Instituto Politécnico Nacional
Centro de Investigación en Computación
Curso de Capacitación

Aprendizaje Automátizado:
Caso de Estudio: Titanic

Programación Python
con Aplicaciones
en el Ámbito Científico

Presenta:

José Luis Domínguez Hernández

Profesor:

Alan Badillo Salas

Ciudad de México **julio de 2023**

Introducción

El aprendizaje automático —también conocido como machine learning, en inglés— es un campo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y tomar decisiones basadas en datos, sin necesidad de programación explícita para tareas específicas. En lugar de seguir instrucciones directas, las máquinas aprenden de forma automática a partir de ejemplos, patrones y experiencias pasadas.

El aprendizaje automático se basa en la idea de que las computadoras pueden analizar grandes cantidades de datos y detectar patrones, tendencias y relaciones ocultas que pueden ser difíciles de identificar para los seres humanos. A través de técnicas y algoritmos específicos, el aprendizaje automático permite a las máquinas extraer información útil, hacer predicciones y tomar decisiones en base a esos datos.

Una de las principales fortalezas del aprendizaje automático es su capacidad para adaptarse y mejorar su rendimiento a medida que se le proporciona más información. Esto se logra mediante la retroalimentación constante y la capacidad de ajustar los modelos en función de los resultados observados.

El aprendizaje automático tiene una amplia gama de aplicaciones en diversas áreas, como reconocimiento de voz, procesamiento de imágenes, detección de fraudes, recomendación de productos, diagnóstico médico, conducción autónoma, entre muchas otras. A medida que la cantidad de datos disponibles sigue creciendo exponencialmente, el aprendizaje automático se vuelve cada vez más relevante y poderoso para aprovechar todo ese potencial y obtener información valiosa a partir de los datos.

Presentación del caso de Estudio

El RMS Titanic, famoso por su trágico naufragio en 1912, proporciona un escenario propicio para aplicar técnicas de aprendizaje automático. Los datos suelen incluir diversos atributos como

- Identificación (id): Número de identificación del usuario
- Nombre: El nombre del pasajero.
- Género: El género del pasajero (masculino o femenino).
- Edad: La edad del pasajero.
- Clase de boleto: La clase de boleto que el pasajero tenía (Primera, Segunda o Tercera clase).
- Número de hermanos o cónyuges a bordo.
- Número de padres o hijos a bordo.
- Tarifa pagada: El precio pagado por el boleto del pasajero.
- Número de cabina: El número de la cabina asignada al pasajero.
- Puerto de embarque: El puerto desde donde el pasajero embarcó (por ejemplo, Southampton, Cherbourg o Queenstown).

Además de estos atributos, la base de datos también incluye una etiqueta que indica si el pasajero sobrevivió o no al naufragio. Esta etiqueta es fundamental para el aprendizaje automático supervisado, ya que permite entrenar y evaluar los modelos para realizar predicciones precisas sobre la supervivencia de los pasajeros en función de los demás atributos.

Planteamiento del problema
Dada información relevante sobre cada uno de los pasajeros del Titanic, descrita en la Tabla 1, el problema se plantea como encontrar una función $f: X \rightarrow y$ que minimiza una función de costo J, que mide la diferencia entre las predicciones del modelo y los valores reales de la variable objetivo.

En este contexto, X representa el conjunto de variables independientes tales como edad, clase en la que viajaba, edad, número de hermanos o conyugues abordo, número de padres o de hijos, la tarifa pagada, puerto en el que abordo.

La variable objetivo y es la variable dependiente que indica si el pasajero sobrevivió o no sobrevivió. Se trata de la variable que se busca predecir mediante el modelo de clasificación. Los valores de y pueden ser 1 para indicar que el pasajero sobrevivió, o 0 para indicar que el pasajero no sobrevivió.

