Predicción de Supervivencia en el Titanic y Análisis Interpretativo de Variables Socioeconómicas

Titanic-Machine learning from Disaster: Predict survival on the Titanic and get familiar with ML Basics

Grupo 3
Universidad de Málaga

Autor: Pablo Antonio García Pastor.

Área de conocimiento y departamento: Área de Análisis Matemático. Departamento de Análisis Matemático, Estadística e Investigación Operativa, y Matemática Aplicada.

Fecha de presentación: Noviembre de 2023

Tema: Investigación operativa

Número de páginas: 16

Índice

1. Introducción: contexto y planteamiento inicial

Hemos estudiado durante el curso conceptos básicos y algunos modelos de Inteligencia Artificial.

Planteamos en este trabajo una situación problemática cuya solución podemos encontrar en el contexto de la Inteligencia Artificial, es decir, resolveremos un problema a partir de un modelo matemático que trabaje con los correspondientes datos para dar una respuesta.

1.1. Situación problemática: ¿qué problema se va a estudiar y por qué?

El hundimiento del *Titanic* en 1912 constituye uno de los episodios más trágicos y, a la vez, más estudiados de la historia marítima moderna. Durante su viaje inaugural, el transatlántico británico colisionó con un iceberg y se hundió en el Atlántico Norte, provocando la muerte de 1.502 de los 2.224 pasajeros y tripulantes a bordo. Más allá del desastre técnico y humano, el suceso reveló profundas desigualdades sociales que influyeron directamente en las probabilidades de supervivencia de las personas a bordo.

Los registros históricos muestran que la supervivencia no dependió únicamente de factores azarosos, sino también de aspectos demográficos y socioeconómicos como el género, la edad o la clase del billete. La estructura jerárquica del Titanic reflejaba fielmente las divisiones de la sociedad eduardiana: los pasajeros de primera clase, pertenecientes en su mayoría a las élites económicas, disfrutaban de mejores camarotes, acceso prioritario a los botes salvavidas y proximidad a la tripulación; mientras que quienes viajaban en tercera clase, principalmente inmigrantes y familias trabajadoras, se encontraban en los niveles inferiores del barco, con un acceso mucho más limitado a las vías de escape.

En este contexto, el presente estudio busca construir un modelo predictivo capaz de responder a la pregunta: **¿qué características personales y sociales influyeron más en las probabilidades de supervivencia durante el naufragio del Titanic?** Para ello, se emplearán datos individuales de los pasajeros (edad, sexo, clase, número de familiares a bordo, entre otros) con el fin de identificar patrones que expliquen las desigualdades observadas en la supervivencia.

2. Metodología usada

Vamos a resolver entonces un problema de clasificación binaria dentro del aprendizaje supervisado, ya que utilizamos un conjunto de entrenamiento del que conocemos la variable respuesta, y con el que enseñaremos a nuestro modelo para generar la salida deseada.

La clasificación utiliza un algoritmo para asignar con precisión datos de prueba a categorías específicas. Los algoritmos de clasificación más comunes son los clasificadores lineales, las máquinas de vectores de soporte (SVM), los árboles de decisiones, el k-vecinos más cercanos y el bosque aleatorio. Nosotros hemos elegido el **Bosque Aleatorio**.

2.1. Bosque Aleatorio

El Bosque Aleatorio es un algoritmo de *Machine Learning* de uso común y registrado por Leo Breiman y Adele Cutler, cuyo resultado procede de la combinación de las salidas de múltiples árboles de decisión.

Los algoritmos de bosque aleatorio tienen cuatro hiperparámetros principales, que deben configurarse antes del entrenamiento:

1. Tamaño del nodo.
2. Cantidad de árboles.
3. Cantidad de características muestradas.
4. Profundidad máxima del árbol.

El algoritmo de bosque aleatorio se compone de un conjunto de árboles de decisión, en el que cada árbol se entrena con la llamada muestra de arranque, esta es una muestra de datos extraída del conjunto de entrenamiento con reemplazo. Este proceso se conoce como *bootstrapping*.

De esa muestra de arranque, un tercio se reserva como datos de prueba, lo que se conoce como muestra fuera de la bolsa (*out of bag*). Luego, se inyecta otra instancia de aleatoriedad a través del agrupamiento de características, ya que solo cogemos un número de características para cada árbol y se eligen de forma aleatoria, esto agrega más diversidad al conjunto de datos y reduce la correlación entre los árboles de decisión.

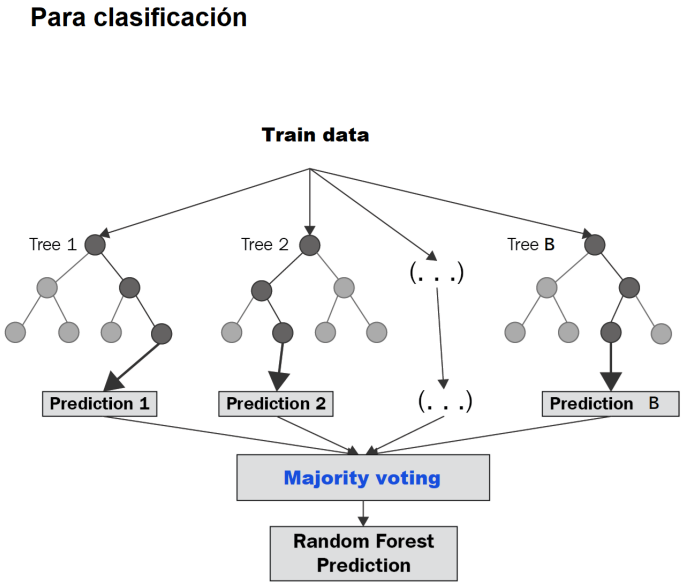


Figura 1: Ejemplo ilustrativo de bosque aleatorio.

