# JAVIER MARTÍNEZ DELGADO

# Práctica 2-Tipología

## 1.-Descripción del dataset

Para esta práctica he elegido el dataset del Titanic que se encuentra en la página web de Kaggle debido a que es uno de los datasets más famosos para principiantes en ciencia de datos ya que contiene variables categóricas y variables numéricas para aprender a tratarlas correctamente. El dataset ofrece una serie de variables como por ejemplo el identificador del camarote de la persona, si era hombre o mujer, la edad, clase social y varias más con la finalidad de poder predecir si dada unas circunstancias concretas, esa persona sobrevivió o no al hundimiento del Titanic, por lo que será un problema de clasificación con dos posibles categorías.

## 2.-Integración y selección de los datos

Desde la propia página de Kaggle podemos descargar el fichero de entrenamiento y el de test sin ningún problema. Deberemos eliminar alguna de las variables que contiene ya que no aportan ninguna información que se pueda llegar a extraer como el Id del pasajero, el número del ticket, el nombre (de donde extraeremos el prefijo) y el camarote (de donde extraeremos la letra del camarote para determinar en qué cubierta se encontraba el pasajero) por lo que trabajaremos con las siguientes variables:

-Survived: Variable target (sólo se encontrará en el conjunto de test). Indicará si el pasajero en cuestión sobrevivió o no sobrevivió al hundimiento.

-Pclass: Status al que pertenecía el pasajero (primera clase, segunda o tercera)

-Name: Prefijo del nombre del pasajero (Actuará como indicador de sexo y/o clase social de la persona)

-Sex: Indica si el pasajero era hombre o mujer

-Age: Edad del pasajero

-SibSp: Indica si el pasajero tenía hermanos o pareja a bordo y cuántos

-Parch: Indica si el pasajero tenía Padres o hijos a bordo y cuántos

-Fare: Precio del ticket del pasajero

-Cabin: Letra del camarote

-Embarked: Puerto desde el que embarcó el pasajero

# 3.-Limpieza de los datos

Al analizar los datos podemos observar que no hay ceros, pero sí que hay elementos vacíos en las variables Age, Cabin, Embarked y Fare. Age es una variable numérica y hay datos faltantes tanto en el conjunto de entrenamiento como en el de test por lo que hemos creado una nueva variable indicando en qué registros se encontraban los valores faltantes y posteriormente hemos imputado con la mediana ya que la media se ve un poco desplazada al no tratarse de una distribución normal ya que es bimodal.

La otra variable numérica es Fare y sólo presenta valores faltantes en el conjunto de test por lo que he imputado los datos con la mediana ya que la variable presenta una distribución asimétrica.

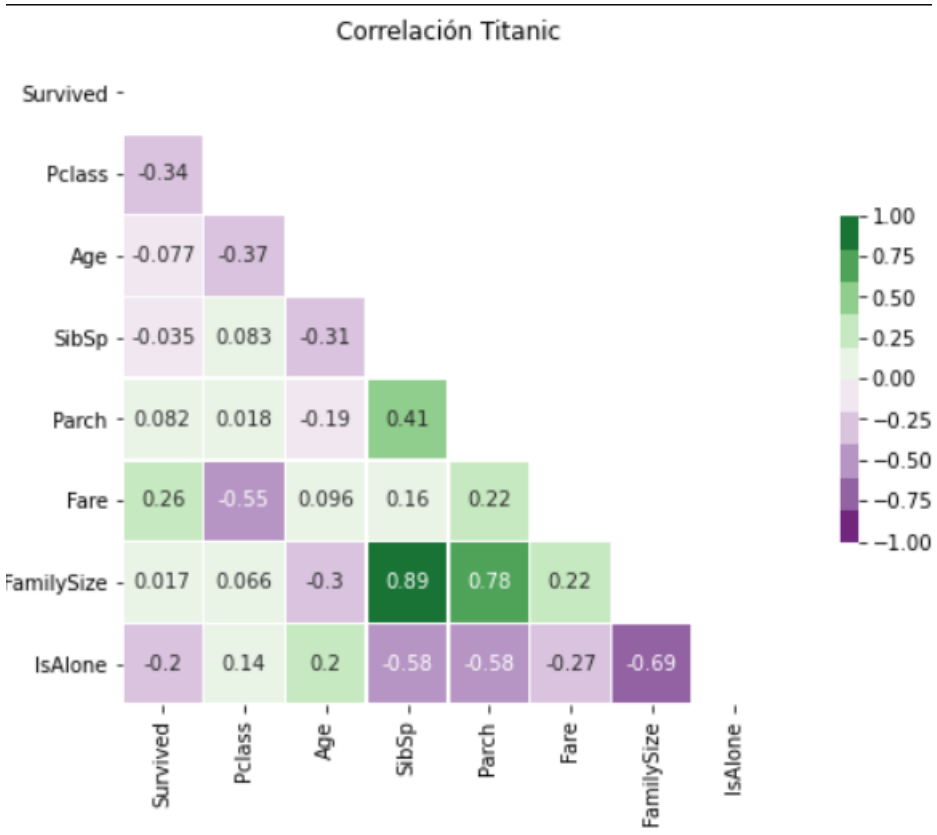
La variable Cabin es categórica y presenta valores faltantes en ambos conjuntos por lo que primero creamos una nueva variable indicadora como en el caso de la variable Age y luego imputamos los datos creando una nueva categoría denominada missing.