Tabla 1 Información relevante sobre cada uno de los pasajeros del Titanic

Encabezado	Descripción	Valores
PassengerId	Número de identificación de los pasajeros	$\{1,2,3,\ldots,891\}$
Survived	Sobrevive: 0: no sobrevive,1: sobrevive	$\{0,1\}$
Pclass	Clase	$\{1,2,3\}$
Age	Edad	$\{1,2,3,\ldots,83\}$
SibSp	Número de hermanos o conyugues abordo	$\{0,1,2,3,4,5\}$
Parch	Número de padres o hijos abordo	$\{0,1,2,3,4,5\}$
Fare	Tarifa o precio pagado	$Número\ real$
Embarked	Puerto en que abordo el pasajero: Southampton, Cherbourg, o Queenstown.	$\{S, C, Q\}$

Para resolver este problema, se pueden aplicar diversas técnicas y algoritmos de aprendizaje automático, como clasificación binaria, regresión logística, árboles de decisión, entre otros. El modelo se entrena utilizando un conjunto de datos de entrenamiento, donde se conocen las etiquetas de supervivencia reales de los pasajeros. Luego, se evalúa el rendimiento del modelo utilizando un conjunto de datos de prueba, donde se comparan las predicciones del modelo con las etiquetas reales de supervivencia para medir su precisión y eficacia.

1.1 Análisis de la información

Sexo
De los 891 pasajeros, solo 342 sobrevivieron (233 mujeres y 109 hombres). Esto representa un 38.4% de supervivencia y un 61.6% de fallecimiento. La tasa de supervivencia de las mujeres fue significativamente mayor, alcanzando el 74.2%, en contraste con el bajo 18.89% de los hombres como se muestra en la Tabla 2.

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 1 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.

Tabla 2 Sobrevivencia por genero

Sobrevive	Mujeres	Hombres	Total
No	81 (25.80%)	468 (81.11%)	549 (61.62%)
Si	233 (74.20%)	109 (18.89%)	342 (38.38%)
Total	314	577	891

Programación en Python:

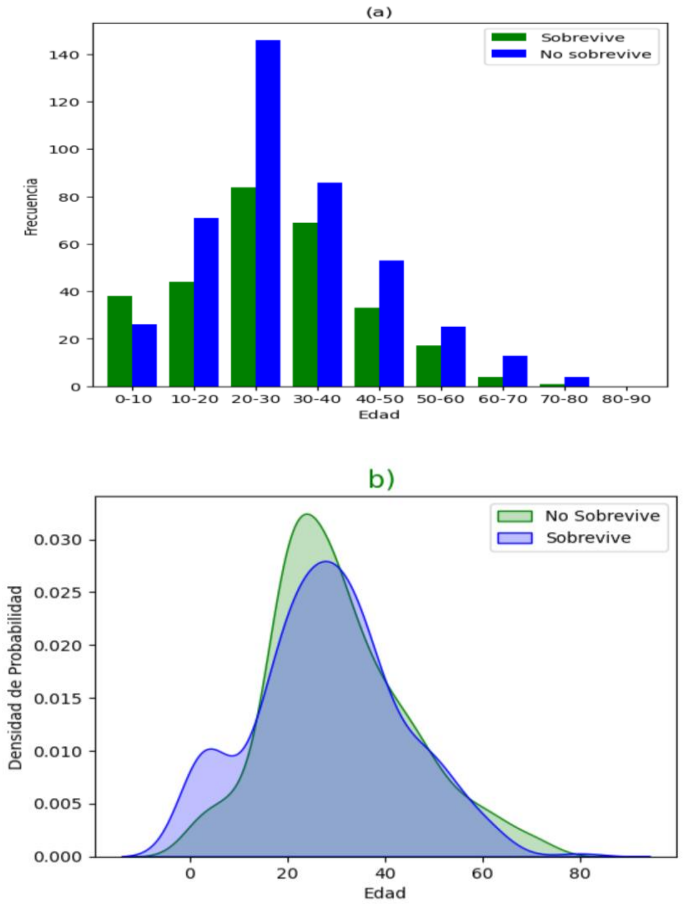

Figura 1 Sobrevivientes por Edad. En a) diagrama de barras y en b) Gráfico de densidad de probabilidad.

Tarifa

La tarifa pagada por cada pasajero está influenciada por la clase en la que viajaban, el puerto de salida y las comodidades contratadas. La Figura 1 muestra la relación entre la tarifa pagada y la edad de los pasajeros, y utiliza colores para distinguir la clase en la que viajaban.

