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

# Descripción del Dataset

## Introducción

Para el desarrollo de esta práctica se ha optado por la elección del dataset: "Titanic: Machine Learning from Disaster" que, a su vez, se puede encontrar en el link: "https://www.kaggle.com/c/titanic". Este dataset, contiene información relacionada con uno de los naufragios más conocidos de la hisotia, donde se tienen datos relativos a sus pasajeros, como su edad, sexo, la clase en que viajaban y, finalmente, si han podido sobrevivir o no.

## Descripción de Columnas

Dicho conjunto de datos contiene un total de 891 observaciones constituidas por 12 columnas:

* 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;
* Emabarked: 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, si al ser de clase alta el pasajero tenía una mayor probabilidad de sobrevivir.

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.

# 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.