Un ejemplo para visualizarlo mejor puede ser el siguiente:

Ejemplo. Imagina que estás tomando una decisión importante: ¿deberías ir de excursión o no?

- Podrías preguntarle a un amigo por su opinión.
- Luego, podrías preguntarle a otro amigo y otro más.
- Cada amigo tiene sus propias ideas y experiencias, por lo que te darán diferentes consejos.

Un bosque aleatorio funciona de manera similar:

- Tienes un montón de amigos llamados *árboles de decisión*. Cada árbol toma decisiones basadas en partes diferentes de la información.
- En lugar de confiar en un solo amigo (un solo árbol), preguntas a todos tus amigos (muchos árboles).
- Luego, se cuentan los votos de tus amigos para tomar una decisión. La mayoría gana.

Entonces, un bosque aleatorio es como consultar a varios amigos (árboles de decisión) para tomar una decisión más segura y acertada. Ayuda a hacer predicciones más precisas en problemas. ¡Es como el "el poder de la multitud" en el aprendizaje automático!

2.2. Ventajas e inconvenientes del método

Algunas ventajas de este método son:

- Para conjuntos de datos grandes resulta un algoritmo eficiente y certero.
- Se puede utilizar para tareas de regresión y clasificación y también es una herramienta eficaz para estimar valores perdidos.
- Riesgo reducido de sobreajuste para un número elevado de árboles y una elección adecuada de hiperparámetros.
- La selección aleatoria de variables explicativas con las que se llevan a cabo las divisiones en los nodos de cada árbol del bosque introduce más aleatoriedad, que de nuevo puede prevenir de sobreajuste o correlación entre árboles.
- Maneja cientos de variables explicativas sin excluir ninguna.

Por otro lado, el método también cuenta con inconvenientes:

- El proceso puede requerir mucho tiempo: al manejar un gran número de árboles de decisión podemos dar predicciones más precisas pero el proceso será más lento.
- Consumo de memoria: cuando trabajamos con bosques con muchos árboles muy profundos y con grandes cantidades de datos, el algoritmo consume mucha memoria.
- En ocasiones puede ser más difícil de interpretar que otros algoritmos como los árboles de decisión o la regresión.
- En caso de tener escasos datos, otros métodos pueden ser más eficientes que el bosque aleatorio.

3. Implementación práctica

3.1. Base de datos

Utilizamos una base de datos obtenida de Kaggle que contiene información sobre los pasajeros a bordo del Titanic. La base de datos incluye datos como la edad de los pasajeros, el número de hermanos, hermanas, esposos o esposas, así como el número de padres e hijos a bordo. Además, se registra la clase del pasajero, clasificada como primera, segunda o tercera, que refleja la calidad del viaje. Otros datos disponibles son el precio del ticket pagado por cada pasajero y el puerto desde el cual embarcaron.

Además, la base de datos también incluía información como el nombre de cada pasajero, el número de ticket asignado y el número de cabina. Sin embargo, desde el principio decidimos descartar estos datos. Considerando que el nombre y el número de ticket no aporta información relevante al modelo. En cuanto a la información de la cabina, la gran mayoría de los pasajeros no contaba con este dato conocido.

Aquí tenemos una leyenda con el significado de cada una de las variables explicativas de nuestra base de datos y un ejemplo de algunos pasajeros de la base de datos.

Variable	**Definición**	**Clave**
`survival`	Supervivencia	0 = No, 1 = Sí
`pclass`	Clase del ticket	1 = primera, 2 = segunda, 3 = tercera
`sex`	Sexo	
`Age`	Edad en años	
`sibsp`	Número de hermanos/esposas a bordo del Titanic	
`parch`	Número de padres/hijos a bordo del Titanic	
`ticket`	Número de ticket	
`fare`	Tarifa del pasajero	
`cabin`	Número del camarote	
`embarked`	Puerto de embarcación	`C` = Cherbourg, `Q` = Queenstown, `S` = Southampton

Cuadro 1: Leyenda de variables.

Cuadro 2: Ejemplo de la base de datos.

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
1	0	3	male	22	1	0	7.25	S
2	1	1	female	38	1	0	712.833	C
3	1	3	female	26	0	0	7.925	S
4	1	1	female	35	1	0	53.1	S
5	0	3	male	35	0	0	8.05	S
6	0	3	male	49	0	0	8.4583	Q
7	0	1	male	54	0	0	518.625	S
8	0	3	male	2	3	1	21.075	S
9	1	3	female	27	0	2	11.1333	S
10	1	2	female	14	1	0	300.708	C

3.1.1. Tratamiento de las variables 'Embarked' y 'Fare'

Un paso esencial en la preparación de los datos es el manejo de valores ausentes. Las variables `Embarked` (puerto de embarque) y `Fare` (tarifa) contenían un número muy reducido de datos faltantes. Dada la baja cantidad de valores nulos, se optó por una estrategia de imputación simple para evitar la pérdida de registros completos, que de otro modo tendrían que ser descartados.