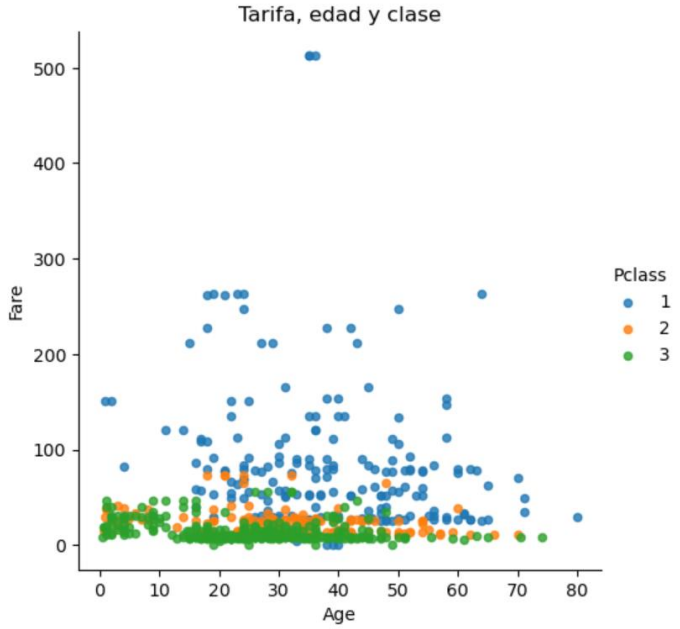

Figura 2 Tarifa por edad y clase

La Figura 3 ilustra las tarifas pagadas por mujeres y hombres durante el viaje del Titanic. Se observa que las mujeres que viajaban en primera clase pagaron en promedio $106.12 mientras que los hombres que viajaban en la misma clase tan sólo $67.22. Por otro lado, tanto hombres como mujeres en segunda y tercera clase pagaron tarifas similares. En segunda clase, las tarifas estuvieron alrededor de $20, mientras que en tercera clase estuvieron en el rango de $14. Estos datos resaltan las diferencias en las tarifas pagadas por género y clase durante el viaje en el Titanic.