Por último, la variable Embarked presenta pocos valores faltantes en el conjunto de entrenamiento por lo que he buscado registros similares para imputar los datos con la moda de los registros similares.

En cuanto a los valores extremos, al hacer el boxplot de las dos variables numéricas se puede ver que hay outliers por lo que los he analizado para ver si correspondían con valores erróneos, algún valor centinela o si se tenían que tener en cuenta. En la variable Age no había nada raro, simplemente la distribución de la variable hacía que esos registros fuesen resaltados como outliers. En la variable Fare los outliers se encuentran muy alejados de la distribución pero no hay nada que nos indique que podría ser un valor centinela o un dato erróneo ya que corresponde a personas de clase alta, supervivientes y que probablemente pagaron mucho por sus billetes por lo que no se han eliminado estos registros

# 4.-Análisis de los datos

Todavía no tengo hecha esta parte pero tengo pensado hacer pruebas estadísticas sobre los resultados obtenidos en el ejercicio 5




La selección de los datos grupos de datos a analizar será examinar cada variable y ver cómo se relaciona con la variable target, en este caso, la variable Survived.

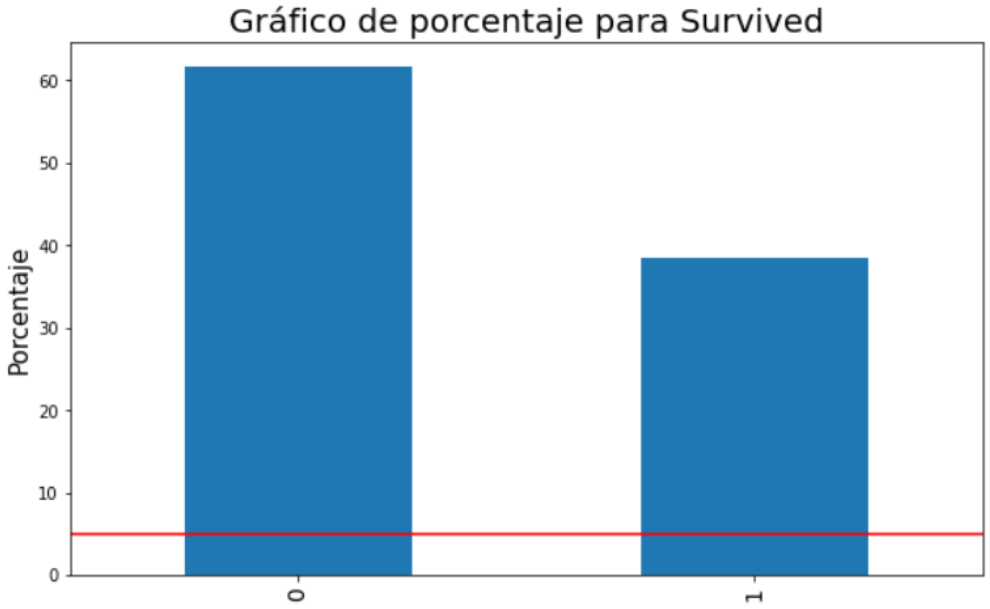
Al realizar el test de Lilliefors para ambas variables numéricas obtenemos que el p-valor es mucho menor que cualquier nivel de significación que utilicemos, aunque normalmente se utiliza un nivel de significación del 5%. El hecho de que el p-valor sea menor que el nivel de significación nos indica que debemos rechazar la hipótesis nula de que la variable sigue una distribución normal y aceptar la hipótesis alternativa de que no siguen una distribución normal.