Puerto de Embarque (`Embarked`): Siendo una variable categórica, la estrategia más adecuada para la imputación es utilizar la **moda**, es decir, el valor más frecuente en el conjunto de datos. Este método asegura que se rellena el valor faltante con la categoría más probable, alterando mínimamente la distribución original de la variable. En este caso, el puerto más común fue Southampton ('S').

Tarifa del Billete (`Fare`): Para la variable numérica `Fare`, se optó por imputar los valores faltantes utilizando la **mediana** en lugar de la media. La elección de la mediana es deliberada: la distribución de las tarifas es muy asimétrica (sesgada a la derecha), con unos pocos pasajeros pagando tarifas extremadamente altas por suites de lujo. La media es muy sensible a estos valores atípicos, mientras que la mediana es un estadístico robusto que no se ve afectado por ellos, representando así de forma más fidedigna el valor çentral"de la tarifa para la mayoría de los pasajeros.

Con estas dos operaciones, el conjunto de datos queda libre de valores nulos en estas columnas, preparándolo para las fases posteriores de entrenamiento del modelo sin introducir un sesgo significativo.

3.2. Variable Edad

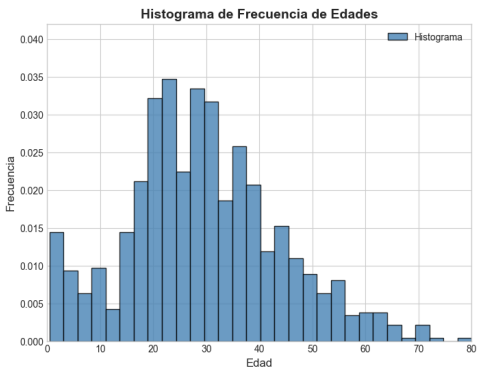
A la hora de trabajar con la variable *edad* tenemos un inconveniente, pues hay pasajeros que no tienen edad, y pensé principalmente en dos soluciones para tratar esta variable en los pasajeros:

1. Fijar una edad común para todos a los que le faltan usando una medida de tendencia central como la media o la mediana.

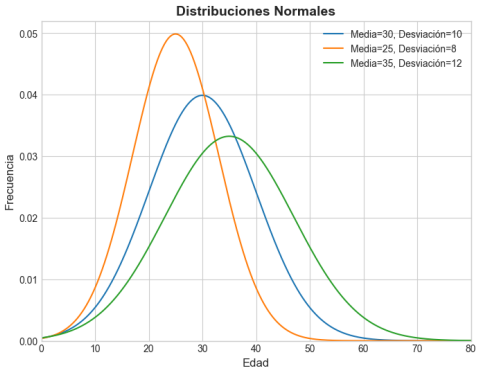
2. Generar valores de una distribución conocida (como puede ser la distribución Normal o Gamma) que se aproxime a como está distribuida la edad conocida de los pasajeros de nuestro problema, y asignarle estos valores a los que le falten.

Terminamos optando por la segunda opción ya que nos parecía una mejor solución pues es más realista al asignar distintos valores a los pasajeros sin edad en vez de atribuir a todos un mismo valor.

Nuestro primer paso era elegir qué distribución conocida aproximaba de forma razonable la variable edad, para ello creamos un histograma de frecuencia de la edad de los pasajeros conocida.



(a) Histograma edad



(b) Distribuciones normales

Figura 2: Análisis de la distribución de la edad.

Como podemos comprobar en las dos gráficas de arriba parece que si consideramos una distribución normal con parámetros adecuados vamos a obtener una buena aproximación de nuestra distribución de la edad.

Para poder hallar esta distribución usaremos el método de estimación por máxima verosimilitud. Si tenemos una muestra $x_1, x_2, \ldots, x_n$ de observaciones independientes, la función de verosimilitud es:

$$L(\theta : x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(\theta : x_i) \quad \theta \in \Theta$$

Para una distribución normal $N(\mu, \sigma)$, la función de verosimilitud es:

$$L(\mu, \sigma : x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} = \frac{1}{(\sigma\sqrt{2\pi})^n} \prod_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}$$

Para maximizar, tomamos logaritmos y derivamos respecto a los parámetros μ y σ:

$$l(\mu, \sigma) = \log(L) = -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \mu}{\sigma}\right)^2$$

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0$$

$$\frac{\partial l(\mu,\sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

Por lo tanto tenemos que resolver el sistema en función de μ y σ:

$$0 = \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma} = \frac{n}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\mu\right) = \frac{n\bar{x}-n\mu}{\sigma} \iff \mu = \bar{x}$$

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \iff n = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Resolviendo el sistema, los estimadores de máxima verosimilitud son la media y la desviación típica de las edades, respectivamente. Calculándolo nos queda que $\mu = 29,7$ y $\sigma = 14,5$. Nuestra distribución que generará las edades es: $N(\mu = 29,7, \sigma = 14,5)$.