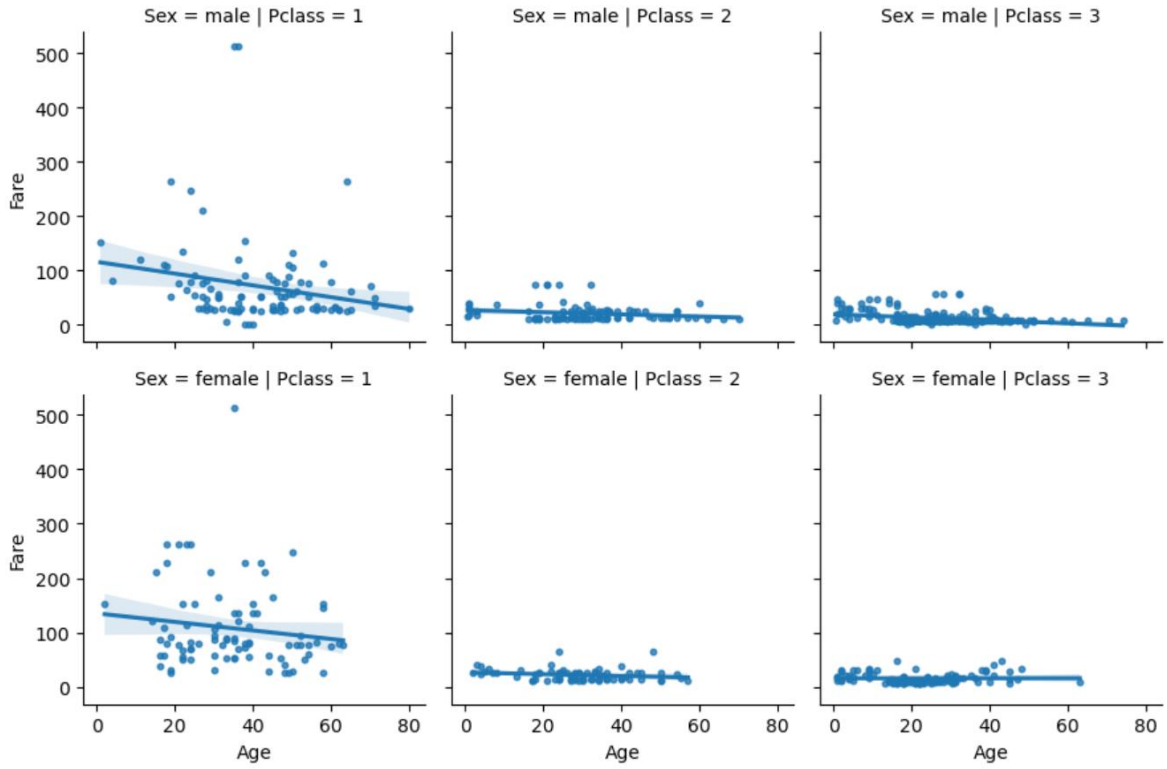

Figura 3 Tarifa por edad, género y clase

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.

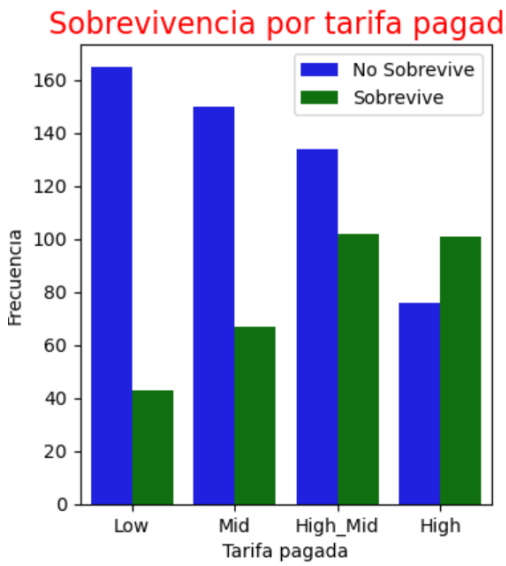

Figura 4 Sobrevivientes por sexo y edad

Clase

La Tabla 3 proporciona los números de pasajeros en cada clase, así como la tasa de sobrevivientes por género. En a) se observa que las mujeres que viajaban en primera y segunda clase fueron las más favorecidas (el 96.81 y 92.11 % respectivamente), Incluso las mujeres que viajaban en tercera clase (50.00 %) superando a la de los hombres. En efecto, en primera clase, el 36.88% de los hombres sobrevivió, mientras que en segunda y tercera clase, las tasas de supervivencia son en bajas en comparación (15.74% y 13.54% respectivamente). Estos datos resaltan las diferencias en las tasas de supervivencia basadas en el género y la clase de los pasajeros en el Titanic.

Tabla 3 Sobrevivencia por clase

a) Mujeres

Situación	Clase 1	Clase 2	Clase 3
Pasajeras	94	76	144
Sobrevivientes	91	70	72
No Sobrevivientes	3	6	72
Sobrevivencia (%)	96.81	92.11	50.00

b) Hombres

Situación	Clase 1	Clase 2	Clase 3
Pasajeros	122	108	347
Sobrevivientes	45	17	47
No Sobrevivientes	77	91	300
Sobrevivencia (%)	36.88	15.74	13.54

Sobrevivencia por sexo y edad

La Figura 5 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.

