decisión vamos a observar dicha condición.

Habíamos valorado la posibilidad de que el billete gratis fuera algún tipo de beneficio de alguna de las clases y que estuviéramos eliminando esta información, pero dado que se distribuyen casi homogéneamente por clase (5, 6 y 4), no parece que el valor pueda ser tomado por cierto. Dado que el precio de un ticket es algo que está fuertemente marcado por la clase, procedemos a imputar el precio medio condicionado a la clase en cada uno de esos ceros.

El caso de los valores NA en edad es bastante complejo. Son demasiados registros como para eliminarlos, pero a su vez no podemos imputar un valor único dado que con semejante volumen estaríamos distorsionando los resultados de eventuales análisis. Por lo que vamos a imputar los valores mediante el paquete caret, que puede funcionar realmente bien en este tipo de conjuntos. Para realizarla vamos a considerar todas las variables de las que disponemos en la creación de un modelo que nos ayudará a predecir valores posibles para esos NA. Posteriormente podremos comprobar si los resultados son razonables a partir de una comparación de las distribuciones de los registros que presentaban NAs y aquellos que no.

En el caso de “Embarked” vamos a proceder de forma parecida al primer punto. Por lógica el valor de la puerta de entrada va a estar fuertemente influenciado por el precio del ticket y la clase del pasajero, así que vamos a ver a través de que puerta embarcaron aquellos con una situación similar a los registros 62 y 830 (aquellos faltantes).

Resulta que aquellos que pagaron de media 80 libras y estaban en primera clase entraron por la puerta C de forma clara, así que vamos a imputar este valor para nuestros registros faltantes y a recalcular los niveles del factor para eliminar el “”.

## 3.2 Identificación y tratamiento de valores extremos

De nuevo vamos a comenzar por plantearnos qué dimensiones pueden presentar valores extremos, que en este caso sólo son las numéricas, es decir: “Age”, “SibSp”, “Fare” y “Embarked”.

A través de la función creada para detectar outliers, vemos que efectivamente el dataset presenta muchos, pero antes de proceder a una imputación debemos comprobar cualitativamente si efectivamente estos valores se pueden considerar outliers, para lo que procedemos a observar los diagramas de caja correspondientes a estas dimensiones.

Vemos que todos los valores extremos quizá a excepción de uno tienen sentido. Tenemos que alguien ha pagado más de 500 libras por una habitación cuando el segundo valor más grande era de menos de 300. Viendo esta circunstancia y la más que segura existencia de información online, hemos hecho una pequeña investigación sobre dicha habitación y resulta que su valor es correcto dado que era una suite triple en la que Thomas Cardeza convivió con su madre y varios sirvientes[[1]](#footnote-1). Debido a estas observaciones, no parece que vaya a ser necesario tratar los outliers.

# 4. Análisis de los datos

## 4.1. Selección de los grupos de datos que se quieren analizar o comparar

Todos conocemos la mítica frase “mujeres y niños primero” gracias a la película homónima, lo que no se decía tan claro en la película es que los ricos también iban primero. Está claro que uno de las selecciones tiene que ser una división de supervivientes y fallecidos, la segunda tiene que ser por sexo y la tercera será por clase. Podríamos categorizar la edad para separar también por grupos, pero corremos el riesgo de perder mucha información al establecer franjas de edad así que no lo vamos a hacer.

Vamos a comprobar la normalidad utilizando el test de shapiro-wilk. Este test consiste en un contraste de hipótesis en el que la hipótesis nula es la distribución normal de los datos. Dado que en nuestro caso los valores p resultantes del análisis para todos los conjuntos numéricos es menor que 0.05, tenemos que rechazar la hipótesis nula y por tanto no se puede confirmar la normalidad de los datos.

1. <https://www.encyclopedia-titanica.org/titanic-survivor/thomas-cardeza.html> (Consultado 20/05/2021). Esta historia también nos sirve para comprobar que el dataset no es de los más completos, la madre de Thomas Cardeza no aparece como viajera y sin embargo sí que aparece la sirvienta de ella Miss. Anna Ward. [↑](#footnote-ref-1)