# ANÁLISIS PREDICTIVO DE SUPERVIVENCIA EN EL TITANIC: APLICACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA LA CLASIFICACIÓN DE PASAJEROS

Juan J. Bonilla, Ricardo Muñoz, Valentina Isaza, Nelcy L. Zapata,

*Estudiantes de Maestría en Inteligencia Artificial y Ciencia de Datos*
*Universidad Autónoma de Occidente, Cll 25 # 115-85 Km 2 Vía Cali - Jamundí · Cali, Colombia,*

***Resumen*— Este trabajo aborda el problema de predecir la supervivencia de los pasajeros del Titanic mediante técnicas de aprendizaje automático. A partir de variables como edad, género, clase social y relaciones familiares, se entrenaron modelos supervisados que permitieron identificar patrones asociados a la supervivencia.**

**Los resultados confirman que ser mujer, viajar en primera clase y pagar una tarifa alta aumentaban significativamente la probabilidad de sobrevivir, reflejando decisiones sociales y desigualdades estructurales de la época. Los modelos también evidencian la desventaja de hombres y pasajeros de tercera clase, posiblemente por su menor acceso a los botes salvavidas.**

**Entre los algoritmos evaluados, XGBoost y SVM ofrecieron el mejor equilibrio entre precisión y sensibilidad, mientras que la Regresión Logística, aunque más simple, mostró un rendimiento competitivo y alta interpretabilidad, ideal para una primera aproximación al problema.**

***Abstract*-- This paper addresses the problem of predicting the survival of Titanic passengers using machine learning techniques. Based on variables such as age, gender, social class and family relationships, supervised models were trained to identify patterns associated with survival.**

**The results confirm that being a woman, traveling first class and paying a high fare significantly increased the probability of survival, reflecting social decisions and structural inequalities of the time. The models also show the disadvantage of men and third-class passengers, possibly due to their lower access to lifeboats.**

**Among the algorithms evaluated, XGBoost and SVM offered the best balance between accuracy and sensitivity, while Logistic Regression, although simpler, showed competitive performance and high interpretability, ideal for a first approach to the problem.**

## I. INTRODUCCION

El análisis predictivo basado en datos históricos ha emergido como una herramienta fundamental en el campo de la ciencia de datos y la inteligencia artificial. Un caso representativo ampliamente utilizado con fines educativos es el desastre del RMS Titanic, ocurrido en 1912, donde más de 1,500 personas perdieron la vida tras el hundimiento del transatlántico. Este evento, si bien es histórico, proporciona un contexto ideal para aplicar y demostrar técnicas modernas de aprendizaje automático en un entorno controlado. El presente estudio aborda el problema de predecir la probabilidad de supervivencia de los pasajeros del Titanic, utilizando como base el conjunto de datos proporcionado por la plataforma Kaggle. A partir de variables como la edad, el sexo, la clase socioeconómica (Pclass), el número de familiares a bordo y el puerto de embarque, se pretende construir un modelo supervisado capaz de identificar patrones ocultos y realizar predicciones precisas sobre la supervivencia de los individuos.

Este problema, más allá de su relevancia histórica, permite a los investigadores y estudiantes practicar conceptos clave del flujo de trabajo en ciencia de datos, incluyendo el análisis exploratorio, la imputación de valores faltantes, la ingeniería de características, el modelado supervisado y la validación cruzada de modelos. Asimismo, se explora el uso de diversos algoritmos de clasificación, tales como regresión logística, bosques aleatorios (Random Forest) y el método de gradiente extremo (XGBoost), evaluando su rendimiento mediante métricas como la precisión, el recall y el F1-score.

## II. OBJETIVO

El propósito de este trabajo es desarrollar e implementar un modelo de aprendizaje automático supervisado que permita predecir la probabilidad de supervivencia de los pasajeros del RMS Titanic, a partir de las variables proporcionadas en el conjunto de datos original distribuido por Kaggle. Para ello, se empleará un enfoque sistemático que incluye la selección y transformación de características relevantes, la imputación de valores faltantes, y la validación de múltiples algoritmos de clasificación. El modelo será entrenado sobre un subconjunto del conjunto de datos (titanic_training.csv) y evaluado utilizando datos reservados, con el objetivo de maximizar su capacidad predictiva a través de métricas de desempeño como la precisión, el recall y el puntaje F1. La propuesta busca garantizar la replicabilidad del proceso y ofrecer una base sólida para futuras mejoras en entornos educativos y profesionales

## III. PROBLEMA A ABORDAR

Este proyecto se propone abordar la siguiente pregunta: ¿qué tipo de personas tenían mayor probabilidad de sobrevivir al hundimiento del Titanic? Para responder a este interrogante, se plantea la construcción de un modelo de aprendizaje automático supervisado que permita predecir, con base en datos históricos, si un pasajero sobrevivió o no al desastre marítimo. La predicción se fundamenta en variables relevantes como la edad, el género, la clase socioeconómica (Pclass) y los vínculos familiares a bordo (SibSp y Parch). El problema se enmarca dentro de una tarea de clasificación binaria, donde la variable objetivo es la supervivencia del pasajero. La solución implica identificar patrones en los datos que expliquen la relación entre los atributos individuales y el resultado observado.

## IV. ANTECEDENTES

El hundimiento del RMS Titanic en 1912 ha sido objeto de numerosos estudios en el ámbito del aprendizaje automático, debido a la disponibilidad de datos detallados sobre los pasajeros y las circunstancias del desastre. Diversas investigaciones han explorado la aplicación de algoritmos de clasificación para predecir la supervivencia de los pasajeros, considerando variables como la edad, el sexo, la clase del boleto y el punto de embarque.

Sherlock et al. [1] utilizaron herramientas de minería de datos, específicamente Weka, para analizar la relación entre características de los pasajeros y su probabilidad de supervivencia. El estudio identificó que la clase de cabina, la edad y el puerto de embarque eran factores significativos que influían en las tasas de supervivencia.