Para realizar el test de heterocedasticidad he utilizado el test Fligner-Killeen ya que como hemos podido ver no se puede asumir la normalidad de los datos por lo que se tiene que realizar un test no paramétrico. Al realizar el test con y sin escalar, podemos ver que al escalar los datos la varianza de las distintas variables es casi la misma.

# 5.-Representación de los resultados a partir de tablas y gráficas

-Survived

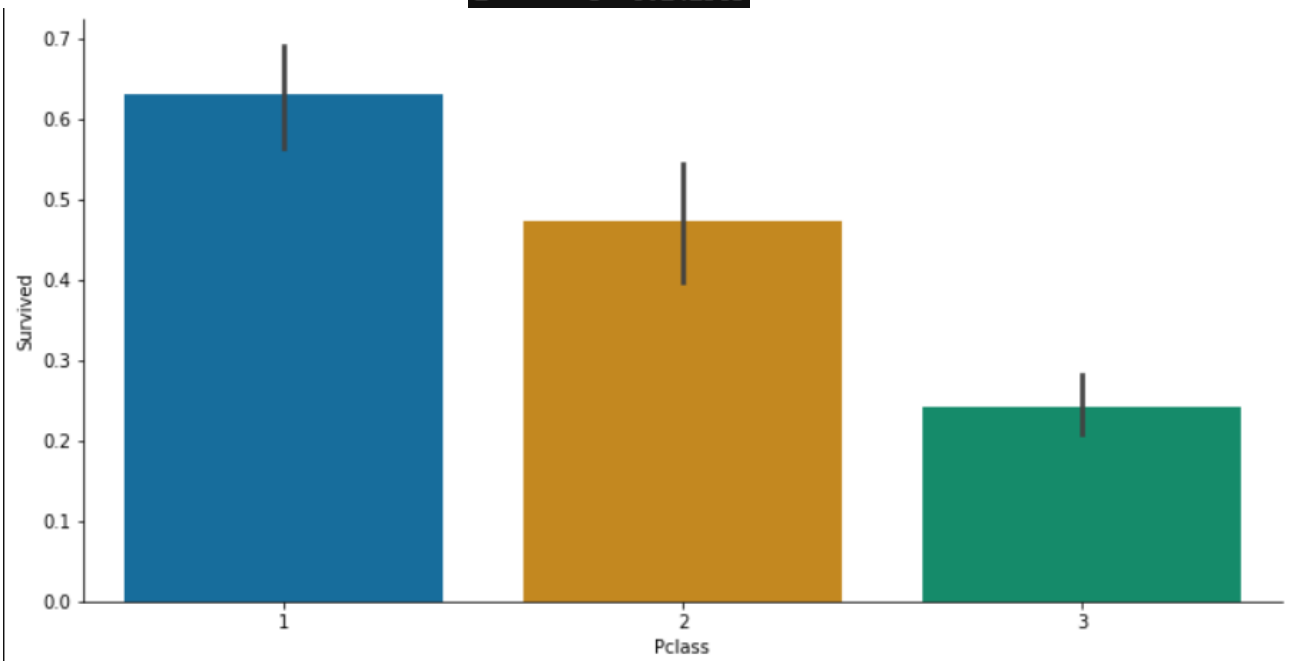
| | Total | Percent |
|---|---|---|
| 0 | 549 | 61.616162 |
| 1 | 342 | 38.383838 |




Podemos apreciar que sólo el 38.39% de las personas sobrevivieron al hundimiento.

