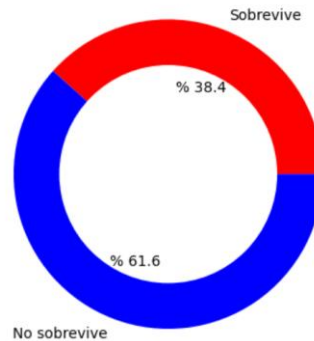


# Prediciendo la supervivencia en el Titanic mediante la aplicación de técnicas de análisis exploratorio de datos y aprendizaje automático

José Luis Domínguez Hernández  
Programación Python con Aplicaciones  
en el Ámbito Científico

## 1.1 Análisis de la información

La Figura 1 muestra que solamente el 38,4 % de los pasajeros logró sobrevivir al hundimiento del Titanic, mientras que el 61,6 % de los pasajeros no tuvo la misma suerte y perdió la vida en el desastre. Estos porcentajes reflejan la trágica magnitud de la tragedia y la gran cantidad de vidas que se perdieron en ese evento histórico.

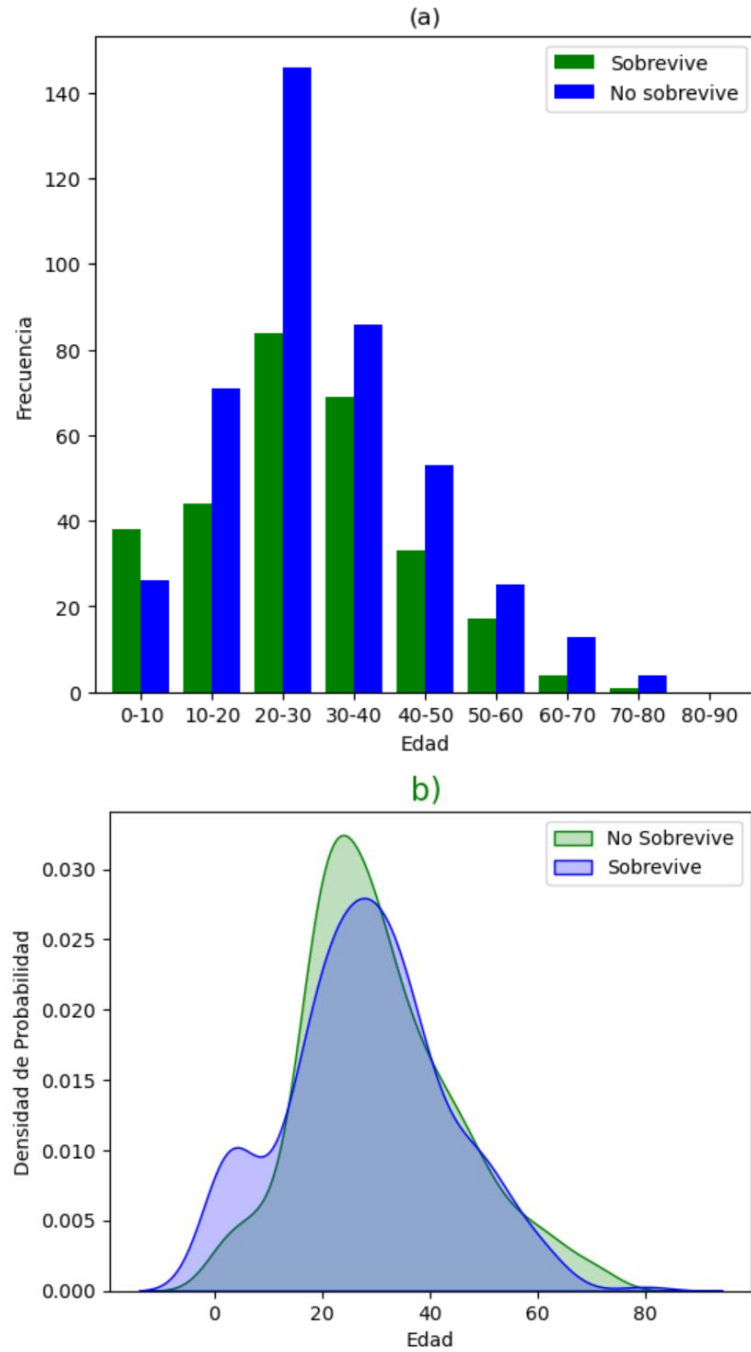


**Figura 1** Sobrevivientes al hundimiento del Titanic

### Edad

El pasajero más joven a bordo del Titanic tenía aproximadamente dos meses de edad, mientras que el pasajero de mayor edad tenía 80 años. La edad promedio de los pasajeros a bordo era ligeramente inferior a los 30 años (29.7 años).