Liang [2] aplicó algoritmos de árboles de decisión y bosques aleatorios para predecir la supervivencia de los pasajeros. Los resultados mostraron que la edad, el sexo y la clase del boleto eran las variables más correlacionadas con la supervivencia. Además, se observó que el algoritmo de árbol de decisión superó al de bosque aleatorio en términos de precisión en este contexto específico.

Amalia y Rahayu [3] compararon múltiples modelos de clasificación, incluyendo regresión logística, bosques aleatorios y XGBoost, para predecir la supervivencia de los pasajeros del Titanic. El estudio concluyó que el modelo de bosques aleatorios alcanzó la mayor precisión, destacando la eficacia de los métodos de ensamblado en conjuntos de datos con variables mixtas.

Liao et al. [4] propusieron un modelo de votación dura que combina regresión logística, bosques aleatorios y árboles de decisión para predecir la supervivencia de los pasajeros. Este enfoque integrador logró una precisión del 87.64%, evidenciando que la combinación de clasificadores puede mejorar la robustez de las predicciones en conjuntos de datos estructurados.

Estos estudios proporcionan una base sólida para el desarrollo de modelos predictivos en el contexto del Titanic, y sirven como referencia para la implementación y evaluación de algoritmos de aprendizaje automático en problemas de clasificación binaria.

## V. ANÁLISIS EXPLORATORIO DE DATOS

**Preparación y Exploración del Conjunto de Datos**

*A. Importación de Librerías*

Para el análisis exploratorio y el preprocesamiento de datos, se emplearon bibliotecas ampliamente utilizadas en el ecosistema de ciencia de datos en Python: pandas y numpy para la manipulación de estructuras de datos, matplotlib y seaborn para la visualización gráfica, y scikit-learn para el modelado predictivo y la transformación de variables. También se incluyó scipy.stats para pruebas estadísticas y google.colab.drive para la integración con Google Drive, que sirvió como fuente de datos.

*B. Carga del Conjunto de Datos*

El conjunto de entrenamiento se obtuvo desde la ruta especificada en Google Drive, correspondiente al archivo titanic_training.csv, que contiene 999 registros con información de pasajeros a bordo del Titanic.

*C. Contexto del Dataset*

El conjunto de datos se enmarca en el evento histórico del hundimiento del RMS Titanic ocurrido entre el 14 y 15 de abril de 1912, durante su viaje inaugural entre Southampton (Reino Unido) y Nueva Escocia (Canadá). Tras una colisión con un iceberg, el navío se hundió en menos de tres horas, causando la muerte de 1502 personas de un total de 2224 entre pasajeros y tripulación. El número insuficiente de botes salvavidas y las condiciones de abordaje influyeron en la tasa de supervivencia.

El dataset contiene información relevante de los pasajeros, con variables tanto demográficas como operativas, que pueden ser determinantes para predecir la supervivencia. La tabla I resume las características incluidas en el conjunto:

| *Variable* | *Descripción* | *Valores* |
|---|---|---|
| *survival* | Sobrevivió? | 0 = No, 1 = Sí |
| *pclass* | Clase del boleto de embarque | 1 = primera, 2 = segunda, 3 = tercera |
| *sex* | Sexo | male = Masculino, female = Femenino |
| *age* | Edad en años | |
| *sibsp* | Cantidad de hermanos o esposas del sujeto a bordo | |
| *parch* | Cantidad de padres o hijos del sujeto a bordo Titanic | |
| *ticket* | Número de pase de abordaje | |
| *fare* | Tarifa de embarque | |
| *cabin* | Número de habitación | |
| *embarked* | Puerto de embarque | C = Cherbourg, Q = Queenstown, S = Southampton |

*Tabla Número 1 – Contexto Dataset*

### *D. Exploración Inicial del Dataset*

El análisis preliminar del DataFrame reveló un total de 999 entradas y 10 columnas. El resumen estadístico mostró la presencia de valores nulos en múltiples variables:

| *N°* | *Columna* | *Valores no nulos* | *Tipo de dato* | *Observaciones* |
|---|---|---|---|---|
| *0* | survived | 999 | float64 | Variable objetivo (0 = No, 1 = Sí) |
| *1* | pclass | 999 | float64 | Clase del boleto (1, 2, 3) |
| *2* | sex | 999 | object | Género (male, female) |
| *3* | age | 804 | float64 | Edad (valores faltantes) |
| *4* | sibsp | 999 | float64 | N° de hermanos/cónyuge a bordo |
| *5* | parch | 999 | float64 | N° de padres/hijos a bordo |
| *6* | ticket | 999 | object | Número de boleto |
| *7* | fare | 998 | float64 | Tarifa pagada |
| *8* | cabin | 227 | object | Cabina asignada (muchos nulos) |
| *9* | embarked | 997 | object | Puerto de embarque (C, Q, S) |

*Tabla Número 2 – Exploración Dataset*

Se observó que las variables age, fare, cabin y embarked presentaban valores faltantes en proporciones variables. Particularmente, cabin tenía una alta proporción de datos ausentes (~77.3%), por lo que su uso fue posteriormente reconsiderado.

Además, se identificó un registro con valor nulo en la variable survived, que representa la etiqueta objetivo. Este registro fue descartado del análisis, dado que su ausencia impide el entrenamiento supervisado del modelo.