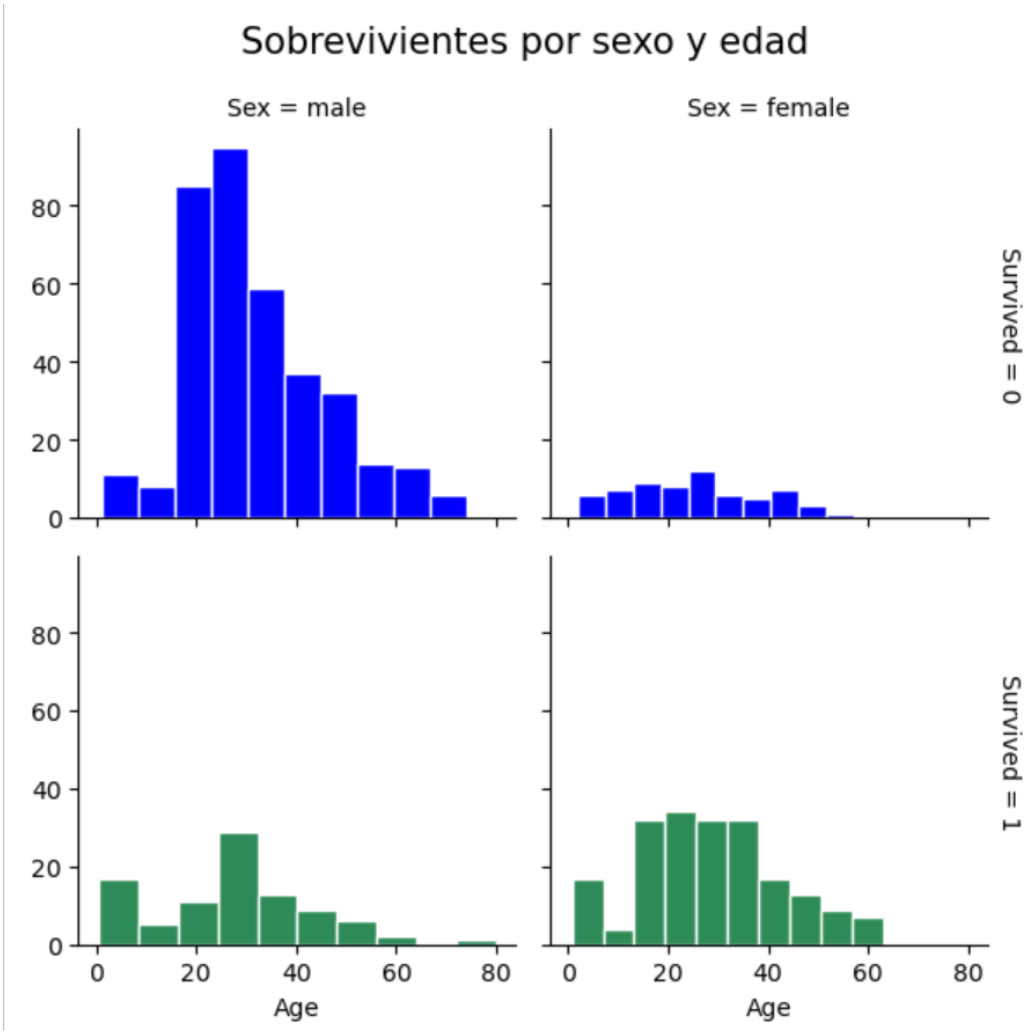

Figura 5 Sobrevivientes por sexo y edad

La Figura 6 presenta la información sobre la sobrevivencia según la edad, el género y la clase en la que viajan los pasajeros. En el caso de aquellos que viajaban en tercera clase, se observa que los 9 pasajeros mayores de 50 años fallecieron, lo que indica una tasa de supervivencia del 0% para este grupo. Del grupo de 89 pasajeros de entre veinte y treinta años de edad en tercera clase, solo 13 lograron sobrevivir, mientras que 76 fallecieron. Por otro lado, se identifica que 75 mujeres viajaban en primera clase, y de ellas, solo dos fallecieron. Una de las mujeres era menor de diez años y la otra tenía entre veinte y treinta años.

Estos datos resaltan las diferencias en las tasas de supervivencia basadas en la edad, género y clase de los pasajeros en el Titanic, donde se observa una mayor tasa de supervivencia entre las mujeres en general y particularmente entre aquellas que viajaban en primera clase.