-Pclass

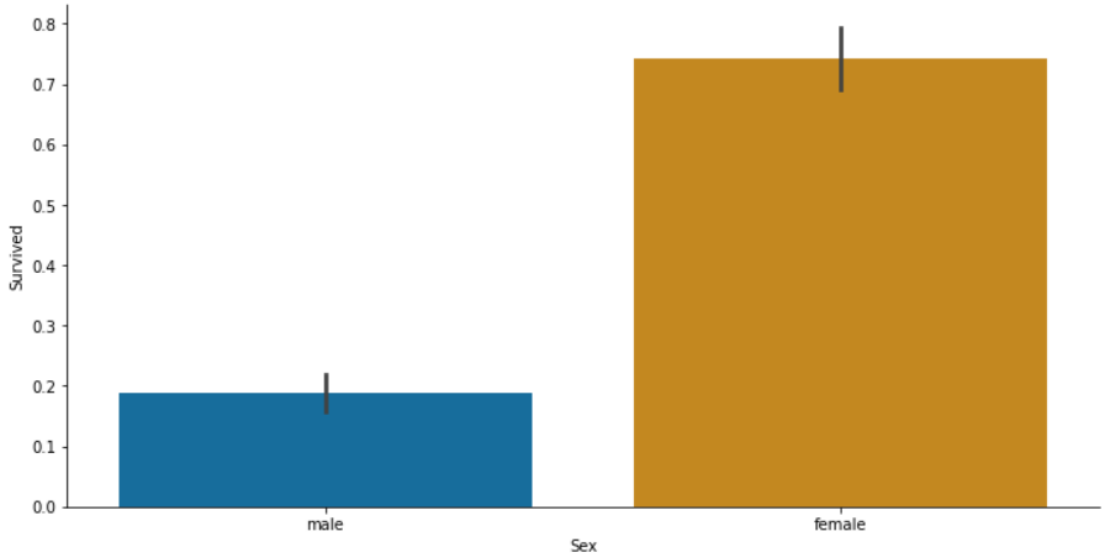
| | Pclass | Survived |
|---|---|---|
| 0 | 1 | 0.629630 |
| 1 | 2 | 0.472826 |
| 2 | 3 | 0.242363 |



A medida que la clase social aumentaba, aumentaba también la probabilidad de supervivencia de los pasajeros

-Sex

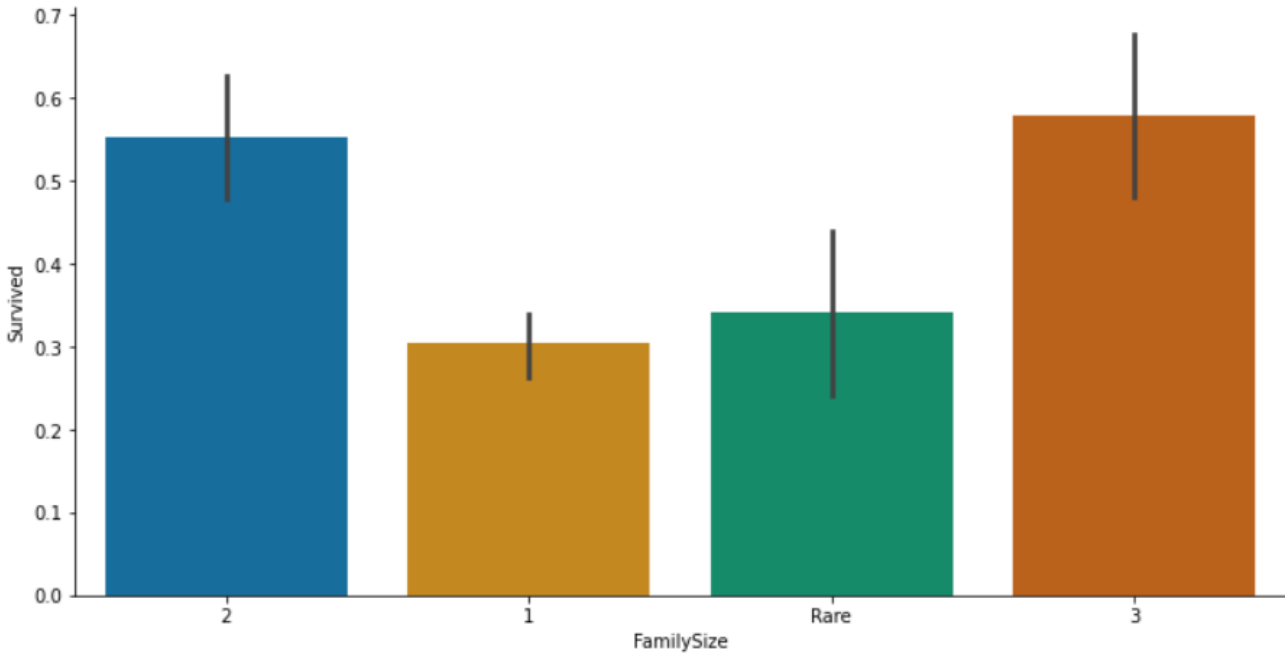
| | Sex | Survived |
|---|---|---|
| 0 | female | 0.742038 |
| 1 | male | 0.188908 |