### *E. Codificación Ordinal de Variables Categóricas*

Dado que los algoritmos de machine learning requieren datos numéricos, las variables categóricas sex, embarked y cabin fueron codificadas utilizando OrdinalEncoder, Con ello se obtiene DataFrame Titanic Codificado y posterior a ello se generaron las estadísticas descriptivas del DataFrame transformado

| *Estadística* | *survived* | *pclass* | *age* | *sibsp* | *parch* | *fare* | *sex_encoded* | *embarked_encoded* | *cabin_encoded* |
|---|---|---|---|---|---|---|---|---|---|
| *count* | 999.0 | 999.0 | 804.0 | 999.0 | 999.0 | 998.0 | 999.0 | 997.0 | 227.0 |
| *mean* | 0.386 | 2.287 | 30.26 | 0.504 | 0.409 | 34.622 | 0.651 | 1.518 | 78.194 |
| *std* | 0.487 | 0.844 | 14.63 | 1.058 | 0.897 | 54.390 | 0.477 | 0.801 | 43.059 |
| *min* | 0.0 | 1.0 | 0.167 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *25%* | 0.0 | 1.0 | 21.0 | 0.0 | 0.0 | 7.896 | 0.0 | 1.0 | 42.5 |
| *50%* | 0.0 | 3.0 | 28.0 | 0.0 | 0.0 | 14.458 | 1.0 | 2.0 | 78.0 |
| *75%* | 1.0 | 3.0 | 39.0 | 1.0 | 0.0 | 32.116 | 1.0 | 2.0 | 113.5 |
| *max* | 1.0 | 3.0 | 80.0 | 8.0 | 9.0 | 512.329 | 1.0 | 2.0 | 152.0 |

*Tabla Número 3 – Codificación Dataset*

## Análisis de Correlación con la Variable Objetivo

Con el fin de identificar asociaciones lineales entre la variable objetivo survived (supervivencia) y otras variables numéricas continuas o de alta cardinalidad, se empleó el coeficiente de correlación punto-biserial, el cual es adecuado para medir la relación entre una variable binaria y una continua. Este coeficiente toma valores entre -1 y 1, donde valores cercanos a los extremos indican una alta asociación (positiva o negativa), y valores cercanos a cero indican una débil o nula asociación.

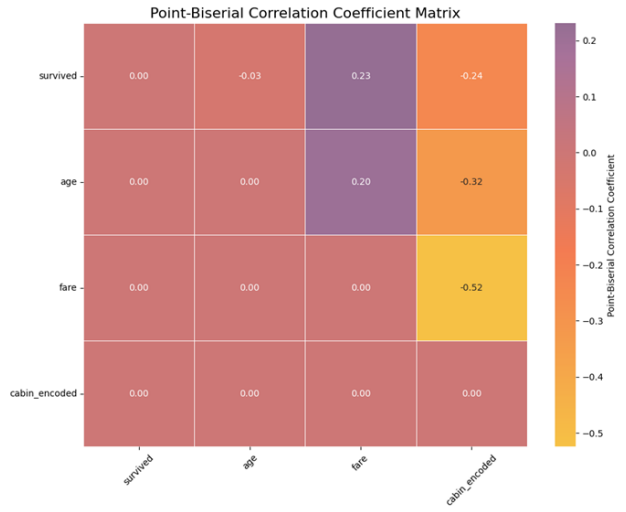
### *A. Manejo de Valores Faltantes*

Dado que el cálculo de correlaciones mediante scipy.stats.pointbiserialr no admite valores nulos, se generó un nuevo DataFrame con imputaciones específicas:

- age y embarked: Imputados con la moda, dado que no hay presencia de decimales en age.
- fare: Imputado con la media, ya que se trata de una variable monetaria continua.
- cabin_encoded: Se le asignó un nuevo valor llamado "Unknown", que puede representar cabinas compartidas o sin asignar, comunes entre los pasajeros de tercera clase.

### *B. Matriz de Correlación Punto-Biserial*

Junto a ello se calculó la correlación punto-biserial entre survived y las siguientes variables: age, fare, y cabin_encoded. A continuación, se presentan los coeficientes de correlación




*Matriz Número 1 – Matriz de Correlación Punto-Biserial*

Con ello se presentan las siguientes hipótesis:

- ***fare (Tarifa)***

**Correlación con survived:** +0.231

**Interpretación:** Existe una correlación positiva moderada entre el valor del pasaje y la probabilidad de sobrevivir. Esto sugiere que los pasajeros que pagaron más por su boleto (posiblemente de clases más altas) tuvieron más probabilidades de sobrevivir. Esto es coherente con los hechos históricos: pasajeros de primera clase tuvieron prioridad en los botes salvavidas.

- ***cabin_encoded (Cabina codificada)***

**Correlación con survived:** -0.241

**Interpretación:** Esta correlación negativa moderada indica que a medida que el número/codificación de cabina aumenta (o cuando no se tiene cabina asignada, como en clase baja), la probabilidad de sobrevivir disminuye. Es decir, los pasajeros con cabinas más simples o sin cabina tendieron a tener menores tasas de supervivencia.

- ***age (Edad)***

**Correlación con survived:** -0.033

**Interpretación:** Esta correlación es muy débil y negativa, lo que indica que la edad tuvo muy poco impacto en la supervivencia según estos datos. Aunque cabría esperar que los niños tuvieran mayor prioridad, la señal es tan pequeña que no se considera significativa.

- ***fare y cabin_encoded***

**Correlación:** -0.525

**Interpretación:** también tienen una correlación negativa fuerte entre sí, lo que puede indicar que cabinas más "codificadas" (o ausentes) estaban asociadas con boletos más baratos. Tiene sentido, ya que pasajeros de tercera clase generalmente no tenían cabina propia.

- ***age y fare***

**Correlación:** 0.204

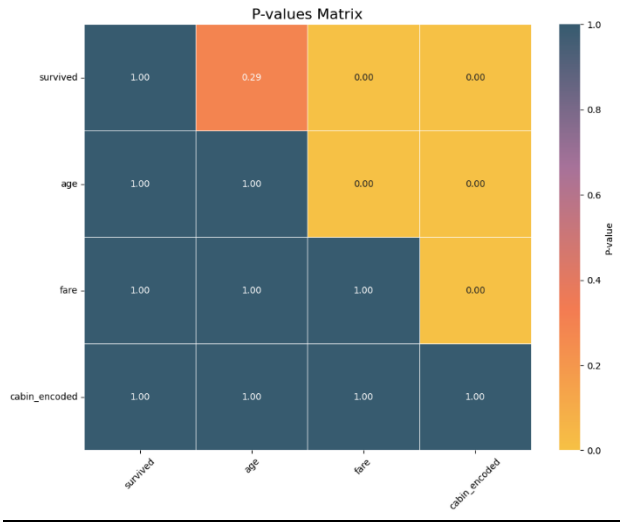
**Interpretación:** tienen una correlación positiva baja (0.204), lo que podría sugerir que personas mayores pagaban pasajes ligeramente más caros (posiblemente porque viajaban en mejores condiciones), aunque la relación es débil.