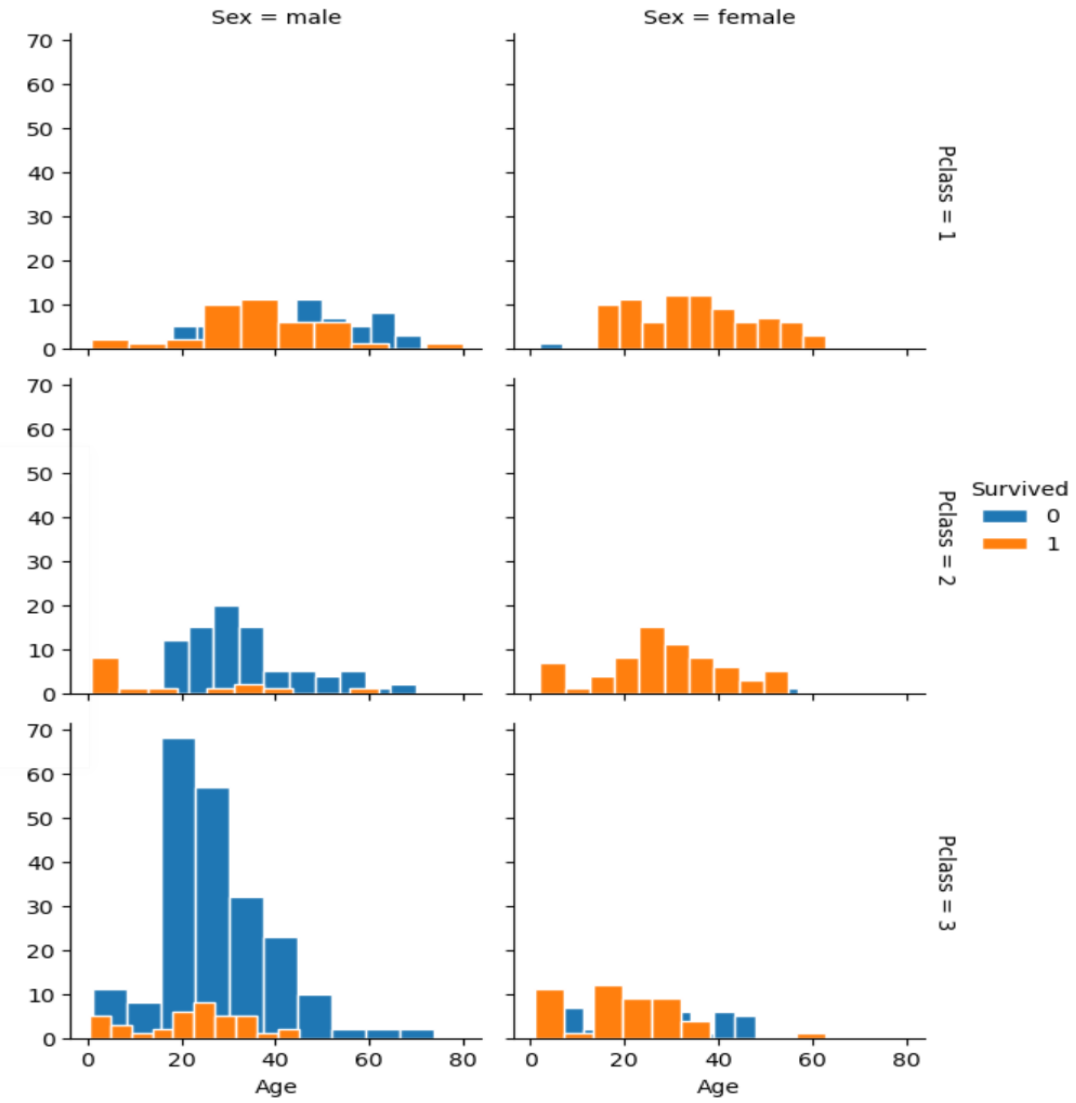

Figura 6 Sobrevivientes por edad, sexo y clase en la que viajaban

Teorema de Bayes
El Teorema de Bayes tiene aplicaciones en diversos campos, como en el análisis de datos, inteligencia artificial, medicina, y otros, donde se utiliza para actualizar probabilidades y hacer inferencias basadas en nueva información. Es especialmente útil en problemas de clasificación y diagnóstico, donde se requiere calcular la probabilidad de un evento dado un conjunto de características o evidencias observadas.

El teorema establece la relación entre dos tipos de probabilidades: la probabilidad condicional y la probabilidad marginal. La probabilidad condicional es la probabilidad de que ocurra un evento A dado que ya sabemos que ha ocurrido un evento B, y se denota como P(A|B). Por otro lado, la probabilidad marginal es la probabilidad de que ocurra el evento B, y se denota como P(B).

