PRÁCTICA 2 - TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

DIEGO CONTRERAS JIMENEZ

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Titanic: Machine Learning from Disaster

Este conjunto de datos está disponible en la plataforma de competiciones de predicciones Kaggle, siendo una competición activa y simple para iniciarse en el mundo del "data science" y "machine learning".

El dataset contiene cierta información de los pasajeros y de la tripulación que viajaron en el Titanic y sufrieron el trágico accidente.

El objetivo del conjunto de datos es conocer qué tipos de personas tenían más probabilidades de sobrevivir. En el caso concreto de la competición, se pide utilizar técnicas de machine learning para predecir qué pasajeros sobrevivieron la tragedia.

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

Los datos vienen separados en dos ficheros (train.csv y test.csv), uno para el entrenamiento y otro para realizar las pruebas de test.

Ambos ficheros cuentan con los mismos atributos, a excepción de la variable objetivo o feature *Survived* que no está presente en el fichero de test.

Columna	Definición	Valores
PassengerId	Identificador único del pasajero	
Survived	Sobrevivió	0 = No, 1 = Sí
Pclass	Clase del Ticket	1 = Primera, 2 = Segunda, 3 = Tercera
Name	Nombre	
Sex	Sexo	male = Hombre, female = Mujer
Age	Edad en años	
SibSp	# de hermanos / parejas abordo del Titanic	
Parch	# de padres / hijos abordo del Titanic	
Ticket	Número de Ticket	
Fare	Tarifa de pasajero	
Cabin	Cabina	
Embarked	Puerto de embarque	C = Cherbourg, Q = Queenstown, S = Southampton

La mayoría de campos tienen relación y pueden aportar información para el modelo que se trata de construir, por lo que nos interesa tenerlos en consideración para su posterior análisis.

Los atributos PassengerId y Ticket son valores aleatorios que no deberían ser relevantes para nuestro modelo, por lo que los excluiremos de nuestros análisis.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

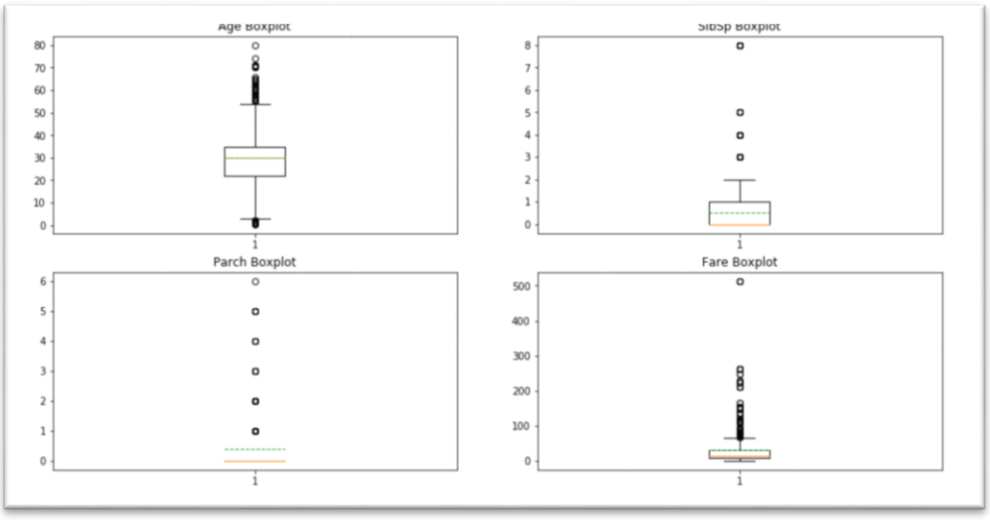
El conjunto de datos, tanto de entrenamiento como de test, contiene elementos vacíos en los siguientes atributos:

- Age: Es una variable continua, por lo que podemos usar la media para sustituir los elementos vacíos.
- Cabin: Este atributo podría darnos información añadida de la zona del barco o de la capacidad económica del pasajero, pero como tenemos demasiados valores a nulo es posible que no añada ningún valor extra y, en tal caso, sería descartado de los próximos análisis.
- Embarked: Es una variable cualitativa por lo que estableceremos un valor por defecto, como la moda, para los elementos que no venga informado este campo.
- Fare: Es otra variable continua, por lo que sustituiremos los elementos vacíos con la media.

Por otra parte, para cada uno de estos atributos que tenemos elementos vacíos podríamos añadir una columna booleana que indique si el elemento está informado o no. De esta forma no perderemos información de cuando un campo no tenía datos.