*C. Matriz de P - Valores*

Teniendo presente los p-valores, se toma a consideración que un p-valor indica la probabilidad de que una correlación observada ocurra por azar.

- Si $p < 0.05$, se considera estadísticamente significativa (es decir, confiamos en que la correlación es real).
- Si $p > 0.05$, la correlación se considera no significativa (podría haber ocurrido por casualidad).

Por lo tanto, se efectúan los siguientes análisis;




*Matriz Número 2 – Matriz de P-valores*

- ***fare vs survived***

**p = 1.62e-13**: Muy significativo

**Interpretación:** La correlación positiva entre el valor del pasaje y la probabilidad de sobrevivir es estadísticamente confiable.

- ***cabin_encoded vs survived***

**p = 1.04e-14**: Muy significativo

**Interpretación:** También podemos confiar en la correlación negativa entre el tipo/nivel de cabina y la supervivencia.

- **age vs survived**

**p = 0.292:** No significativa

**Interpretación:** No hay evidencia estadística suficiente para afirmar que la edad influye en la probabilidad de sobrevivir en este conjunto de datos. Puede haber sido un efecto aleatorio o demasiado débil.

A continuación, examinaremos cómo ciertas características individuales de los pasajeros del Titanic se relacionaron con su probabilidad de sobrevivir al desastre. Se detallan los resultados de la correlación para la tarifa pagada (Fare), el número de cabina (Cabin_encoded) y la edad (Age). Para cada variable, se presenta el coeficiente de correlación, el p-valor asociado y una interpretación de los hallazgos, junto con posibles explicaciones contextuales basadas en la estructura social y física del barco.

**1. Fare (Tarifa de embarque)**

- Existe una correlación positiva moderada entre la tarifa pagada y la probabilidad de supervivencia.
- Esto es estadísticamente significativo, lo que indica que este patrón no es producto del azar.

- **Posible explicación:** Los pasajeros que pagaron tarifas más altas solían viajar en primera clase, donde estaban ubicados en zonas más cercanas a los botes salvavidas y recibieron atención prioritaria durante la evacuación. Además, el personal del Titanic estaba condicionado a proteger primero a los pasajeros de clases altas, reflejando una estructura social jerárquica típica de la época.

**2. Cabin_encoded (Número de cabina codificado)**

- Existe una correlación negativa moderada entre el número de cabina y la supervivencia.
- Esto sugiere que las personas con valores altos en cabin_encoded, lo cual puede corresponder a cabinas con numeración elevada, compartidas o incluso desconocidas, tuvieron menos probabilidades de sobrevivir.
- Posible explicación: Muchos pasajeros de tercera clase no tenían una cabina individual, y si la tenían, no era identificada en el manifiesto. Estos pasajeros estaban ubicados en los niveles inferiores del barco, lejos de las salidas y botes. Esto refleja un factor estructural de exclusión y acceso desigual durante la emergencia.

**3. Age (Edad)**

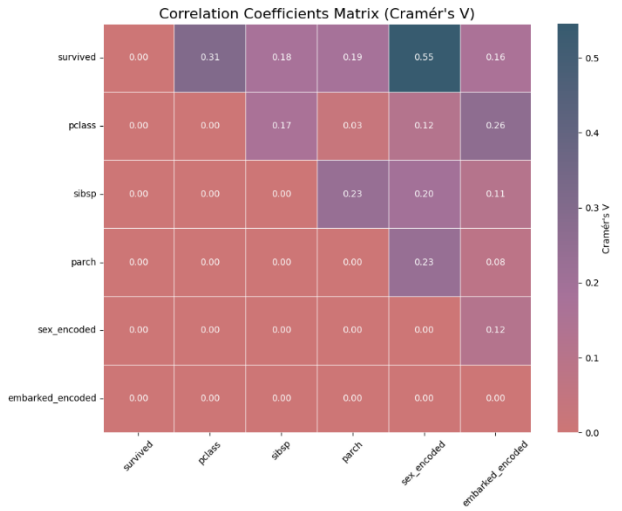
- La edad tiene una correlación muy débil y no significativa con la supervivencia. Esto significa que, en promedio, la edad no tuvo un efecto claro o sistemático sobre quién sobrevivió o no.
- Posible explicación: Aunque había una consigna de "mujeres y niños primero", en la práctica no siempre se cumplió de manera sistemática. Además, la variable de edad por sí sola no distingue entre niños pequeños, adolescentes o adultos jóvenes, lo cual puede estar diluyendo el efecto.

## VI. ANÁLISIS VARIABLES CATEGÓRICAS

Con el fin de entender la fuerza de la relación entre las variables categóricas del conjunto de datos del Titanic y la variable objetivo (survived), se calculó el coeficiente de asociación de Cramér's V. Este coeficiente mide la fuerza de asociación entre dos variables categóricas, tomando valores entre 0 (sin asociación) y 1 (asociación perfecta). Se complementó el análisis con los p-valores correspondientes para evaluar la significancia estadística de cada relación.

*A. Matriz de Coeficientes Cramér's V*

Para entender la relación entre variables categóricas y la probabilidad de supervivencia en el Titanic, se utilizó el coeficiente de Cramér's V, que cuantifica la fuerza de asociación entre pares de variables categóricas. Este análisis permite identificar qué características tienen mayor capacidad explicativa sobre el desenlace de los pasajeros (sobrevivió o no).




