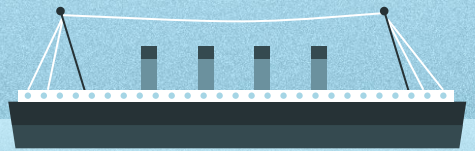
TP1 — GRUPO 8

TITANIC

EXPLORACIÓN, LIMPIEZA Y ANÁLISIS DE DATOS



Dataset

Cantidad de filas: 891

Cantidad de columnas: 12

Variables numéricas: “PassengerId” - ”Survived” - ”Pclass” - ”Age” - ”SIbSp” - ”Parch” - ”Fare”

Variables descriptivas: “Name” - “Sex” - “Ticket” - “Cabin” - “Embarked”

Valores vacios: 177 en la columna “Age”, 687 en la columna “Cabin”, 2 en la columna “Embarked”

Sobrevivientes: 342

Cantidad de hombres y mujeres: 577 hombres y 314 mujeres

— Proceso de limpieza —

1. Columna : “Cabin”

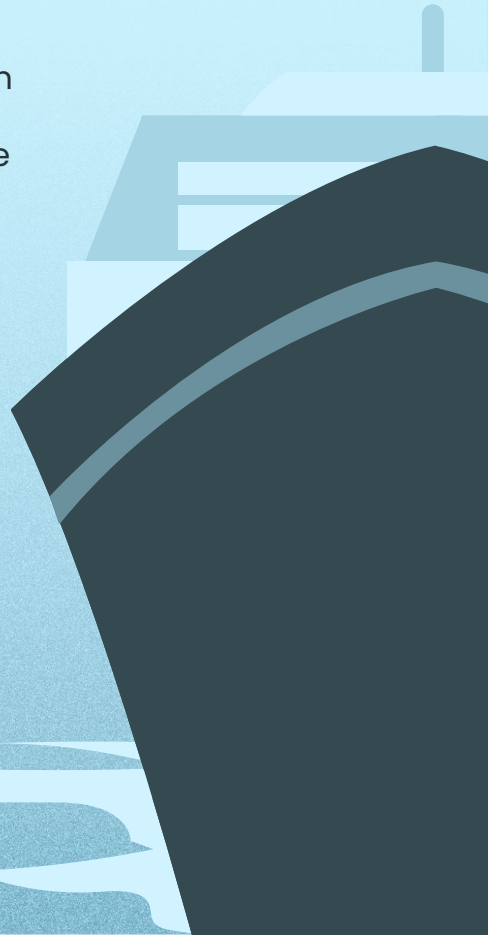
Analizando esta columna vimos que los que tenían registros tuvieron un mayor índice de supervivencia, por lo cual codificamos con 1 donde había registros y con 0 donde no los había. Esta nueva codificación nos aportará un poder predictivo mayor

2. Columna : “Age”

Analizando esta columna vimos que no había una correlación fuerte entre la edad del pasajero y su índice de supervivencia. Pero, al igual que en la columna “cabin”,los que tenían registrada la edad tenían un mayor grado de supervivencia. Por lo cual procedimos a codificarla de la misma manera que la columna “cabin”

3. Columna : “Embarked”

Al ser tan solo 2 datos faltantes procedimos a eliminar dichas filas ya que representan menos de 0,3% del dataset



— Visualización y análisis de datos—

- Para la predicción de supervivencia decidimos utilizar una regresión logística.
- Eliminamos algunas columnas que no nos aportaban información y codificamos otras para poder trabajar correctamente
- Columnas eliminadas: “PassengerId” - ”Name” - ”Ticket”
- Columnas codificadas: “Age” - “Sex” - ”Cabin” - ”Embarked” - ”SibSp” - ”Parch”

— Codificación —

- **“Age” y “Cabin”**: como se comentó previamente estas columnas se codificaron con un 1 dónde habìa registros y con un 0 donde no los había
- **“Sex”**: esta columna se codificó con un 1 si era mujer y con un 0 si era hombre
- **“Embarked”**: esta columna se codificó de forma binaria, al tener tres valores se codificó con dos columnas (is.C y is.S)
- **“SibSp” y ”Parch”:** estas dos columnas decidimos reducirlas a una, donde esta nueva columna tiene un 1 si el pasajero viajaba acompañado y con un 0 si viajaba solo