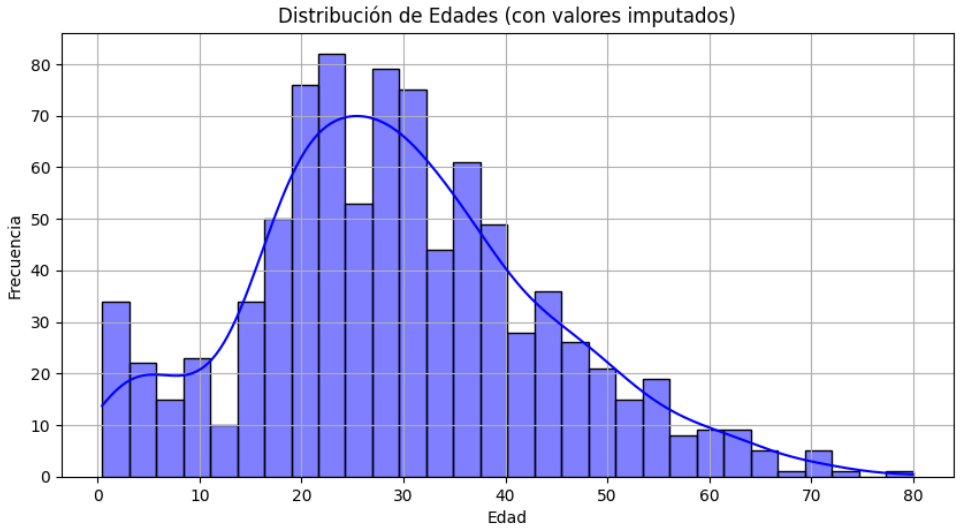


Figura 3: Distribución Normal e Histograma.

Luego efectivamente en esta gráfica podemos ver que la distribución normal parece una buena aproximación, se aprecia mejor en los siguientes dos histogramas.

3.4. Implementación y Optimización del Modelo

Para el desarrollo del modelo predictivo se empleó el lenguaje de programación **Python**, apoyado en bibliotecas clave como **pandas** para la manipulación de datos y **scikit-learn** para la implementación del `RandomForestClassifier` y las métricas de evaluación.

Proceso de Optimización en Dos Fases

Tras obtener una precisión inicial, se abordó un riguroso proceso de optimización de hiperparámetros en dos fases principales para mejorar el rendimiento del modelo.

Fase 1: Búsqueda Exhaustiva con `GridSearchCV` Primero, se utilizó `GridSearchCV` de scikit-learn para realizar una búsqueda exhaustiva y sistemática de la mejor combinación de hiperparámetros. Se definió un espacio de búsqueda (`param_grid`) que incluía:

- **Número de árboles** (`n_estimators`)**:** `[10, 20, 47, 50, 60]`
- **Profundidad máxima** (`max_depth`)**:** `[5, 10, 11, 20, 30, 40, 50, 60]`
- **Muestras mínimas por hoja** (`min_samples_leaf`)**:** `[1, 2, 4]`
- **Muestras mínimas para dividir** (`min_samples_split`)**:** `[2, 5, 10]`

`GridSearchCV` fue configurado para evaluar cada combinación utilizando una **validación cruzada de 5 pliegues** (`cv=5`), asegurando que los resultados fueran robustos y no dependieran de una única partición de los datos. Este proceso automatizado permitió explorar una amplia gama de configuraciones para encontrar la más prometedora.

Lo cual nos terminó dando: `n_estimators`: 11, `max_depth`: 1, `min_samples_leaf`: 5, `min_samples_split`: 20.

Fase 2: Análisis Visual y Optimización Fina con Out-of-Bag (OOB) Score Para profundizar en la relación entre los dos hiperparámetros más influyentes —`n_estimators` y `max_depth`—, se realizó un segundo análisis enfocado en la visualización. Se programó un doble bucle para iterar sobre rangos específicos:

- **Profundidad** (`max_depth`)**:** de 1 a 40.
- **Número de árboles** (`n_estimators`)**:** de 10 a 40.

Para cada par de valores, se entrenó un modelo y se calculó su precisión utilizando la métrica **Out-of-Bag (OOB) score**. Esta métrica evalúa el rendimiento del modelo utilizando los datos que no fueron seleccionados para el entrenamiento de cada árbol, funcionando como un método de validación interna muy eficiente.

Los resultados de este análisis se representaron en un gráfico de superficie 3D, que muestra cómo varía la precisión en función de los dos hiperparámetros. Donde nos terminó por confirmar que: `n_estimators`: 11, `min_samples_split`: 21. Con un **OOB Score: 0.836**

Resultados y Mejora Final

El análisis visual permitió identificar con claridad la combinación óptima de hiperparámetros. Como se observa en el gráfico, el punto de máxima precisión (marcado en rojo) se alcanzó con `n_estimators`: 11, `min_samples_split`: 21.

na vez seleccionados los hiperparámetros óptimos mediante validación cruzada, se procedió a entrenar el modelo final con la totalidad del conjunto de entrenamiento. Posteriormente, su rendimiento se evaluó de forma definitiva utilizando el conjunto de prueba, datos que el modelo no había visto previamente.

3.2.1. Métricas de Rendimiento del Modelo Inicial

La precisión global (*accuracy*) obtenida en el conjunto de prueba fue del **76.87 %**. Sin embargo, un análisis más detallado de la matriz de confusión y el reporte de clasificación reveló un rendimiento desigual entre las clases.