La formulación matemática del Teorema de Bayes es la siguiente:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Donde:

$P(A/B)$	probabilidad condicional de A dado B e indica la probabilidad de que el evento A sabiendo que ha ocurrido B.
$P(B/A)$	probabilidad condicional de B dado A (la probabilidad de que ocurra B sabiendo que A es verdadero).
$P(A)$	probabilidad de que ocurra A
$P(B)$	probabilidad de que ocurra B

Ejemplo 1 Estimar la probabilidad de sobrevivir dado que se es mujer (usar regla de Bayes).

Se plantea estimar una probabilidad condicional y se puede determinar directamente.

Se tiene información de 891 pasajeros de los cuales 314 son de género femenino y 577 del sexo masculino. Es decir, entre los pasajeros:

$$P(M) = \frac{314}{891} * 100 = 35.24\ \%$$

$$P(H) = \frac{577}{891} * 100 = 64.75\ \%$$

La tasa de sobrevivencia por genero son:

$$P(S/M) = 233/314 = 74.20\%$$

$$P(S/H) = 109/577 = 18.89\%$$

Ejemplo 2 Dado que se es sobreviviente ¿Cuál es la probabilidad de que sea una mujer?

Solución. Sobrevivieron 233 mujeres y 109 hombres. Entonces, para que un sobreviviente tomado al azar sea una mujer $P(M/S) = 233/342 * 100 = 68.12\%$ y de que sea hombre $P(H/S) = 109/342 * 100 = 31.87\,\%$.

Otra forma es aplicando el teorema de Bayes,

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Entonces sustituyendo valores queda:

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)} = \frac{0.7420 \times 0.3524}{0.7420 \times 0.3524 + 0.1889 \times 0.6425} = 0.6813$$

$$P(H/S) = \frac{P(S/H)P(H)}{P(S)} = \frac{0.6476 \times 0.1889}{0.7420 \times 0.3524 + 0.1889 \times 0.6425} = 0.3187$$

Ejemplo 3 ¿Cuál es la probabilidad de que una sobreviviente haya viajado en primera clase?

Solución. Se trata de calcular una probabilidad condicional; Dado que un pasajero sobrevivió siendo del sexo femenino se plantea calcular que hubiera viajado en primera clase. Se define:

A) El pasajero es una mujer
B) El pasajero sobrevivió
C) La pasajera viajó en primera clase

Sea $S = A \cap B$

Se sabe que 314 mujeres viajaban en el Titanic. De las cuales 94 viajaban en primera clase, 76 en segunda clase y 144 en tercera clase. Entonces las probabilidades de ocurrencia de que viajaban en cada clase son:

$$P(C_1) = 94/314 = 0.2993$$
$$P(C_2) = 76/314 = 0.2420$$
$$P(C_3) = 144/314 = 0.4585$$

Ahora bien, la probabilidad de que una mujer sobreviviera dado que viajaba en las clases C_1, C_2, C_3

$$P(S/C_1) = 91/94 \ = 0.9681$$
$$P(S/C_1) = 70/76 \ = 0.9211$$
$$P(S/C_1) = 72/144 \ = 0.9681$$

La fórmula del teorema de Bayes en este caso:

$$P(C_1/S) = \frac{P(S/C_1)\,P(C_1)}{P(S)}$$

Donde

$$P(S) = \sum_{i=1}^{3} P(S/C_i)\,P(C_i) = 0.9681 * 0.2993 + 0.9211 * 0.2420 + 0.50 * 0.4585$$
$$= 0.74190.$$

Entonces

$$P(C_1/S) = \frac{P(S/C_1)\,P(C_1)}{P(S)} = \frac{0.2993 \times 0{,}9681}{0.74189} = 0.3905$$

Otra forma de verlo; Sobrevivieron 314 pasajeros de sexo femenino de los cuales 94 viajaban en primera clase. Entonces, la probabilidad de que una sobreviviente viajara en primera clase es

$$\left(\frac{91}{233}\right) * \ 100 \ = \ 39.05\,\%$$

Conclusiones

El análisis de los datos del Titanic revela claras diferencias en las tasas de supervivencia basadas en la edad, el sexo y la clase de los pasajeros. Dos hallazgos destacados son que los hombres mayores de cincuenta años que viajaban en tercera clase fallecieron en su totalidad, lo cual indica una menor probabilidad de supervivencia para este grupo específico. Por otro lado, más del 90% de las mujeres que viajaban en primera y segunda clase lograran sobrevivir, lo cual destaca la mayor tasa de supervivencia para las mujeres en general y para las que tenían acceso a una clase más alta en particular. Estos datos subrayan la importancia de la edad, el sexo y la clase en las probabilidades de supervivencia en el desastre del Titanic.