La Figura 2 muestra que los niños menores de 10 años sobrevivieron en mayor cuantía que los que fallecieron. Sin embargo, para los otros grupos de edad, el número de víctimas superó al número de sobrevivientes. Específicamente, más de 140 personas en el grupo de edad de 20 a 30 años fallecieron, en contraste con solo alrededor de 80 personas de la misma franja de edad que lograron sobrevivir.

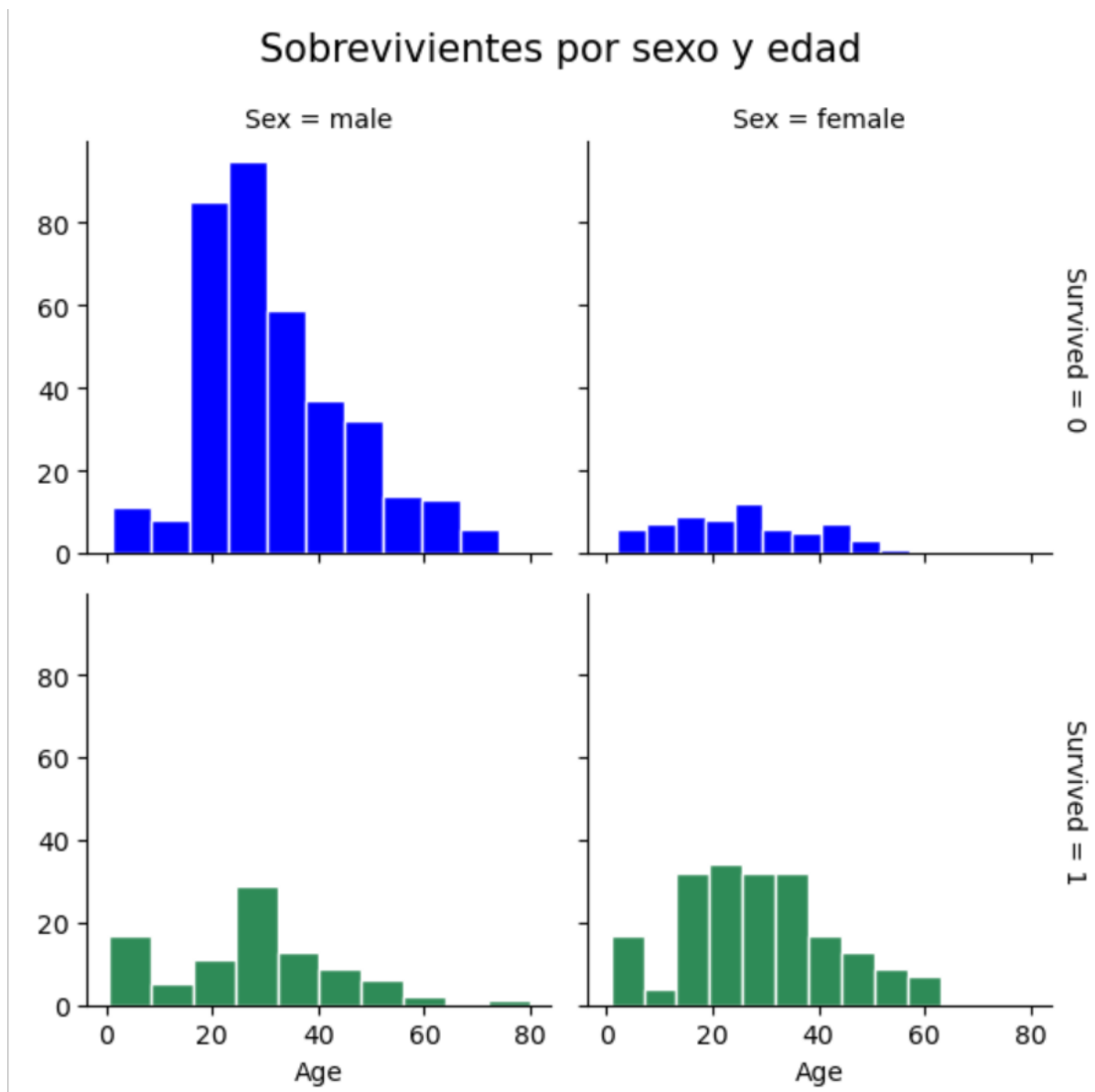


**Figura 2** Sobrevivientes por Edad. En a) diagrama de barras y en b) Gráfico de densidad de probabilidad.

La Figura 3 permite comparar visualmente los niveles de supervivencia por grupos de edades para hombres y mujeres. En el eje x se encuentran los diferentes grupos de edad, mientras que en el eje y se presenta el número de sobrevivientes. Cada barra representa un grupo de edad y su altura representa el número bien sea de hombres o mujeres que sobrevivieron o no en ese grupo de edad en particular.