La probabilidad de supervivencia era mucho mayor si el pasajero era mujer

-FamilySize

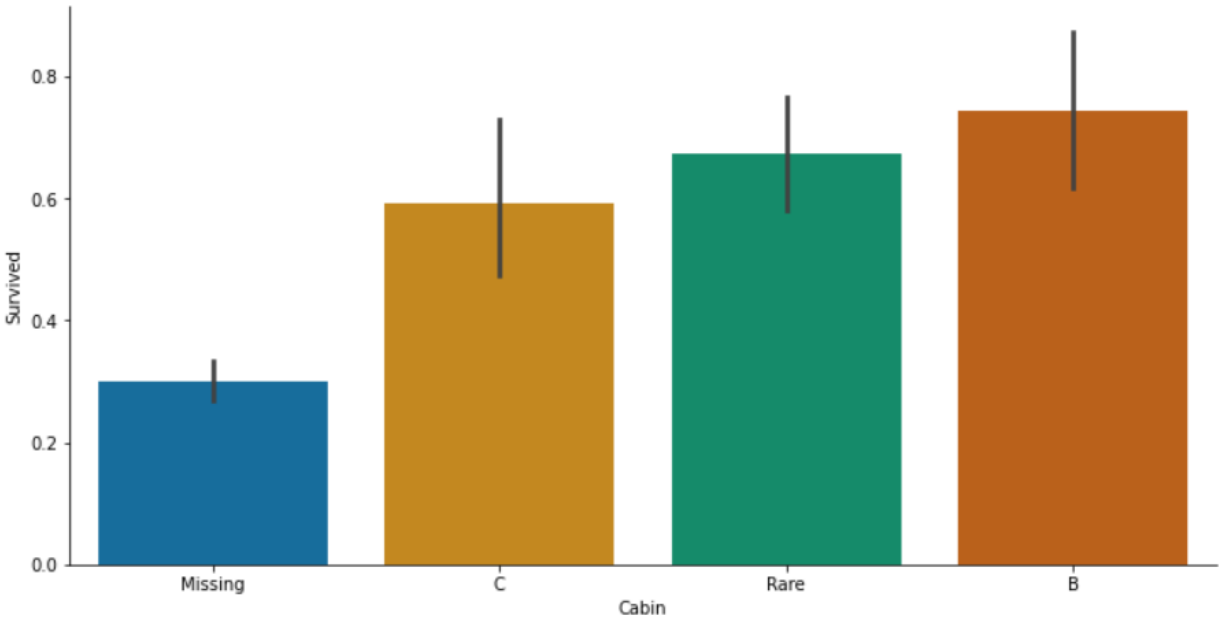
| | FamilySize | Survived |
|---|---|---|
| 0 | 1 | 0.303538 |
| 1 | 2 | 0.552795 |
| 2 | 3 | 0.578431 |
| 3 | Rare | 0.340659 |



Esta es una variable curiosa ya que si la familia estaba compuesta por 2 o 3 miembros la probabilidad de supervivencia aumentaba mientras que si el pasajero viajaba sólo o si la familia estaba compuesta por más de 3 miembros, disminuía