La matriz de confusión resultante fue la siguiente:

El análisis del reporte de clasificación arrojó las siguientes métricas clave:

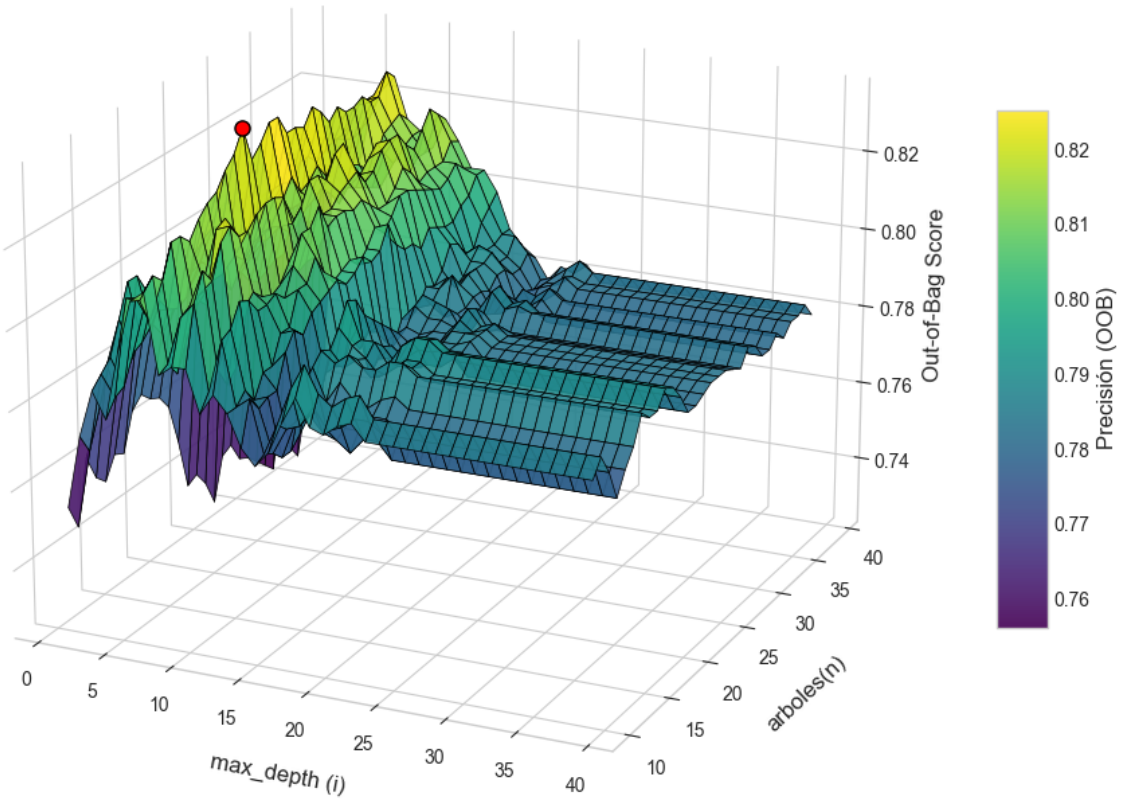


Figura 4: Ejemplo de Random Forest con un dato.

	Clase Predicha	
Clase Real	No Sobrevivió (0)	Sobrevivió (1)
No Sobrevivió (0)	65 (VN)	13 (FP)
Sobrevivió (1)	18 (FN)	38 (VP)

Cuadro 3: Matriz de confusión en el conjunto de prueba.

- **Clase 0 (No Sobrevivió):** Precisión de 0.78 y un *Recall* de 0.83.
- **Clase 1 (Sobrevivió):** Precisión de 0.75 y un *Recall* de solo **0.68**.

3.2.2. Diagnóstico del Problema: Desbalance de Clases

El análisis de estas métricas reveló un desbalance en el rendimiento del modelo. Mientras que el modelo era muy eficaz identificando a los pasajeros que no sobrevivieron (*Recall* del 83 %), mostraba una debilidad significativa para identificar a los que sí sobrevivieron (*Recall* de solo el 68 %).

Esta baja sensibilidad para la clase positiva se tradujo en un número elevado de **Fal-**

sos Negativos (18 casos), donde el modelo predijo incorrectamente que un pasajero no sobrevivió. Este comportamiento es característico de modelos entrenados con datos donde las clases no están perfectamente equilibradas, lo que provoca un sesgo del modelo hacia la predicción de la clase mayoritaria.

3.2.3. Solución Implementada: Ajuste de Pesos por Clase

Para mitigar este sesgo y mejorar la capacidad del modelo para identificar correctamente a la clase minoritaria, se decidió re-entrenar el modelo final aplicando un ajuste de pesos. Esto se logró configurando el hiperparámetro `class_weight='balanced'` en la instancia del clasificador `RandomForestClassifier`.

Esta técnica penaliza de forma más severa los errores de clasificación en la clase minoritaria (supervivientes), obligando al modelo a prestarle mayor atención durante el entrenamiento y, en consecuencia, a mejorar su *Recall*. La aplicación de esta técnica resultó ser altamente efectiva, elevando la precisión final del modelo al **82 %** y logrando un rendimiento más equilibrado entre ambas clases.

3.3. Resultados del Modelo Final Optimizado

Tras aplicar la técnica de ajuste de pesos de clase (`class_weight='balanced'`) para mitigar el sesgo del modelo, se realizó una evaluación final sobre el conjunto de prueba. Los resultados muestran una mejora significativa tanto en la precisión general como en el equilibrio del rendimiento entre las clases.

La precisión global (*accuracy*) del modelo optimizado alcanzó el **82.09 %**.