El gráfico de barras resalta que las niñas menores de 10 años tuvieron una tasa de supervivencia muy alta en comparación con otros grupos de edad. A medida que aumenta la edad, la tasa de supervivencia tiende a disminuir ligeramente, pero sigue siendo considerablemente alta en comparación con los grupos de edad correspondiente a los hombres.

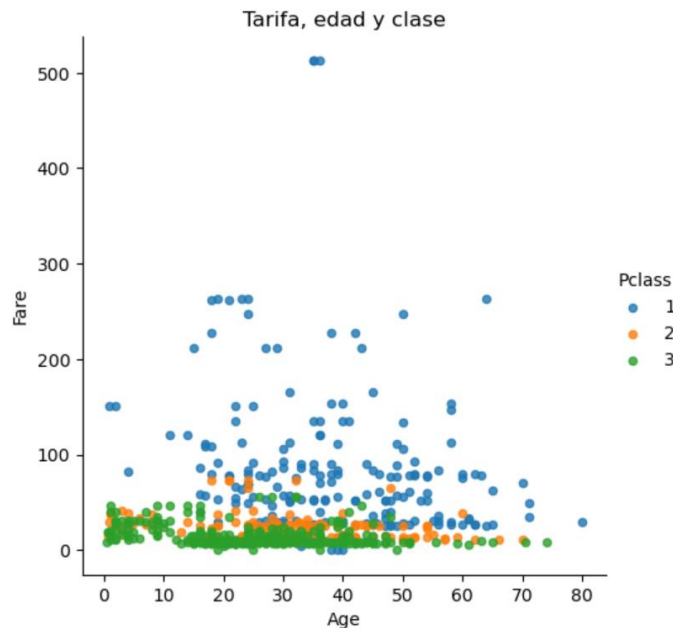
Este gráfico de barras proporciona una representación visual clara de cómo la tasa de supervivencia varía según la edad en el caso del sexo femenino. Muestra la importancia del factor de edad en la probabilidad de supervivencia de las mujeres durante el naufragio del Titanic.



**Figura 3** Sobrevivientes por sexo y edad

## Tarifa

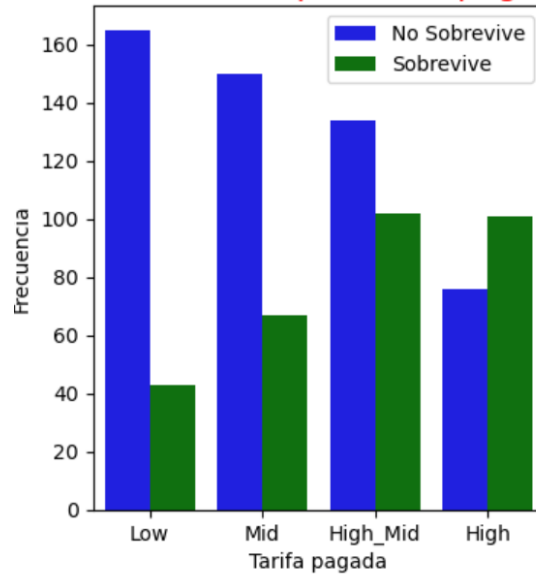
La tarifa pagada esta asociada a la clase, la edad y el puerto de salida y las comodidades contratadas como se muestra en la Figura 4.