*Matriz Número 3 – Coeficientes Cramér's V*

**Relación con la Variable Objetivo (survived):**

- **sex_encoded (sexo del pasajero):**

**Cramér's V = 0.545**

Se trata de la asociación más fuerte observada. Esto refleja una diferencia clara y marcada entre hombres y mujeres en cuanto a la probabilidad de supervivencia. El valor alto sugiere que el sexo fue un factor determinante durante la evacuación, reforzando el conocido protocolo histórico de "mujeres y niños primero".

- **pclass (clase del boleto):**

**Cramér's V = 0.306**

Muestra una asociación moderada con la supervivencia. Los pasajeros de primera clase tenían mayor acceso a los botes salvavidas y mejor ubicación dentro del barco, lo cual influyó en su probabilidad de sobrevivir. Este hallazgo coincide con patrones de desigualdad social durante el naufragio.

- **parch (número de padres o hijos a bordo):**

**Cramér's V = 0.187**

Aunque débil, indica que la presencia de familiares cercanos podría haber afectado el comportamiento en situaciones de emergencia, por ejemplo, el deseo de permanecer juntos o el apoyo mutuo.

- **sibsp (número de hermanos o cónyuges a bordo):**

**Cramér's V = 0.177**

Al igual que parch, sugiere una leve relación con la supervivencia. Podría influir tanto de forma positiva (cooperación familiar) como negativa (dificultad para evacuar en grupo).

- **embarked_encoded (puerto de embarque):**

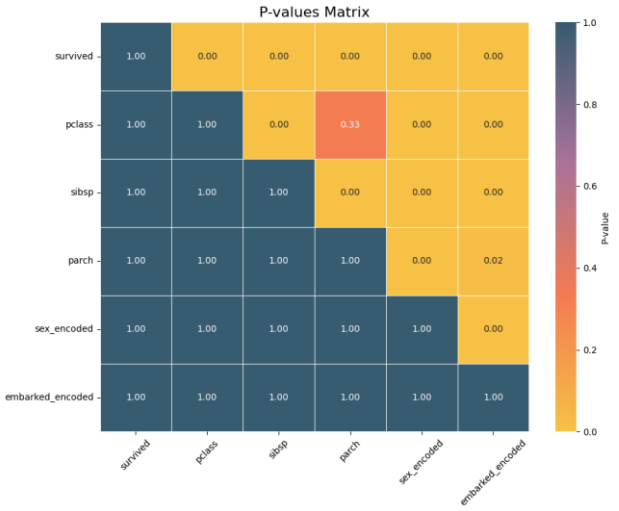
**Cramér's V = 0.164**

Asociación débil. No parece tener un peso determinante, aunque puede reflejar diferencias indirectas en el perfil socioeconómico de los pasajeros según el lugar de embarque.

Este análisis refuerza la importancia de **sex** y **pclass** como variables clave para predecir la supervivencia en modelos de clasificación. Su inclusión es esencial tanto para el rendimiento predictivo como para la interpretabilidad histórica del fenómeno.

*B. Matriz de P-valores para Cramér's V*

Tras calcular los coeficientes de Cramér's V que estiman la fuerza de asociación entre variables categóricas, es necesario verificar si esas asociaciones son estadísticamente significativas. Para ello, se utilizaron p-valores, que indican la probabilidad de observar la relación detectada (o una más extrema) si no existiera realmente una asociación en la población.




*Matriz Número 4 – P-valores para Cramér's V*

- **sex_encoded: p ≈ 9.26e-67:**

La asociación entre sexo y supervivencia es extremadamente significativa. Este valor refuerza cuantitativamente lo que la historia y el coeficiente de Cramér's V ya sugerían: el sexo tuvo un impacto determinante en las probabilidades de sobrevivir al naufragio.

- **Pclass: p ≈ 1.68e-21:**

También altamente significativo. Indica que la clase del pasajero influyó fuertemente en el desenlace. El acceso a botes y la ubicación en el barco, asociadas a la clase, probablemente explican esta influencia.

- **parch, sibsp, embarked_encoded: p < 0.00001:**

Aunque su fuerza de asociación era baja (ver análisis de Cramér's V), los p-valores muestran que estas relaciones no son aleatorias. Es decir, tener familiares a bordo o el lugar de embarque puede tener una influencia sutil pero real en la supervivencia.

- **pclass y parch: p ≈ 0.33:**

Esta es la única relación no significativa en toda la matriz. Indica que la clase no está estadísticamente asociada con la cantidad de padres/hijos a bordo. Lo cual tiene sentido: pasajeros de cualquier clase podían viajar con o sin familia.

- **Generalidades**

el resto de combinaciones entre variables también muestran p-valores extremadamente bajos, lo que respalda la idea de que las variables categóricas utilizadas contienen señales relevantes y no aleatorias sobre la estructura de los datos.

## VII. MODELAMIENTO

*A. Evaluación de Modelos Predictivos*

Con el objetivo de evaluar el rendimiento de distintos enfoques de clasificación para predecir la supervivencia de pasajeros del Titanic, se implementaron y compararon seis modelos supervisados: Regresión Logística, Random Forest, XGBoost, Support Vector Machines (SVM), K-Nearest Neighbors (KNN) y Naive Bayes. Además de considerar métricas tradicionales como la exactitud, el análisis se centró especialmente en el comportamiento de cada modelo frente a la clase minoritaria (supervivientes), utilizando medidas como recall, precisión y F1-score.

| *Modelo* | *Exactitud* | *Recall (Clase 1)* | *Precision (Clase 1)* | *F1-score (Clase 1)* |
|---|---|---|---|---|
| *Regresión Logística* | 0.795 | 0.704 | 0.714 | 0.709 |
| *Random Forest* | 0.790 | 0.718 | 0.699 | 0.708 |
| *XGBoost* | 0.805 | **0.761** | 0.711 | **0.735** |
| *SVM* | **0.810** | 0.676 | **0.762** | 0.716 |
| *KNN* | 0.785 | 0.676 | 0.706 | 0.691 |
| *Gaussian Naive Bayes* | 0.780 | 0.747 | 0.671 | 0.707 |

*Tabla Número 4 – **Modelos Predictivos***

Este enfoque es crucial en problemas con clases desbalanceadas, donde el costo de los errores de clasificación varía según la clase. A continuación, se presentan observaciones detalladas sobre el desempeño de cada modelo, con énfasis en su capacidad para identificar correctamente a los sobrevivientes (clase 1), así como sus posibles ventajas y limitaciones contextuales.