— Dataset Final —

Survived	Pclass	Sex	Age	Fare	Cabin	is.alone	is.C	is.S
0	3	0	1	7.25	0	0	0	1
1	1	1	1	71.2	1	0	1	0
1	3	1	1	7.9	0	1	0	1
0	3	0	1	8.05	0	1	0	1
1	1	1	1	53.1	0	1	0	1

— Regresión Logística —

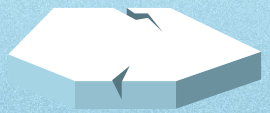
Se realizó un test de significancia a través de una regresión logística obteniendo los siguientes resultados:

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Pclass	-0.6875	0.0974	-7.0579	0.0000	-0.8785	-0.4966
Sex	2.5969	0.1917	13.5493	0.0000	2.2212	2.9725
Age	0.3028	0.2406	1.2585	0.2082	-0.1688	0.7743
Fare	-0.0001	0.0020	-0.0258	0.9794	-0.0040	0.0039
Cabin	0.6272	0.2491	2.5179	0.0118	0.1390	1.1153
is.alone	-0.0160	0.1782	-0.0895	0.9287	-0.3652	0.3333
is.C	0.0750	0.3292	0.2279	0.8197	-0.5702	0.7203
is.S	-0.4877	0.2874	-1.6973	0.0896	-1.0509	0.0755

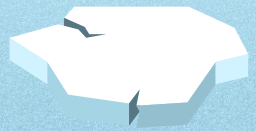
— Variables Seleccionadas —

Luego del ajuste del modelo nos quedamos con las tres variables que resultaron ser más significativas

1. Pclass



2. Sex

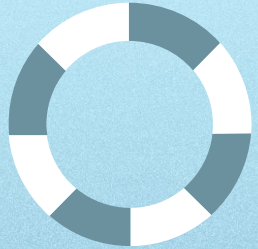


3. Cabin

— Entrenamos el modelo —

Separamos el dataset en una muestra de entrenamiento y una de prueba y obtenemos la siguiente matriz de confusión. Con tan solo 3 variables pudimos predecir con un 77% de precisiòn

	Positivos Predichos	**Negativos Predichos**
Positivos Observados	132	25
Negativos Observados	37	73