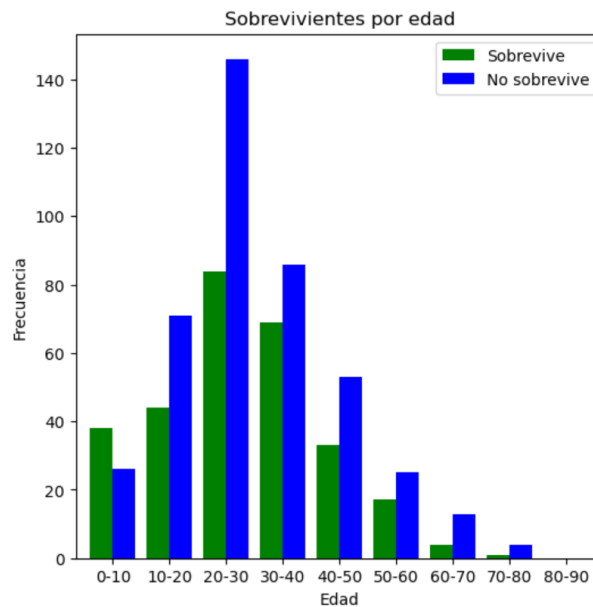
**Figura 4** Sobrevivientes por sexo y edad

Al dividir el monto de la tarifa en cuatro categorías, se hizo evidente que existía una fuerte asociación entre el precio pagado y la supervivencia. La Figura 4 muestra claramente que cuanto más alto era el monto pagado por un pasajero, mayores eran sus posibilidades de sobrevivir. Las categorías más altas de tarifa tienen una proporción mucho mayor de sobrevivientes en comparación con las categorías más bajas. Esto sugiere que aquellos pasajeros que pagaron una tarifa más alta posiblemente tuvieron acceso a mejores servicios y ubicaciones en el momento del naufragio, lo que les brindó mayores oportunidades de sobrevivir.

### Sobrevivencia por tarifa pagada

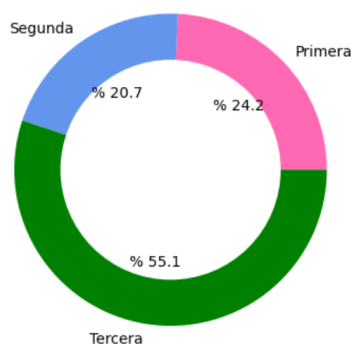


**Figura 5** Sobrevivientes por sexo y edad



**Figura 6** Sobrevivientes por sexo y edad

La Figura 7 muestra la distribución de pasajeros por clase. Es interesante observar que casi la mitad de los pasajeros viajaban en tercera clase.



**Figura 7 Distribución de pasajeros por clase**

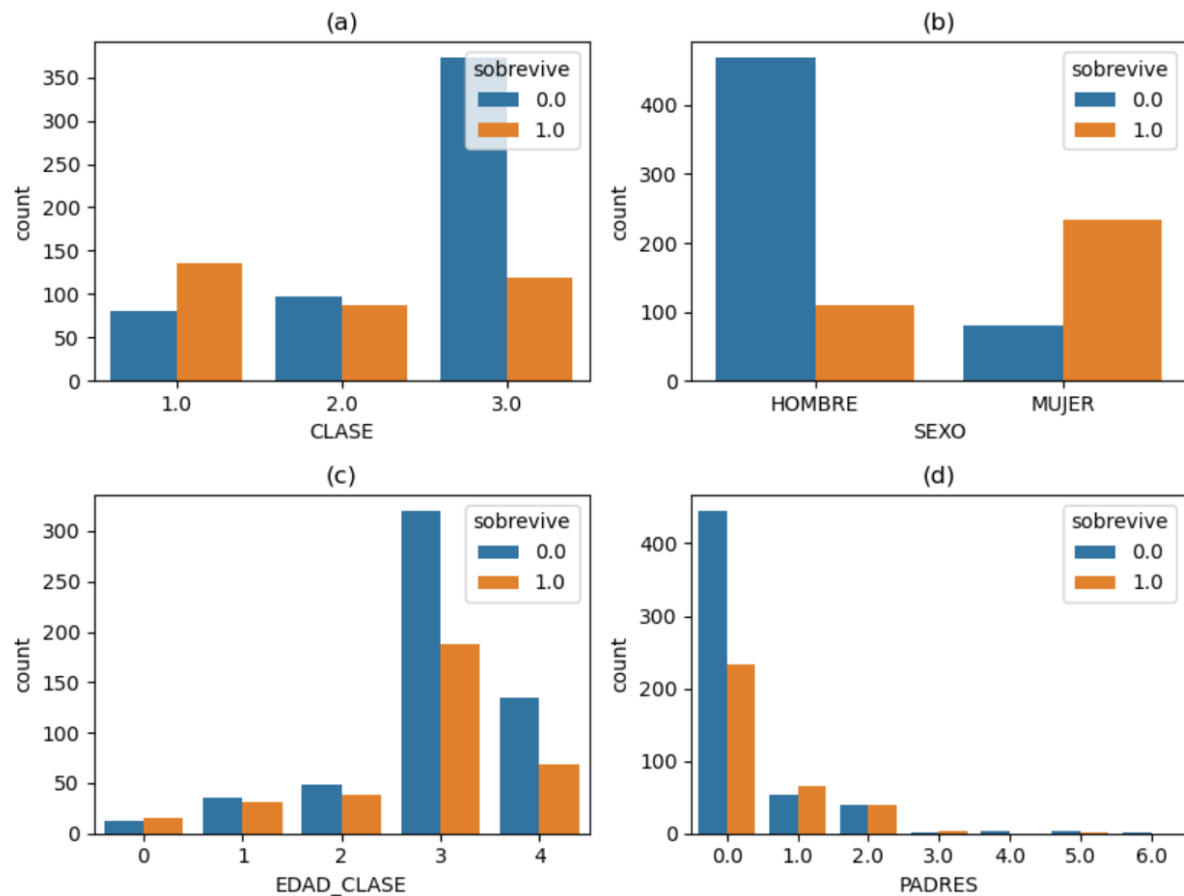
En la Figura 8 a), que es un gráfico de barras comparativo, se puede visualizar de manera más clara la diferencia en las tasas de supervivencia entre las tres clases. Las barras correspondientes a la primera clase son las más altas, seguidas de las barras de la segunda clase y, finalmente, las barras más bajas representan la tercera clase.