- **Regresión Logística**

Este modelo se presenta como la base interpretable del análisis. Su desempeño general es sólido, con una precisión aceptable y coeficientes fácilmente explicables. Sin embargo, su capacidad para detectar correctamente a los sobrevivientes es limitada: aproximadamente un 30% de ellos no son identificados (recall bajo), lo cual podría ser problemático en escenarios donde los falsos negativos tienen un alto costo.

- **Random Forest**

Ofrece un rendimiento similar al de la regresión logística, pero muestra mayor robustez frente a relaciones no lineales entre las variables. Presenta una ligera mejora en el recall de la clase 1, lo que indica un mejor reconocimiento de los sobrevivientes, aunque aún de forma moderada.

- **XGBoost**

Destaca como el modelo con el mejor balance entre precisión y recall. Su arquitectura permite capturar patrones complejos en los datos, lo que se traduce en una identificación más eficiente de los sobrevivientes sin sacrificar demasiado la precisión. Este equilibrio lo posiciona como una opción particularmente adecuada en contextos donde minimizar falsos negativos es prioritario.

- **Support Vector Machine (SVM)**

El modelo SVM demuestra una mayor precisión en la predicción de sobrevivientes, lo que sugiere una baja tasa de falsos positivos. No obstante, su recall es el más bajo entre los modelos evaluados, lo que indica una limitada capacidad para detectar a todos los individuos de la clase positiva. Esta característica lo hace más apropiado en escenarios donde los falsos positivos son más costosos que los falsos negativos.

- **K-Nearest Neighbors (KNN)**

Aunque simple y altamente dependiente de la estructura local de los datos, el modelo KNN ofrece un rendimiento razonable. A pesar de su sensibilidad al ruido y su menor capacidad general de generalización, logra un desempeño comparable al de otros modelos más complejos, aunque no alcanza los niveles de precisión o recall observados en XGBoost o SVM.

- **Naive Bayes**

Pese a su simplicidad y a la fuerte suposición de independencia condicional entre características, Naive Bayes se muestra competitivamente efectivo. El modelo logra identificar una proporción significativa de sobrevivientes (buen recall), aunque lo hace a costa de una menor precisión, lo que implica una mayor cantidad de falsos positivos. Esta compensación puede ser aceptable dependiendo del contexto de aplicación.

*B. Optimización De Hiperparámetros*

Se seleccionaron los modelos Regresión Logística, XGBoost y Support Vector Machine (SVM) para la etapa de optimización de hiperparámetros debido a su relevancia teórica, capacidad de generalización y su desempeño observado en las pruebas iniciales:

- **Regresión Logística**

Aunque es un modelo simple, la Regresión Logística mostró un rendimiento competitivo (F1-score = 0.709) comparable al de modelos más complejos. Se destaca por su interpretabilidad y bajo número de hiperparámetros, lo que permite evaluar de forma clara el efecto de su optimización.

- **XGBoost**

Este modelo presentó uno de los mejores desempeños globales, con una alta exactitud (0.805) y el mayor recall (0.761), lo que indica su capacidad para identificar correctamente a los sobrevivientes (clase 1). Esto lo convierte en un candidato ideal para optimización, ya que pequeños ajustes en sus múltiples hiperparámetros (como número de árboles, profundidad, tasa de aprendizaje) pueden maximizar aún más su rendimiento.

- **Support Vector Machine (SVM)**

SVM obtuvo la mayor precisión (0.762) para la clase 1, lo que indica que cuando predice un sobreviviente, generalmente acierta. Aunque su recall fue menor (0.676), esta característica lo hace valioso en contextos donde los falsos positivos son más costosos que los falsos negativos. Dado que SVM es muy sensible a la configuración de hiperparámetros como el kernel, C y gamma, su inclusión permite explorar su potencial de mejora mediante una optimización adecuada.

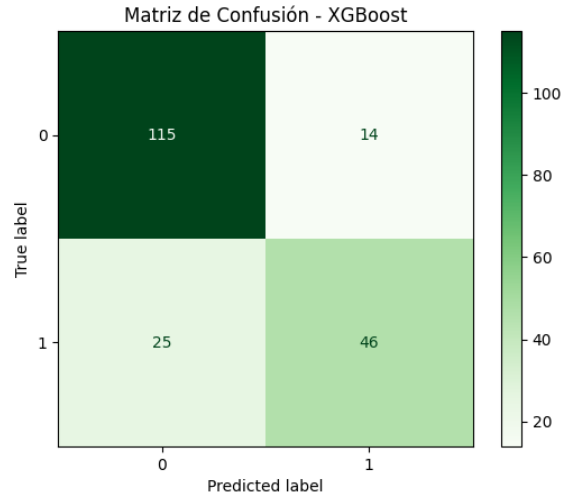
Tras aplicar **GridSearchCV** para afinar los hiperparámetros de tres modelos seleccionados: **Regresión Logística**, **XGBoost** y **SVM**, se observó que las mejoras en el rendimiento fueron marginales o inexistentes.

| | *Modelo* | *Exactitud (Test)* | *Recall (Clase 1)* | *Precision (Clase 1)* | *F1-score (Clase 1)* | *Recall (Clase 0)* | *Precision (Clase 0)* | *F1-score (Clase 0)* |
|---|---|---|---|---|---|---|---|---|
| *0* | Regresión Logística (GridSearchCV) | 0.795 | 0.6901 | 0.7206 | 0.705 | 0.8527 | 0.8333 | 0.8429 |
| *1* | XGBoost (GridSearchCV) | 0.805 | 0.6479 | 0.7667 | 0.7023 | 0.8915 | 0.8214 | 0.855 |
| *2* | SVM (GridSearchCV) | 0.81 | 0.662 | 0.7705 | 0.7121 | 0.8915 | 0.8273 | 0.8582 |

*Tabla Número 5 –* ***Optimización***

- **Regresión Logística** mostró un rendimiento casi idéntico tras la optimización, lo cual sugiere que ya operaba cerca de su punto óptimo o que su capacidad de mejora mediante hiperparámetros es limitada dada la estructura lineal del modelo.