3.3.1. Matriz de Confusión Final

La matriz de confusión detalla la distribución de aciertos y errores del modelo:

Cuadro 4: Matriz de confusión del modelo optimizado en el conjunto de prueba.

	Clase Predicha	
Clase Real	No Sobrevivió (0)	Sobrevivió (1)
No Sobrevivió (0)	69 (VN)	9 (FP)
Sobrevivió (1)	15 (FN)	41 (VP)

Se observa una notable reducción en los Falsos Negativos (de 18 a 15) en comparación con el modelo inicial, indicando una mejor capacidad para identificar a los supervivientes.

3.3.2. Reporte de Clasificación Detallado

El reporte de clasificación confirma el rendimiento equilibrado del modelo. Las métricas de precisión (*precision*) y sensibilidad (*recall*) son ahora más consistentes entre ambas clases.

El *Recall* para la clase 1 (Sobrevivió) aumentó del 68 % al **73 %**, solucionando la debilidad principal identificada en el modelo anterior. La precisión para ambas clases se estabilizó en un sólido 82 %, lo que demuestra la robustez del clasificador final.

Cuadro 5: Reporte de clasificación detallado del modelo optimizado.

Clase	Precision	Recall	F1-Score	Support
0 (No Sobrevivió)	0.82	0.88	0.85	78
1 (Sobrevivió)	0.82	0.73	0.77	56
Promedio Ponderado	*0.82*	*0.82*	*0.82*	*134*

3.4. Interpretación de los resultados

El análisis de los resultados nos permite comprender qué factores fueron determinantes para la supervivencia en el desastre del Titanic, basándonos en la información visual proporcionada por las figuras 6 y 7.

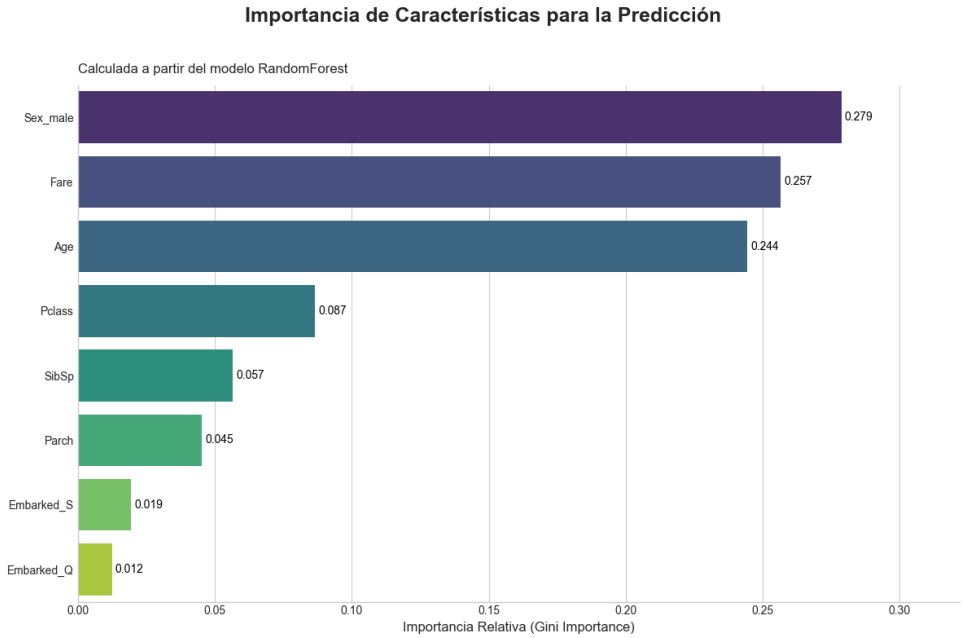


Figura 5: Importancia de las variables en el modelo.

La **Figura 6: Importancia de Características para la Predicción** nos ofrece una visión jerárquica de las variables según su poder predictivo dentro del modelo *Random Forest*. Esta gráfica no mide la tasa de supervivencia directamente, sino la influencia que cada variable tiene en las decisiones del modelo. Los puntos clave son:

- **Sexo del pasajero (`Sex_male`):** Es, con diferencia, la variable más influyente. El modelo aprendió que conocer el sexo de una persona era el factor más decisivo para predecir si sobreviviría o no.
- **Tarifa (`Fare`) y Edad (`Age`):** Estas dos variables ocupan el segundo y tercer lugar con una importancia muy similar. Indican que el estatus socioeconómico (reflejado en el precio del billete) y la edad del pasajero fueron también factores cruciales.
- **Clase (`Pclass`):** Aunque relevante, su importancia es considerablemente menor que las tres anteriores. Esto sugiere que, si bien la clase era importante, la tarifa pagada (que está correlacionada con la clase pero es más granular) ofrecía un poder predictivo mayor.

- **Variables familiares y de embarque:** El número de familiares a bordo (`SibSp`, `Parch`) y el puerto de embarque (`Embarked`) tuvieron una importancia predictiva mucho menor en comparación.