-Cabin

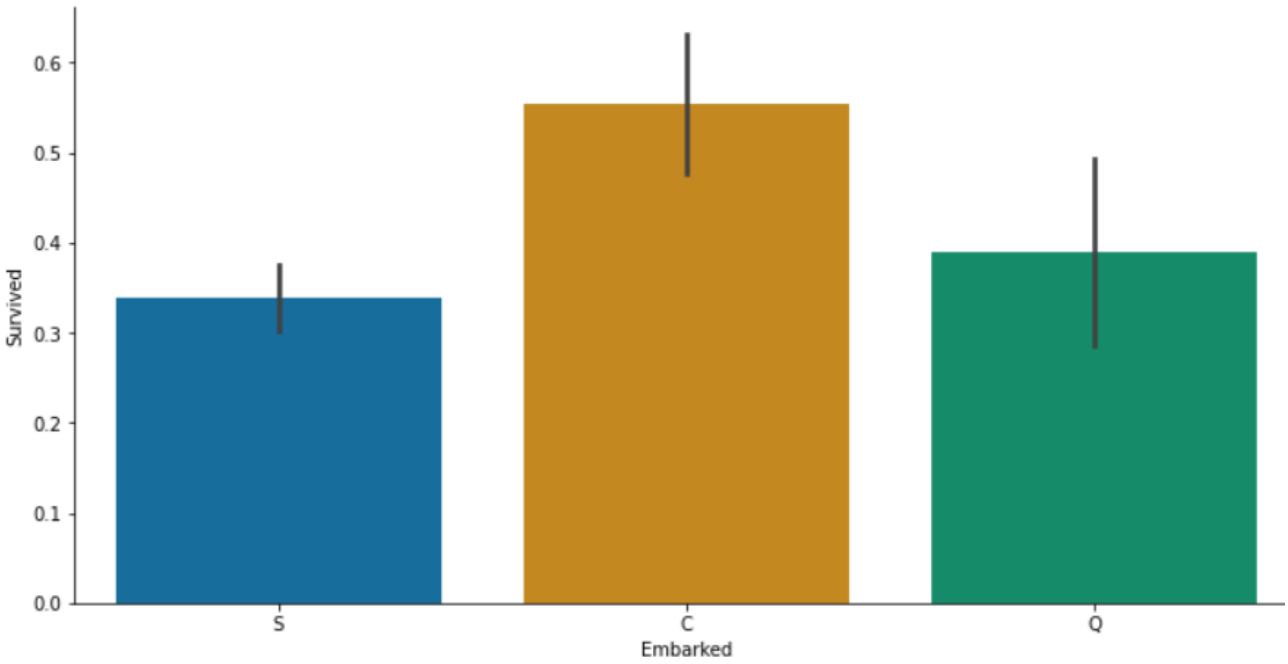
| | Cabin | Survived |
|---|---|---|
| 0 | B | 0.744681 |
| 1 | C | 0.593220 |
| 2 | Missing | 0.299854 |
| 3 | Rare | 0.673469 |




Podemos ver que aquellos registros en los que teníamos valores faltantes, tuvieron una proporción mucho menor de supervivencia por lo que puede estar relacionado

-Embarked

| | Embarked | Survived |
|---|---|---|
| 0 | C | 0.553571 |
| 1 | Q | 0.389610 |
| 2 | S | 0.339009 |




Los pasajeros que embarcaban en Cherbourg tenían una probabilidad de supervivencia mayor que los demás pasajeros.

# 6.-Resolución del problema

Se ha resuelto el problema con aproximadamente un 83% de precisión. Se ha entrenado un modelo de random forest, se ha optimizado los hiperparámetros del modelo y se ha llevado a cabo la predicción del conjunto de test para determinar si una serie de pasajeros pudieron sobrevivir a la catástrofe. Este problema en concreto lo hice hace tiempo en kaggle obteniendo top 3% aunque actualmente me encuentro sobre el top 10%.

https://www.kaggle.com/arcticai/top-3-titanic-dataset

# 7.-Código

El código se encuentra dentro de la carpeta código del repositorio de github