3.2. Identificación y tratamiento de valores extremos.

Gracias a la representación gráfica de boxplot, podemos detectar fácilmente si existen valores extremos en el conjunto de datos.



En el caso de la tarifa (Fare) hemos detectado un posible caso de valor extremo, pero explorando los datos vemos que son tres personas que viajan en primera clase, por lo que asumimos que esos valores puedan ser correctos.

Así pues, podemos determinar que en este conjunto de datos no existe ningún atributo que contenga valores extremos.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Los grupos de datos que se quieren analizar/comparar serán los siguientes:

- Age
- SibSp
- Parch
- Fare

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Hemos aplicado el test de Anderson-Darling para comprobar si los datos siguen una distribución normal.

Nos hemos basado sobre las variables anteriormente citadas y el resultado que hemos encontrado es que en ningún caso el valor de p es menor que el nivel de significancia ($p < 0.05$).

```
Age:
------
Statistic: 15.318
15.000: 0.573, los datos no son normales (rechazamos H0)
10.000: 0.653, los datos no son normales (rechazamos H0)
5.000: 0.784, los datos no son normales (rechazamos H0)
2.500: 0.914, los datos no son normales (rechazamos H0)
1.000: 1.087, los datos no son normales (rechazamos H0)

SibSp:
------
Statistic: 147.365
15.000: 0.573, los datos no son normales (rechazamos H0)
10.000: 0.653, los datos no son normales (rechazamos H0)
5.000: 0.784, los datos no son normales (rechazamos H0)
2.500: 0.914, los datos no son normales (rechazamos H0)
1.000: 1.087, los datos no son normales (rechazamos H0)

Parch:
------
Statistic: 175.659
15.000: 0.573, los datos no son normales (rechazamos H0)
10.000: 0.653, los datos no son normales (rechazamos H0)
5.000: 0.784, los datos no son normales (rechazamos H0)
2.500: 0.914, los datos no son normales (rechazamos H0)
1.000: 1.087, los datos no son normales (rechazamos H0)

Fare:
------
Statistic: 122.170
15.000: 0.573, los datos no son normales (rechazamos H0)
10.000: 0.653, los datos no son normales (rechazamos H0)
5.000: 0.784, los datos no son normales (rechazamos H0)
2.500: 0.914, los datos no son normales (rechazamos H0)
1.000: 1.087, los datos no son normales (rechazamos H0)
```

Para el análisis de la igualdad de las varianzas hemos empleado el test de Flinger-Killeen.

```
Age vs Fare: statistic 80.331, pvalue 0.000
Age vs SibSp: statistic 648.724, pvalue 0.000
Age vs Parch: statistic 698.022, pvalue 0.000
Fare vs SibSp: statistic 1017.054, pvalue 0.000
Fare vs Parch: statistic 1045.984, pvalue 0.000
SibSp vs Parch: statistic 8.778, pvalue 0.003
```

Puesto que hemos obtenido un p-valor inferior a 0.05 en todos los casos, rechazamos la hipótesis de que las varianzas de ambas muestras son homogéneas. Asi pues, podemos afirmar que las varianzas tienen una diferencia significativa entre sí.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

¿Es igual o diferente la media de edad de las personas que sobrevivieron al naufragio según el sexo?

En primer lugar, vamos a comparar las medias de las edades en el colectivo de personas que sobrevivieron al naufragio según sean hombres o mujeres.

Así planteamos el contraste de hipótesis de dos muestras sobre la diferencia de medias:

$H_0 : \mu_1 - \mu_2 = 0$ (La edad media de los sobrevivientes es igual en hombres que en mujeres)

$H_1 : \mu_1 - \mu_2 < 0$ (La edad media de los sobrevivientes es diferentes en hombres que en mujeres)

Hemos aplicado la prueba de contraste del t-test de Welch, que es una adaptación del t-test de Student, pero es más fiable para comparar dos muestras con tamaños y varianzas distintos. Aunque como vimos en el apartado anterior, la edad no seguía una distribución normal, como la muestra es superior a 30 podemos sí que podremos aplicar este test.

Los resultados obtenidos son los siguientes:

```
Edad hombres sobrevivientes: media 27.632, 15.258
Edad mujeres sobrevivientes: media 28.979, 13.033
t-statistic: -0.796, p value: 0.427
```

En este caso el p-value es mayor que el valor de significación fijado ($\alpha = 0.05$) por lo que no podemos rechazar la hipótesis nula y por lo tanto podremos decir que las medias son iguales.

¿Qué variables cuantitativas influyen más en el resultado que una persona sobreviva o no?