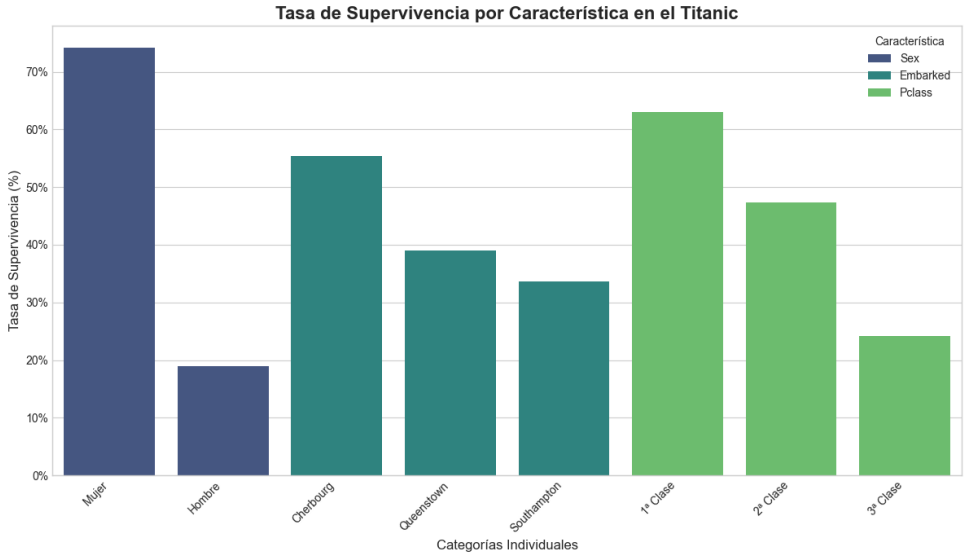


Figura 6: Histograma del porcentaje de supervivencia por categorías.

La **Figura 7: Tasa de Supervivencia por Característica en el Titanic** complementa el análisis anterior, ya que visualiza las tasas de supervivencia reales para las categorías más importantes, explicando *por qué* el modelo les dio esa importancia.

- **Sexo:** La gráfica confirma de manera contundente la importancia de esta variable. La tasa de supervivencia de las **mujeres** fue extraordinariamente alta (superior al 70 %), mientras que la de los **hombres** fue drásticamente baja (alrededor del 20 %). Esto valida la hipótesis histórica del protocolo de "mujeres y niños primero".

- **Clase (Pclass):** Se observa una clara correlación entre la clase socioeconómica y la supervivencia. Los pasajeros de **1ª Clase** tuvieron una tasa de supervivencia superior al 60 %, los de **2ª Clase** se situaron cerca del 50 %, y los de **3ª Clase** tuvieron la menor probabilidad, con una tasa de apenas el 25 %.

- **Puerto de Embarque (Embarked):** Los pasajeros que embarcaron en **Cherbourg (C)** muestran una tasa de supervivencia notablemente más alta (cercana al 55 %) en comparación con los de **Queenstown (Q)** y, especialmente, **Southampton (S)**. Esto a menudo se correlaciona con el hecho de que una mayor proporción de pasajeros de primera clase embarcó en Cherbourg.

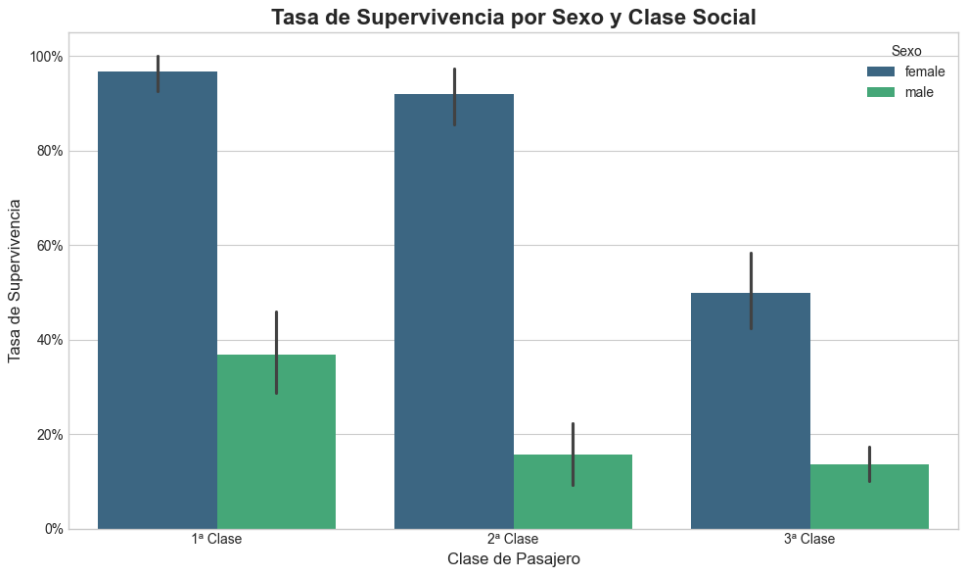


Figura 7: Histograma del porcentaje de supervivencia por clase y sexo.

3.4.1. Interacción entre Sexo y Clase Social

El primer gráfico **Figura 8: Tasa de Supervivencia por Característica en el Titanic** de interacción, *Tasa de Supervivencia por Sexo y Clase Social*, analiza si el beneficio de ser mujer era uniforme en todas las clases sociales. Los resultados son reveladores:

- **Privilegio condicionado por la clase:** Aunque las mujeres tuvieron sistemáticamente una tasa de supervivencia mayor que los hombres en todas las clases, el estatus socioeconómico fue un factor crítico. Las mujeres de **1ª y 2ª Clase** tuvieron una probabilidad de supervivencia altísima, superior al 90 %.
- **La vulnerabilidad de la 3ª Clase:** La tasa de supervivencia para las mujeres de **3ª Clase**, en cambio, se desplomó hasta aproximadamente el 50 %. Si bien seguía siendo más del doble que la de los hombres de su misma clase, demuestra que el protocolo de "mujeres y niños primero"no se aplicó con la misma eficacia para las pasajeras de menor estatus.
- **Gradiente en los hombres:** Entre los hombres también se observa una clara jerarquía. Un pasajero de 1ª Clase tuvo una probabilidad de sobrevivir significativamente mayor que uno de 2ª, y este a su vez, que uno de 3ª, cuya tasa fue inferior al 15 %.