Accuracy: 77%

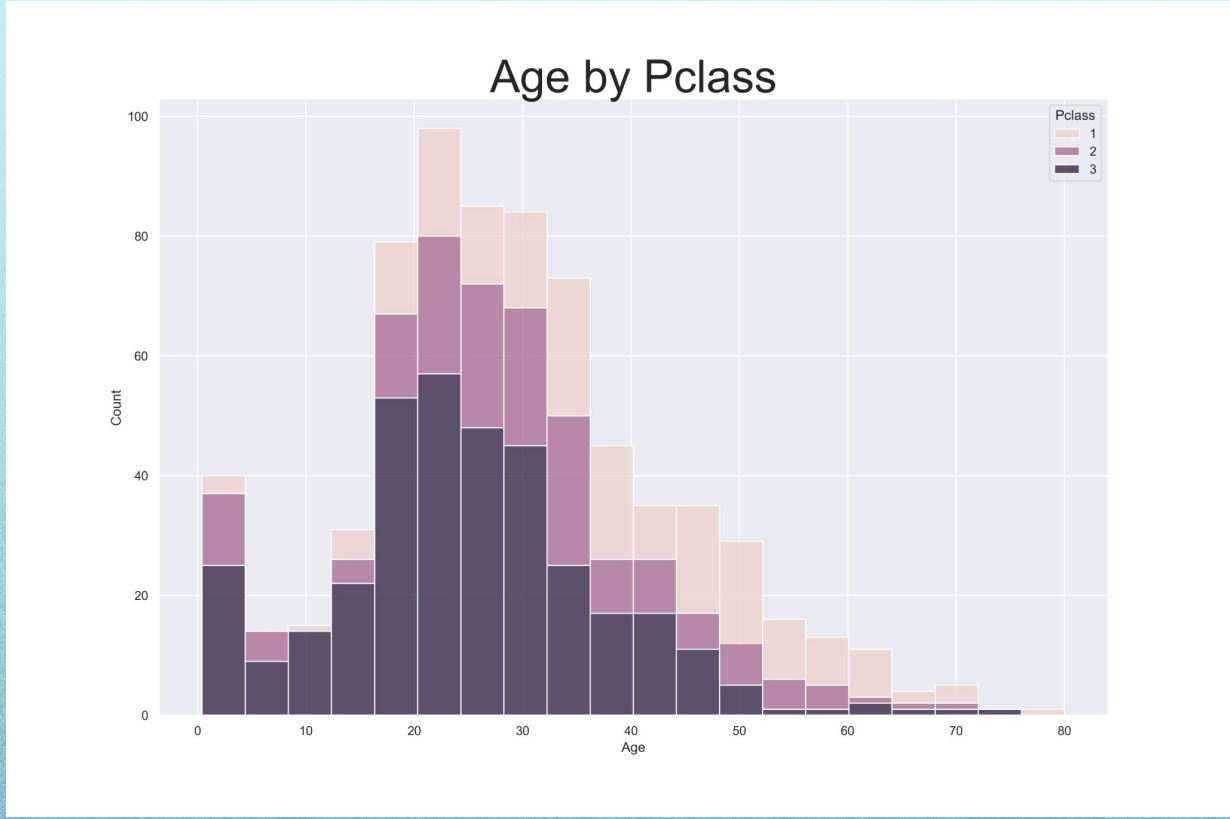
Correlación entre variables

No hay una gran correlación entre las variables, la de mayor correlación es Fare vs Pclass. A mayor valor del ticket menor es el nro de clase. Algo esperable a priori

	Survived	Pclass	Age	SibSp	Parch	Fare
Survived	1.00	-0.34	-0.08	-0.04	0.08	0.26
Pclass	-0.34	1.00	-0.37	0.08	0.02	-0.55
Age	-0.08	-0.37	1.00	-0.31	-0.19	0.10
SibSp	-0.04	0.08	-0.31	1.00	0.41	0.16
Parch	0.08	0.02	-0.19	0.41	1.00	0.22
Fare	0.26	-0.55	0.10	0.16	0.22	1.00