Para obtener qué variables influyen más en este resultado vamos a obtener la matriz de correlaciones y nos fijaremos en la variable objetivo Survived.

```
SibSp       0.035322
Age         0.069809
Parch       0.081629
Fare        0.257307
Pclass      0.338481
Survived    1.000000
Name: Survived, dtype: float64
```

Como podemos observar, la variable Pclass es la que mayor valor tiene y por lo tanto podemos decir que es la que más influye en el hecho que una persona haya sobrevivido al naufragio.

Modelo de regresión lineal

Por último, vamos generar un modelo con el que podamos predecir si una persona va a sobrevivir o no según una serie de características.

Para ello crearemos un modelo de regresión lineal y en primera instancia incluiremos únicamente las variables cuantitativas. El coeficiente de correlación R_2 y los coeficientes para los predictores del modelo son los siguientes:

```
R^2: 0.088
       Coefficient
Age      -0.004317
SibSp    -0.056920
Parch     0.032993
Fare      0.002709
```

En una segunda ejecución hemos añadido las variables categóricas para ver si mejora el modelo anterior. Hemos obtenido los siguientes resultados:

```
R^2: 0.395
             Coefficient
Age            -0.005816
SibSp          -0.042867
Parch          -0.020274
Fare            0.000385
Pclass         -0.167176
Sex             0.513233
Embarked        0.018309
```

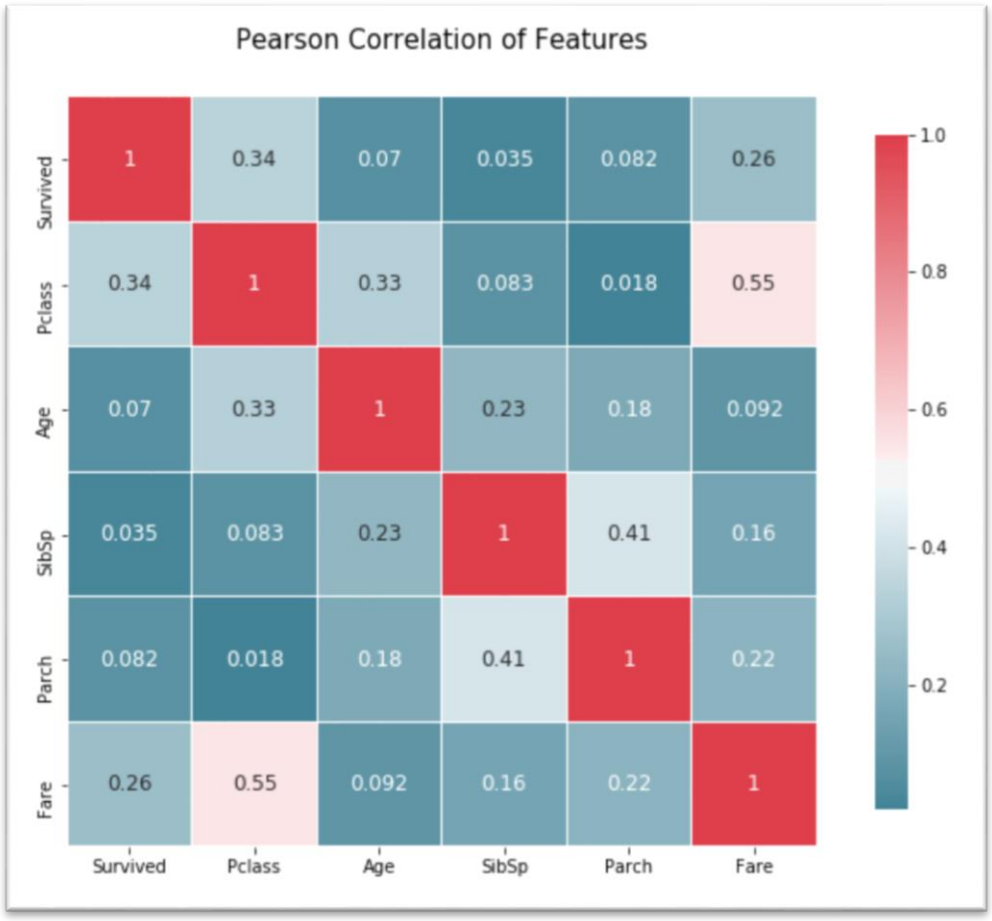
Como se puede observar, el resultado mejora considerablemente. Esta mejora se consigue gracias a las variables Sex y Pclass que son las que tienen un coeficiente mayor.