Este gráfico demuestra que la supervivencia no dependía de un único factor, sino de la intersección de privilegios sociales, donde el género y la clase se reforzaban mutuamente.

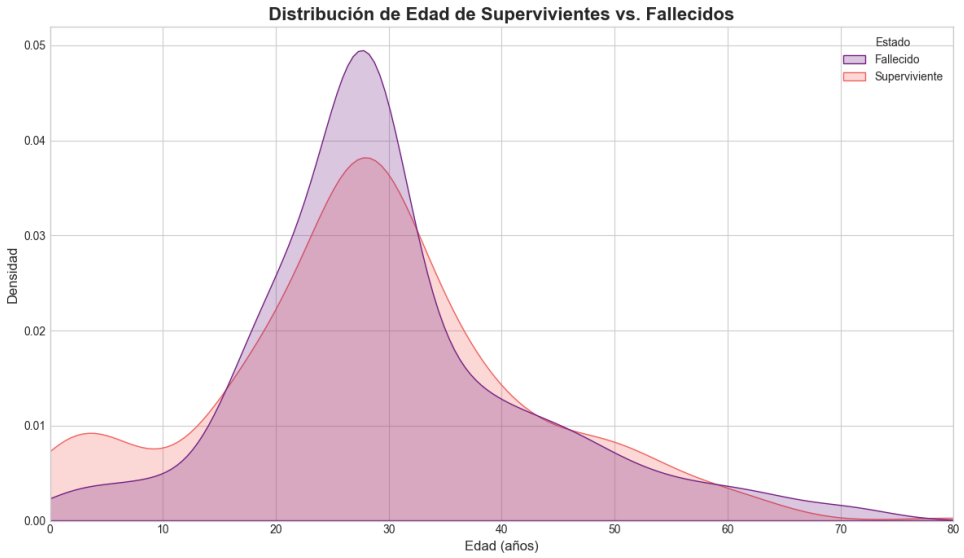


Figura 8: Distribución de Edad de Supervivientes vs. Fallecidos.

3.4.2. Impacto de la Edad en la Supervivencia

El segundo gráfico, **Figura 8: Distribución de Edad de Supervivientes vs. Fallecidos**, muestra el impacto de la edad, confirmando el papel crucial de esta variable:

- **Prioridad absoluta a la infancia:** La distribución de los supervivientes muestra un pico extremadamente pronunciado para los niños de entre 0 y 10 años. Esto ofrece una evidencia visual contundente de que la parte de "niños primero"del protocolo de evacuación se siguió rigurosamente.
- **La generación sacrificada:** La mayor concentración de víctimas se encuentra en el rango de edad de 20 a 40 años. Este grupo, compuesto por adultos jóvenes y de mediana edad (principalmente hombres), representaba el grueso de los pasajeros pero tuvo la menor prioridad durante el rescate.
- **Vulnerabilidad en la vejez:** Aunque había menos pasajeros de edad avanzada a bordo, la distribución sugiere que su tasa de supervivencia fue también muy baja, probablemente debido a dificultades de movilidad en una situación de pánico.

En conclusión, la interpretación conjunta de ambas gráficas demuestra que el modelo no solo identificó correctamente las variables más predictivas, sino que estas se alinean con patrones lógicos y observables en los datos. La supervivencia en el Titanic no fue un suceso aleatorio, sino que estuvo fuertemente determinada por el sexo, el estatus socioeconómico (reflejado en la tarifa y la clase) y la edad de los pasajeros.

4. Conclusiones

Este estudio ha logrado no solo construir un modelo predictivo con una precisión final del **82.09 %**, sino también desentrañar la compleja red de factores humanos y sociales que determinaron quién vivió y quién murió en el desastre del Titanic.

Las conclusiones principales revelan una clara jerarquía de supervivencia, donde el destino de un pasajero estaba fuertemente predeterminado por sus características demográficas:

1. **El Sexo fue el factor más determinante**, con una abrumadora ventaja para las mujeres, lo que confirma la aplicación del código de conducta marítimo "mujeres y niños primero".

2. **La Edad fue el segundo pilar del protocolo de rescate**, donde se priorizó de forma explícita a los niños, resultando en su alta tasa de supervivencia en comparación con los adultos.

3. **La Clase Social actuó como un modulador implacable de los dos factores anteriores.** El privilegio de ser mujer o niño se magnificaba enormemente en la 1ª y 2ª Clase y se atenuaba drásticamente en la 3ª. Para los hombres, la clase alta ofrecía la única esperanza tangible de supervivencia.

En definitiva, la tragedia del Titanic no fue un ecualizador social. Por el contrario, la presión extrema de la catástrofe exacerbó las divisiones sociales existentes, demostrando que en la lucha por la vida, las normas sociales y la jerarquía económica fueron tan decisivas como los botes salvavidas.

Fin