Estos datos reflejan los patrones de supervivencia observados durante el desafortunado naufragio del Titanic en 1912, donde se evidenció una mayor prioridad y acceso a los botes salvavidas para los pasajeros de clases más altas, especialmente los de primera clase.

Las Figura 8 b y c) muestran que las mujeres y los bebés tuvieron las tasas más altas en comparación a sus respectivas clases. En porcentaje se detecta un 74.20% para las mujeres, mientras que para los hombres fue de solo el 18.89%. Por otro lado, los bebés tuvieron la tasa de supervivencia más alta, con un 64%. Los adultos tuvieron una tasa de supervivencia del 37.85%, mientras que los ancianos tuvieron la tasa más baja, con un 33.50%.

Estos resultados podrían estar relacionados con la política de "mujeres y niños primero" seguida durante el naufragio. Es posible que los bebés y los niños pequeños recibieran prioridad en la evacuación y asignación de botes salvavidas. Esto podría explicar la diferencia significativa en las tasas de supervivencia entre hombres y mujeres, donde las mujeres y bebés tuvieron una tasa mucho más alta de supervivencia en comparación con los hombres, adultos y viejos.

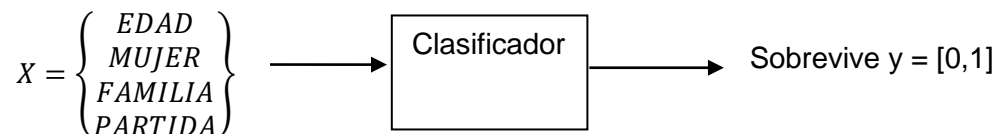
La Figura 7 d) sobrevivieron 233 de 678 individuos sin familiares a bordo. Es decir, 34.36 %.



**Figura 8** Comparativo por sector de sobrevivientes al hundimiento del Titanic

## 1.2 Clasificación

Dadas las características asociadas a un pasajero se plantea definir si la variable de salida sobrevivencia es cero o uno.



En este ejercicio, se realiza la división de los datos en una muestra de entrenamiento y otra de prueba. La muestra de entrenamiento contendrá el 75% de las observaciones, mientras que los datos de prueba representarán el 25% restante.

Cuando se aplica un método de clasificación, es posible que se produzcan dos tipos de errores. El primer error es cuando el valor estimado ( $\hat{y}$ ) es igual a cero, mientras que en



realidad el valor real (y) es igual a uno. Esto se conoce como un falso negativo o un error de tipo I. Por otro lado, el segundo error ocurre cuando el valor estimado es igual a uno, mientras que el valor real es cero. Esto se conoce como un falso positivo o un error de tipo II.

Predicción:

Edad	Sexo	Familiares	Tarifa	Clase	Embarcación	Resultado
42	1	0	7.25	3	1	0
5	0	1	11.25	1	1	1