1. 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 de Kaggle (<https://www.kaggle.com/c/titanic>) es posible descargar los siguientes 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.

2. 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.

Texto

Descripción generada automáticamente con confianza media

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.

Texto

Descripción generada automáticamente

Finalmente, éste es el dataset que obtenemos y con el que vamos a trabajar.

Texto, Carta

Descripción generada automáticamente

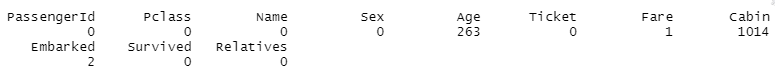
3. Limpieza de los datos.

3.1 Elementos vacíos

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

Texto

Descripción generada automáticamente con confianza baja



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.

Vamos a intentar completar nuestro dataset con valores estimados para la tarifa y la edad en los registros con datos desconocidos. En todos los casos vamos a interpretar que simplemente son valores desconocidos, es decir, 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.

En el caso de la tarifa únicamente tenemos un registro desconocido. 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.

Texto

Descripción generada automáticamente

Para la edad tenemos 263 por lo que ir uno a uno implicaría demasiado tiempo. En este caso, vamos a agrupar los datos en función de la clase y el número de familiares embarcados y vamos a calcular la media de edad de estos grupos. Esta media será la que asignemos a los valores desconocidos en función del grupo al que pertenecen.

Texto

Descripción generada automáticamente

El número de cabina y el puerto de embarque no tienen una relación directa con ninguna de las columnas restantes, por lo que en principio no es posible estimar su valor. Por lo tanto, no haremos nada con ellos y los dejaremos con NA.

3.2 Outliers

Los outliers o valores 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).

Captura de pantalla de un celular

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Imagen que contiene Diagrama

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Como podemos comprobar, aunque sí que hay valores que no están dentro de los valores más “normales” de estos atributos, no existen datos realmente anómalos. Para las edades y el número de familiares, los valores que vemos más alejados de los valores medios son razonables y no tendrían por qué se incorrectos. Respecto a las tarifas, sí que vemos valores mucho más elevados de los normal. Sin embargo, analizando los datos, vemos que corresponden a pasajes de primera clase por lo que es razonable pensar que también son correctos. Además, hay algunos valores a 0 que asumiremos como invitaciones.

4. Análisis de los datos

4.1 Selección de grupos de datos

Lo primero que vamos a realizar es una discretización de los datos relativos al sexo, el puerto de embarque, la clase del billete y la supervivencia, que, aunque vienen representados en enteros o cadenas, en realidad pertenecen a grupos que deben ser factorizados.

Texto

Descripción generada automáticamente

La estructura de los datos queda de la siguiente manera.

Texto, Carta

Descripción generada automáticamente

En general, lo que vamos a querer analizar es la relación que existe entre los pasajeros que sobreviven con el resto de atributos. Queremos comprobar si existen ciertas características que hacen que los pasajeros tengan más probabilidades de sobrevivir. En concreto, en los próximos apartados vamos a analizar si el sexo, la clase del billete, la edad o la tarifa son determinantes para decir que un pasajero sobrevivió.

4.2. Comprobación de normalidad y homogeneidad de la varianza

La normalidad y la homogeneidad de la varianza únicamente tienen sentido con variables numéricas, por lo que vamos a utilizar la edad y la tarifa.

Con el objetivo de verificar la suposición de normalidad, vamos a realizar el test de Shapiro-Wilk.



Texto

Descripción generada automáticamente

Asumimos como hipótesis nula que la población está distribuida normalmente. En ambos casos, obtenemos p-values muy bajos, mucho menores que 0.05 que podríamos tomar como nivel de significancia, por lo que vamos a rechazar la hipótesis nula, es decir, no podemos asumir normalidad.

Para la homocedasticidad (igualdad de varianza entre dos grupos), vamos a comprobar la edad y la tarifa para los grupos que sobrevivieron y los que no.