Para valorar la precisión de este último modelo, primero debemos normalizar los resultados de las predicciones en valores booleanos, ya que las predicciones que nos aporta el modelo son de tipo decimal. Así pues, simplemente redondearemos el valor de los resultados para obtener unas predicciones con valor 0 o 1.

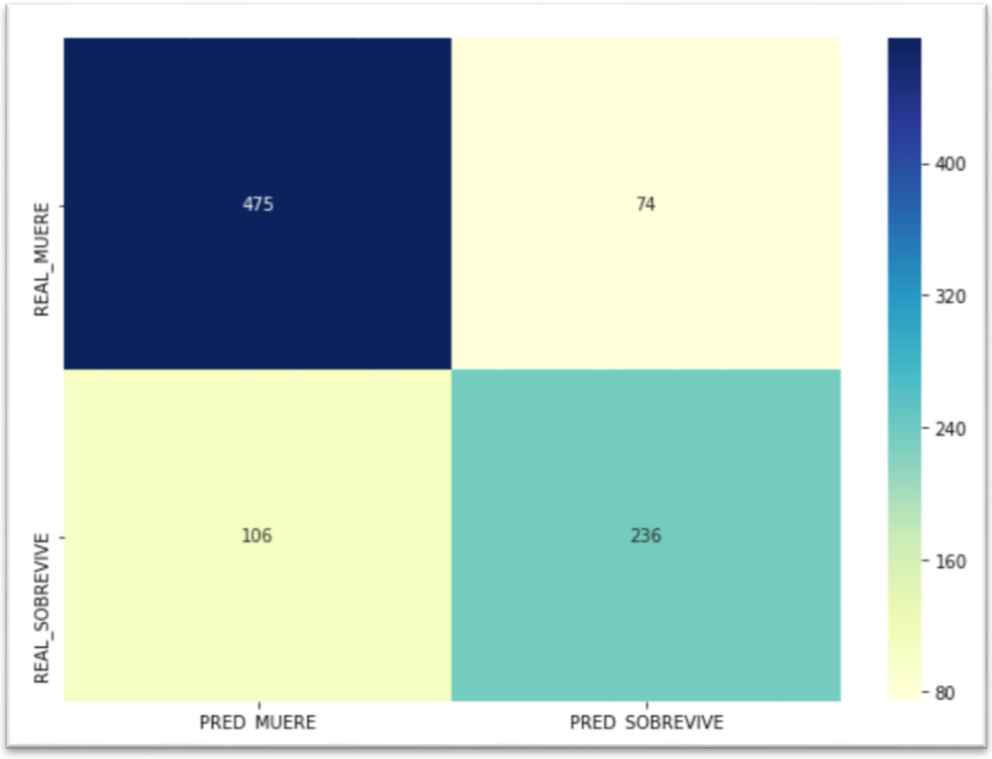
Una vez hecho esto, calcularemos la precisión y obtendremos una puntuación de 0.798.

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

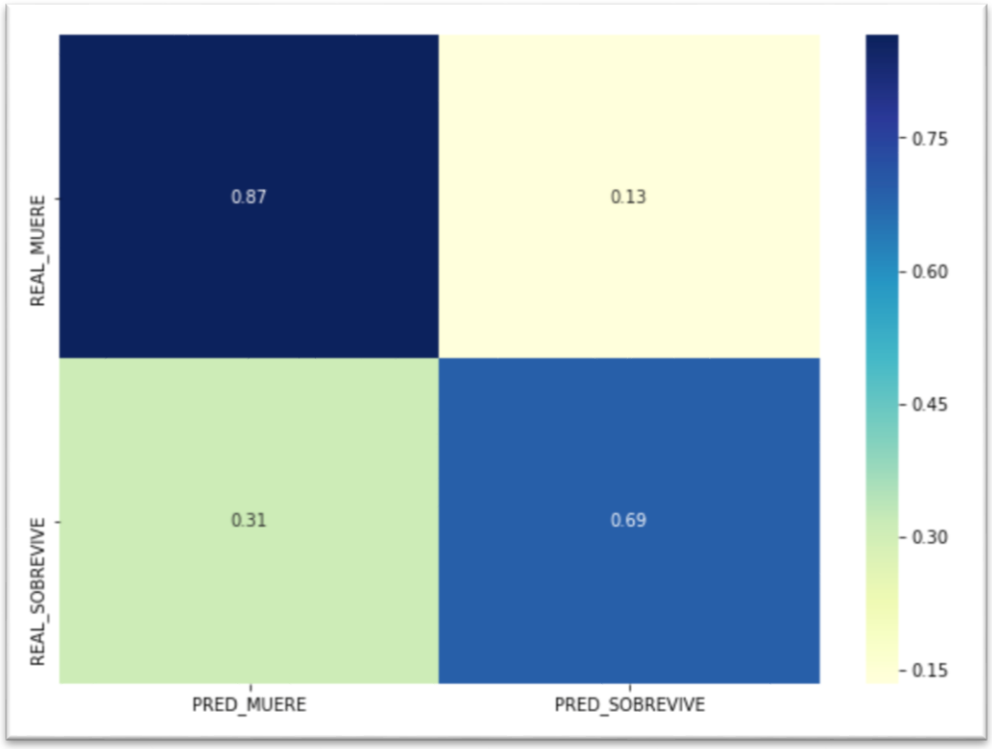
Matriz de correlación de todos los atributos del conjunto de datos



