Descripción del dataset.

El juego de datos elegido para la realización de esta práctica es el propuesto en el enunciado de la misma, sobre el conjunto de pasajeros del Titanic.

Desde la página es posible descargar ficheros:

* train.csv. Contiene el conjunto de datos “a entrenar”
* test.csv. Contiene el conjunto de datos para validar los modelos generados durante el entrenamiento. Contiene las mismas columnas que el dataset train.csv pero sin la variable que indica si el pasajero sobrevivió o no.
* gender\_submission.csv. Contiene un conjunto de datos en el que solo se informa del identificador del pasajero y si este ha sobrevivido o no.

A continuación se explican brevemente los campos del fichero “train.csv”.

1. PassengerId -> Identificador numérico único del pasajero.
2. Survived -> 0 si no sobrevivió, 1 si sobrevivió.
3. Pclass -> 1 si es primera categoría, 2 si segunda, 3 si tercera.
4. Name -> Nombre del pasajero
5. Sex -> “Male” si era hombre, “Female” si mujer
6. Age -> Atributo numérico con la edad del pasajero. Si es desconocida y, por tanto, estimada, se informa en el formato “XX.5”.
7. SibSp -> Atributo numérico con el número de hermanos más esposo/a a bordo.
8. Parch -> Atributo numérico con el número de padres o hijos a bordo.
9. Ticket -> Identificador alfanumérico del billete.
10. Fare -> Tarifa del billete
11. Cabin -> Número de cabina ocupada
12. Emabarked -> Puerto donde embarcó (C = Cherbourg, Q = Queenstown, S = Southampton)

El principal objetivo de este dataset es responder a la pregunta ¿qué características hacían más probable que un pasajero sobreviviera?

Es posible que este ejemplo no sea el mejor para hacer predicciones sobre la posibilidad de sobrevivir en otro posible naufragio. Las características del barco pueden ser diferentes y, por ejemplo, que las cabinas de cierta clase estén en una posición que dificulte la supervivencia mientras que en este caso la facilitaran. No obstante, algunas propiedades sí que pueden resultar de interés y se tratarán de estudiar en capítulos posteriores.

Integración y selección de los datos de interés a analizar.

Dado que disponemos de tres ficheros, vamos a realizar una tarea de integración para tener todos los datos en un solo dataset.

Como hemos indicado, en el conjunto de test no se proporciona el atributo “survived”. Sin embargo, se puede obtener del fichero “gender\_submission”.

Por tanto, será posible crear un único juego de datos con todos los pasajeros, sus características y su información de si sobrevivió o no.

Se puede ver el código y la ejecución en el fichero adjunto, pero se muestran algunas capturas a continuación.

En el dataset resultante, tenemos 1309 observaciones con 12 propiedades.

Por trabajar un poco más con la manipulación de datos, vamos a agrupar las columnas de “SibSp” y “Parch” en una nueva columna “Relatives”, que sea la suma de los valores de las columnas anteriores.

Limpieza de los datos.

Vamos a revisar si existen valores nulos para cada columna y ver el tratamiento que le debemos dar en cada caso.

Tras realizar la comprobaciones anteriores vemos que tenemos:

263 casos en los que no disponemos de la edad

1014 casos en los que no disponemos del número de cabina

1 caso en el que no disponemos de la tarifa

2 casos en los que desconocemos el puerto de embarque.

En todos los casos vamos a interpretar que simplemente son valores desconocidos; por ejemplo, no disponer de tarifa, no necesariamente implica que la tarifa era 0 (situación para la que también existen algunos registros), sino que no disponemos del dato. Podemos intentar estimarlo en función de la clase y el puerto de embarque, ya que parece que la tarifa y la clase guardan relación.

Por tanto, vemos que es un registro con clase 3 y embarque Southampton. Obtenemos la media de los datos de todos los registros con esas características y lo asignamos.

Para el resto de registros, no haremos nada y los dejaremos con NA

Los valores outliers o extremos son aquellos que llaman la atención por su evidente diferencia con respecto a la mayoría de datos de otros registros. Vamos a analizar con diagramas de cajas los atributos para intentar encontrar los valores outlier de cada propiedad numérica (tarifa, edad y familiares a bordo).

Como podemos comprobar, no existen datos realmente anómalos (las edades son razonables todas y las tarifas, salvo por el hecho de que algunas son gratis y tres muy elevadas no llaman demasiado la atención).