- En el caso de **XGBoost**, si bien mejoró la precisión en la clase positiva (sobrevivientes), se produjo una disminución considerable en el recall, lo que implica que se identificaron menos sobrevivientes correctamente. Esta pérdida de sensibilidad reduce su utilidad en escenarios donde los falsos negativos son costosos, como es el caso del presente problema.
- **SVM**, por su parte, mantuvo prácticamente el mismo nivel de desempeño que antes de la optimización, con ligeras variaciones en precisión y recall, pero sin impacto significativo en las métricas globales.

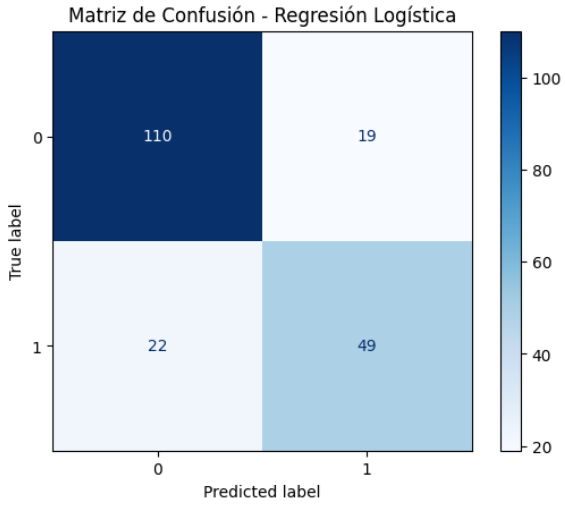



*Matriz Número 6 – Matriz Confusion – XGBoost*

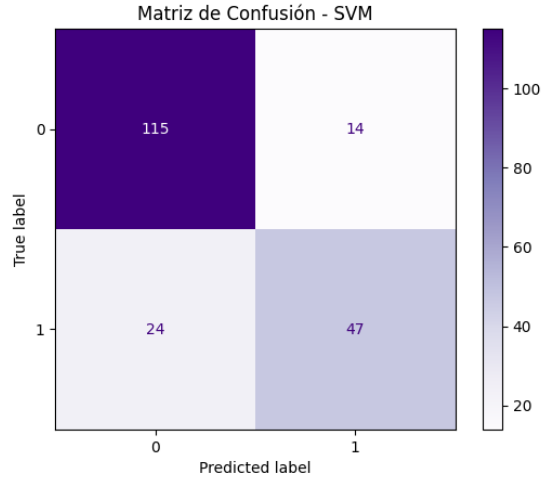
## VIII. VISUALIZACIÓN E INTERPRETABILIDAD

### *A. Matriz de Confusión*

Las matrices de confusión permiten observar con claridad los errores tipo I y tipo II (falsos positivos y falsos negativos), lo cual es crucial en contextos como la predicción de supervivencia, donde no detectar a un sobreviviente puede ser más grave que predecir uno incorrectamente.




*Matriz Número5 – Matriz Confusion – Regresión Logística*




*Matriz Número 7 – Matriz Confusion – SVM*

**Regresión Logística** logra identificar un mayor número de sobrevivientes (49), lo que es relevante dado que, en el contexto del Titanic, los sobrevivientes fueron una minoría y su identificación es clave para comprender patrones de supervivencia (por ejemplo, sexo, clase o edad).

**XGBoost y SVM** identifican menos sobrevivientes correctamente, pero son más conservadores al etiquetar a alguien como tal (menor tasa de falsos positivos).

*B. Curvas ROC y AUC*

Las curvas ROC y el AUC indican cuán bien distingue un modelo entre sobrevivientes y no sobrevivientes, sin necesidad de fijar un umbral específico.

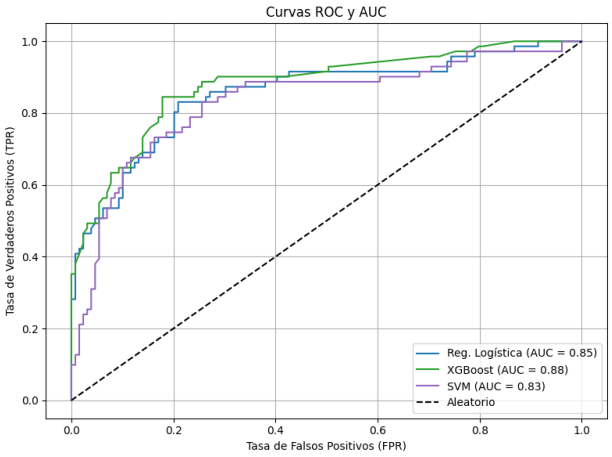



*Gráfico Número 1 – Curvas ROC y AUC*

**XGBoost alcanza el mejor AUC**, lo que sugiere que tiene una alta capacidad para detectar correctamente tanto a sobrevivientes como a no sobrevivientes, incluso con distintos umbrales. Esto es importante si se desea ajustar el modelo para distintos contextos (por ejemplo, simulaciones históricas o decisiones hipotéticas sobre prioridades de evacuación).

**Regresión logística**, aunque más simple, también ofrece una alta discriminación, reforzando su utilidad como modelo interpretable para entender qué factores históricos influyeron en la supervivencia.

**SVM**, aunque tiene un AUC menor, aún se desempeña razonablemente bien, y sugiere que el margen máximo que traza entre clases sigue siendo útil, pero tal vez menos ajustado al patrón social y demográfico de los pasajeros del Titanic.

*C. Importancia de Características con SHAP*

El análisis SHAP revela qué variables fueron más determinantes en las decisiones de los modelos. En el Titanic, esto ayuda a confirmar o refutar hipótesis históricas sobre quiénes tenían mayor probabilidad de sobrevivir.