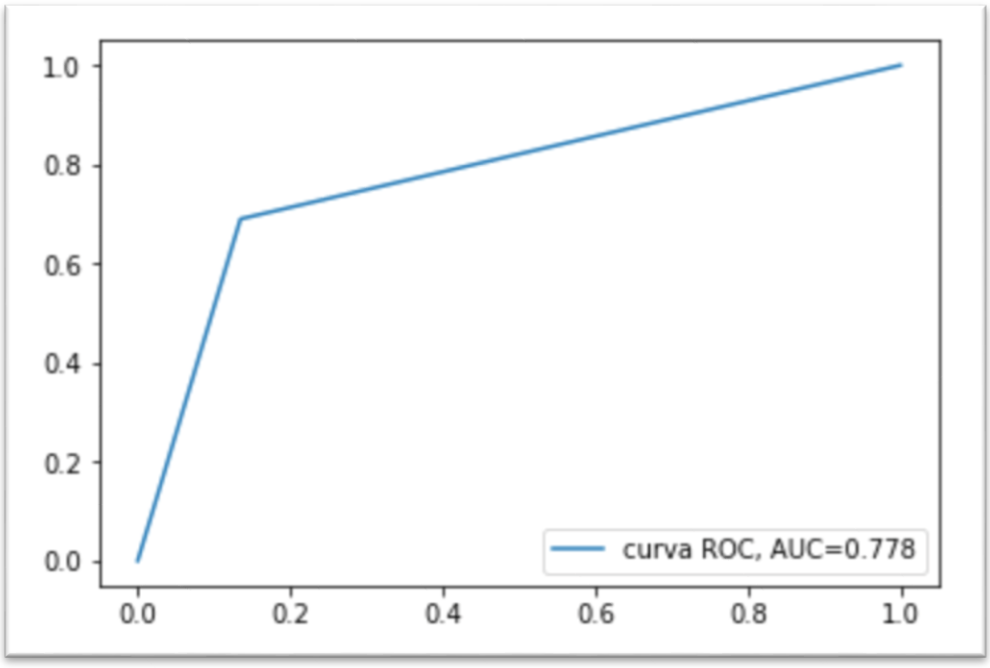
Matriz de confusión



Matriz de confusión normalizada



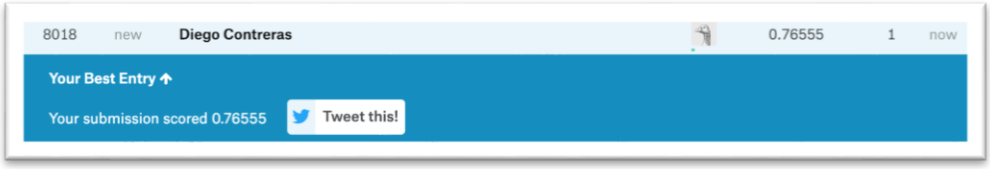
Curva ROC



6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Hemos conseguido obtener un modelo que permite predecir con una precisión de aproximadamente el 80% si una persona ha sobrevivido o no a la tragedia del Titanic según una serie de atributos, por lo que podemos afirmar que los resultados permiten responder al problema.

Como se comentó en un inicio, este conjunto de datos pertenecía a una competición de Kaggle. Tras realizar la predicción sobre el conjunto de test y enviar la solución al concurso se ha obtenido la siguiente clasificación:



Para llegar a este modelo hemos estudiado previamente las características del conjunto de datos, se han identificado los datos relevantes y se han limpiado los elementos que tenían valores a cero o que contenían valores extremos.

Además, se ha representado con las matrices de confusión y la curva ROC para mostrar gráficamente la calidad obtenida con este modelo.