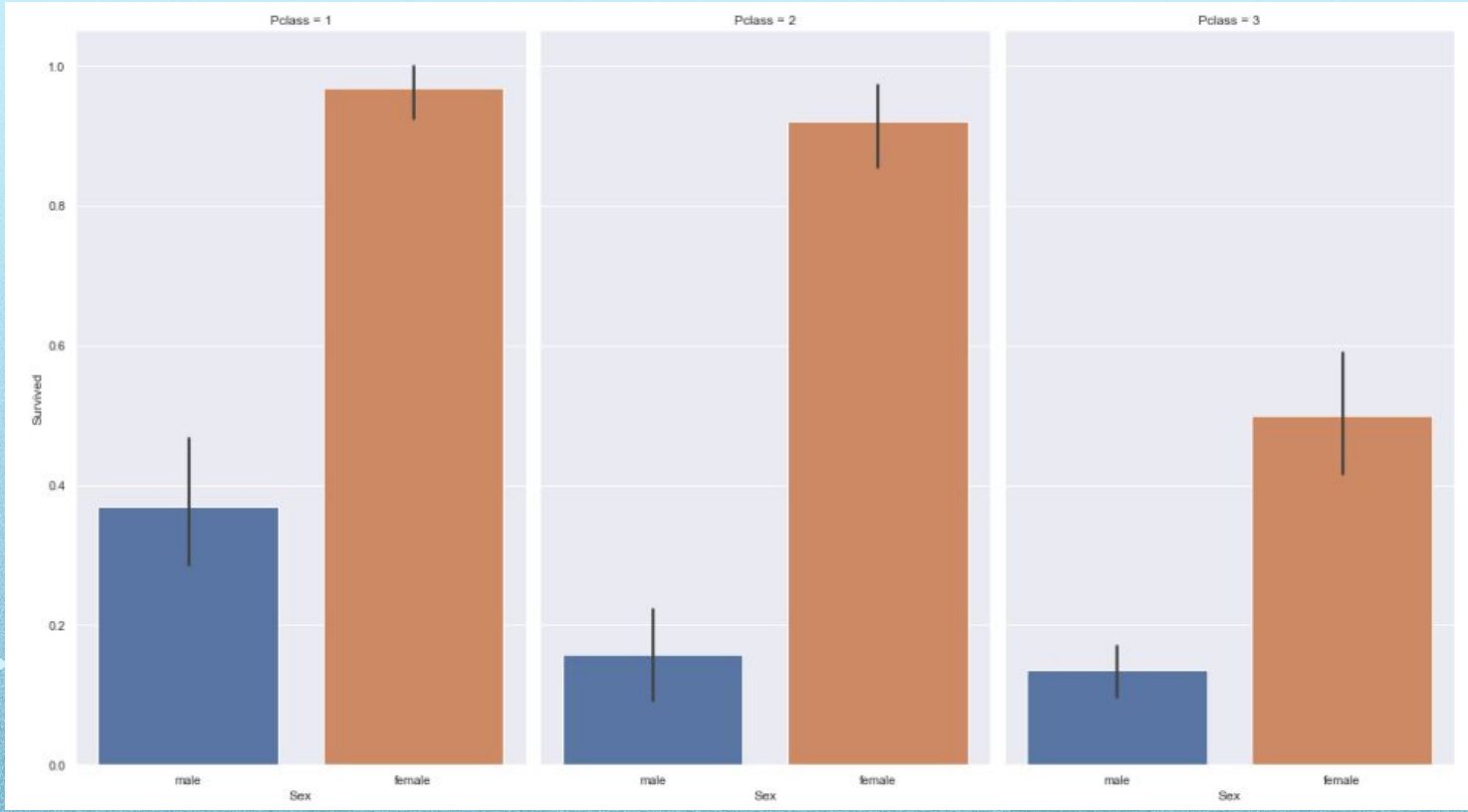
Gráficos

Este gráfico nos permite ver como estaban distribuidas las edades en cada clase. Los de clases más altas solían tener mayor edad que los de clases más bajas



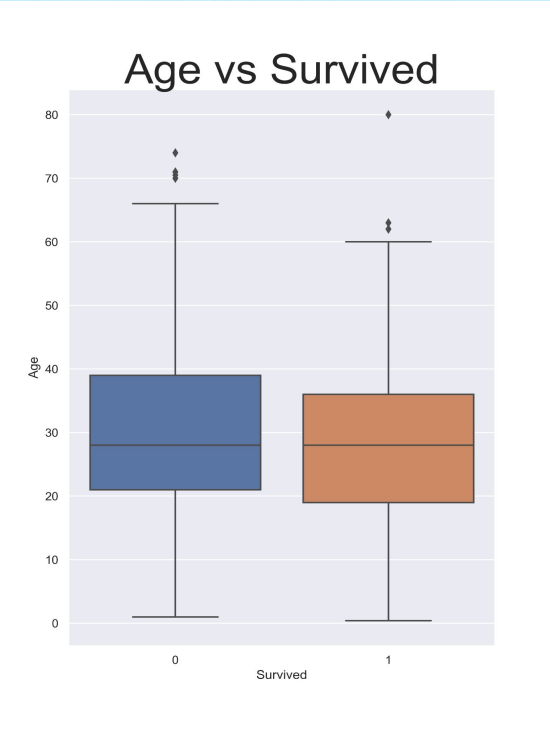
Gráficos

Este gráfico nos permite ver el índice de supervivencia por clase y sexo. Se puede observar que las mujeres de las clases más altas tuvieron un mayor índice de supervivencia



Gráficos

Este diagrama de cajas nos permite ver que la edad no fue un factor determinante a la hora de la supervivencia



— El capitán se hunde con su barco—

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0	B22	S

Gracias!