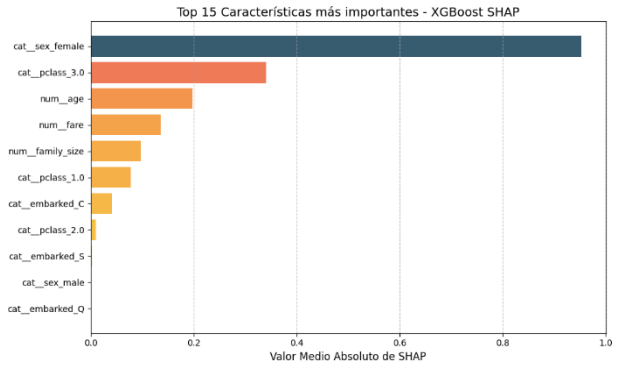



*Gráfico Número 2 – SHAP*

**Característica:** sex_female
**SHAP medio:** 0.953
**Interpretación:** Ser mujer fue el factor más decisivo. Esto confirma la política de "mujeres y niños primero", que predominó durante la evacuación.

**Característica:** pclass_3.0
**SHAP medio:** 0.341
**Interpretación:** Viajar en tercera clase se asoció con una menor probabilidad de sobrevivir, posiblemente por ubicación en el barco y acceso limitado a los botes salvavidas.

**Característica:** age
**SHAP medio:** 0.198
**Interpretación:** La edad también influye, ya que los niños tuvieron más prioridad.
**Característica:** fare
**SHAP medio:** 0.136
**Interpretación:** Tarifas más altas (indicativas de primera clase) correlacionan con una mayor supervivencia.

**Característica:** family_size
**SHAP medio:** 0.097
**Interpretación:** Pasajeros con familias grandes pueden haber enfrentado más dificultades para evacuar juntos.

**Característica:** pclass_1.0
**SHAP medio:** 0.078
**Interpretación:** Refuerza que la primera clase tuvo más oportunidades de sobrevivir.

El análisis visual y la interpretabilidad permiten verificar que los modelos no solo predicen con buena precisión, sino que también capturan de forma coherente los factores históricos que determinaron la supervivencia en el Titanic. Este enfoque añade un valor narrativo y explicativo a los modelos, facilitando su uso como herramienta de análisis histórico y sociológico.

## IX. ANÁLISIS DE RESULTADOS Y CONCLUSIÓN

**1. Factores clave de supervivencia:** A través de los modelos predictivos e interpretabilidad con SHAP, se confirma que el sexo (ser mujer), la clase de pasajero (primera clase) y, en menor medida, la edad y la tarifa pagada fueron los factores más influyentes en la probabilidad de supervivencia. Esto refleja las decisiones sociales y estructurales de la época, en particular la priorización de mujeres y pasajeros de clase alta durante la evacuación.

**2. Evidencia de desigualdad estructural:** Los resultados refuerzan un hallazgo crítico: los pasajeros de tercera clase y los hombres tuvieron una probabilidad significativamente menor de sobrevivir, lo que sugiere un acceso desigual a los botes salvavidas. Los modelos captan estas disparidades históricas de forma clara y consistente, resaltando cómo las decisiones tomadas durante el desastre se vieron influenciadas por el estatus socioeconómico y el género.

**3. XGBoost y SVM: los modelos más balanceados:** Tras optimización, los modelos XGBoost y SVM ofrecieron un equilibrio superior entre exactitud, precisión y sensibilidad, logrando predecir correctamente una mayor proporción de sobrevivientes sin comprometer significativamente la tasa de falsos positivos. Esto es especialmente relevante en contextos donde identificar correctamente a quienes necesitan asistencia es prioritario, como en simulaciones de evacuación.

**4. La Regresión Logística: simple, interpretable y competitiva:** A pesar de su simplicidad, la Regresión Logística mostró un rendimiento competitivo y una interpretabilidad valiosa. Aunque su capacidad para detectar sobrevivientes fue ligeramente inferior, es un excelente punto de partida por su claridad en la relación entre variables y resultado. Es útil para construir una narrativa clara del fenómeno.

**5. Importancia de la interpretabilidad en eventos históricos:** Las técnicas de interpretabilidad (SHAP, matrices de confusión, curvas ROC) no solo ayudaron a entender cómo funcionan los modelos, sino que también permitieron conectar patrones numéricos con decisiones humanas históricas reales. Esta trazabilidad entre datos, modelo y contexto refuerza el valor educativo y analítico del enfoque.

## X. REFERENCIAS


**[1]** J. Sherlock, M. Muniswamaiah, L. Clarke, y S. Cicoria, "Classification of Titanic Passenger Data and Chances of Surviving the Disaster," arXiv preprint arXiv:1810.09851, 2018. [En línea]. Disponible en: https://arxiv.org/abs/1810.09851

**[2]** W. Liang, "Titanic Disaster Prediction Based on Machine Learning Algorithms," BC Publication, 2023. [En línea]. Disponible en: https://bcpublication.org/index.php/BM/article/view/4860

**[3]** A. E. Amalia y C. Rahayu, "Comparison of Machine Learning Classification Models in Predicting The Titanic Survival Rate," International Journal of Computer Science and Human-Aided Intelligence, vol. 3, no. 2, pp. 52–58, 2023. [En línea]. Disponible en: https://journal.binus.ac.id/index.php/ijcshai/article/view/12163

**[4]** Y. Liao, S. Zhang, y Z. Zhang, "Research on Titanic Survival Prediction Based on Machine Learning Method," Proceedings of the Asia Education Management and Public Service Conference, 2023. [En línea]. Disponible en: https://www.ewadirect.com/proceedings/aemps/article/view/19451

**[5]** Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

**[6]** Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies*, 2(1), 37–63.

**[7]** Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. In *Advances in Neural Information Processing Systems* (NeurIPS 2017).

**[8]** Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD* (pp. 785–794).