El Apéndice A contiene las respuestas a Preguntas Especificas.

Páginas consultadas

Titanic Data (consultada el 3 de julio de 2023).
https://raw.githubusercontent.com/datasciencedojo/datasets/master/data.csv

A beginner's guide to Kaggle's Titanic problem | by Sumit Mukhija | Towards Data Science

Apéndice A
Respuestas a Preguntas Especificas

```
print('Total de sobrevivientes        : {}'.format(data.Survived.eq(1).sum()))
print('Total de sobrevivientes mujeres : {}'.format((data.Survived.eq(1) & data.Sex.eq('female') ).sum()))
print('Total de sobrevivientes hombres : {}'.format((data.Survived.eq(1) & data.Sex.eq('male') ).sum()))
print('Total de supervivientes mayores a 18 años : {}'.format((data.Survived.eq(1) & data.Age.gt(18) ).sum()))
print('Total de supervivientes menores a 18 años : {}'.format((data.Survived.eq(1) & data.Age.lt(18) ).sum()))
print('Total de supervivientes mayores a 50 años : {}'.format((data.Survived.eq(1) & data.Age.gt(50) ).sum()))
print('Total de muertos         : {}'.format(data.Survived.eq(0).sum()))
print('Total de muertos mujeres : {}'.format((data.Survived.eq(0) & data.Sex.eq('female') ).sum()))
print('Total de muertos hombres : {}'.format((data.Survived.eq(0) & data.Sex.eq('male') ).sum()))
print('Total de supervivencia de la clase 1 : {}'.format((data.Survived.eq(1) & data.Pclass.eq(1) ).sum()))
print('Total de supervivencia de la clase 2 : {}'.format((data.Survived.eq(1) & data.Pclass.eq(2) ).sum()))
print('Total de supervivencia de la clase 3 : {}'.format((data.Survived.eq(1) & data.Pclass.eq(3) ).sum()))
```

```
Total de sobrevivientes        : 342
Total de sobrevivientes mujeres : 233
Total de sobrevivientes hombres : 109
Total de supervivientes mayores a 18 años : 220
Total de supervivientes menores a 18 años : 61
Total de supervivientes mayores a 50 años : 22
Total de muertos         : 549
Total de muertos mujeres : 81
Total de muertos hombres : 468
Total de supervivencia de la clase 1 : 136
Total de supervivencia de la clase 2 : 87
Total de supervivencia de la clase 3 : 119
```

P(M/S)=(P(S/M)P(M))/(P(S))

```
Ns   = data.Survived.eq(1).sum()
Ns_m = (data.Survived.eq(1) & data.Sex.eq('female')).sum()

print('Aplicando la regla de Bayes se obtiene:')
print('Probabilidad de sobrevivir dado que se es mujer    : {:.2f} %'.format
      ((data.Survived.eq(1) & data.Sex.eq('female')).sum()/Ns*100))
print('Probabilidad de sobrevivir dado que se es hombre   : {:.2f} %'.format
      ((data.Survived.eq(1) & data.Sex.eq('male')).sum()/Ns*100))
print('Probabilidad de sobrevivir dado que se es mujer y está en la clase 1 : {:.2f} %'.format
      ((data.Survived.eq(1) & data.Sex.eq('female') & data.Pclass.eq(1)).sum()/Ns_m*100))
```

```
Aplicando la regla de Bayes se obtiene:
Probabilidad de sobrevivir dado que se es mujer    : 68.13 %
Probabilidad de sobrevivir dado que se es hombre   : 31.87 %
Probabilidad de sobrevivir dado que se es mujer y está en la clase 1 : 39.06 %
```