Análisis Predictivo de Supervivencia en el Titanic Mediante Clasificadores de Machine Learning

Esteban Sierra Baccio
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Monterrey, México
a00836286@tec.mx

Sergio Aarón Hernández Orta
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Monterrey, México
a01613878@tec.mx

Javier Jorge Hernández Verduzco
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Monterrey, México
a01722667@tec.mx

Sergio Omar Flores García
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Monterrey, México
a01285193@tec.mx

Diego Esparza Ruíz
Escuela de Ingeniería y Ciencias
Tecnológico de Monterrey
Monterrey, México
a00837527@tec.mx

***Abstract*—Abstract—Contexto: En el ámbito de la ciencia de datos, la predicción de resultados binarios como la supervivencia es un problema clásico que combina relevancia histórica con desafíos modernos de machine learning. El naufragio del RMS Titanic, con su detallado registro de pasajeros, ofrece un caso de estudio ideal para explorar la aplicación de modelos predictivos y sus implicaciones éticas.**

Objetivo: El propósito de este trabajo es desarrollar un modelo de machine learning para predecir la probabilidad de supervivencia de cada pasajero del Titanic, abordando simultáneamente la necesidad de interpretabilidad y equidad algorítmica.

Métodos: Se utilizó un conjunto de datos público que incluye información demográfica y de viaje de los pasajeros. Se aplicaron diversas técnicas de preprocesamiento, como la imputación de datos y la ingeniería de características. Se entrenaron y evaluaron múltiples modelos de clasificación, incluyendo árboles de decisión, regresión logística y XGBoost, comparando su rendimiento.

Resultados: El modelo XGBoost alcanzó la mejor métrica de rendimiento, superando a otros clasificadores en exactitud. Sin embargo, un análisis de interpretabilidad reveló que el "sexo" y la "clase de pasajero" fueron las características más influyentes en la predicción, lo que plantea preocupaciones sobre la justicia y los sesgos inherentes a los datos.

Conclusiones: Si bien es posible construir un modelo predictivo altamente preciso, el estudio subraya la importancia crítica de la interpretabilidad y la ética en el machine learning. Los hallazgos demuestran cómo la selección de características puede reflejar y amplificar sesgos históricos, destacando la necesidad de auditorías algorítmicas para asegurar la equidad en la toma de decisiones.

Index Terms—machine learning, fairness, interpretability, Titanic, classification, ethical AI

***Index Terms*—machine learning, fairness, interpretability, Titanic, classification, ethical AI**

I. Introducción

A. Apertura Impactante

En la madrugada del 15 de abril de 1912, el RMS Titanic se hundía en el Atlántico Norte, llevando consigo más de 1,500 vidas y creando uno de los datasets más estudiados en ciencia de datos. Más allá de la tragedia, este evento capturó un microcosmos de la sociedad eduardiana, donde las jerarquías sociales determinaron literalmente quién vivió y quién murió.

B. Contexto del Problema

El dataset del Titanic trasciende el análisis histórico para convertirse en un laboratorio natural para estudiar sesgos algorítmicos. Las estructuras sociales de 1912 —género, clase social, nacionalidad— se manifiestan directamente en patrones de supervivencia, ofreciendo insights críticos sobre cómo los sesgos históricos pueden perpetuarse en sistemas de inteligencia artificial modernos.

C. Motivación

En una era donde algoritmos de machine learning toman decisiones que afectan vidas humanas —desde sistemas de justicia penal hasta aprobación de créditos— entender cómo los sesgos sociales se codifican en datos es fundamental. El Titanic proporciona un caso de estudio donde las consecuencias de estos sesgos fueron literalmente de vida o muerte.

D. Preguntas de Investigación

- RQ1: ¿Qué factores fueron más determinantes para la supervivencia en el Titanic?
- RQ2: ¿Cómo se manifiestan los sesgos sociales eduardianos en patrones algorítmicos?
- RQ3: ¿Qué trade-offs existen entre accuracy predictivo y fairness algorítmico?
- RQ4: ¿Qué lecciones podemos extraer para el diseño ético de sistemas de IA modernos?

E. Contribuciones

1) Análisis sistemático de sesgos sociales en datos históricos del Titanic
2) Comparación de múltiples algoritmos de ML con enfoque en interpretabilidad

Si aplica, aquí la nota de financiamiento.

3) Framework metodológico para evaluación de fairness en contextos históricos
4) Insights sobre optimización algorítmica versus equidad social

F. Estructura del Paper

Este trabajo progresa desde contextualización histórica y análisis exploratorio, hacia modelado predictivo y evaluación de fairness, concluyendo con implicaciones para IA ética moderna.

II. Revisión de Literatura

A. Trabajos Previos sobre el Dataset Titanic

Se realizó una investigación sobre los acontecimientos del desastre del titanic, obteniendo información importante para la formulación de hipótesis, entre ellas el mapa del recorrido del barco, la visualización de dónde se encontraban las cabinas con respecto al barco.

B. Machine Learning Interpretable

El campo del Machine Learning Interpretable (MLI) ha cobrado gran relevancia ante la necesidad de comprender cómo los modelos de caja negra, como las redes neuronales y los modelos de boosting, toman sus decisiones. Entre una de las técnicas de MLI más importantes se encuentran los Shapley Additive Explanations (SHAP) ayudan a entender el comportamiento del modelo en general, mostrando cómo una característica impacta la predicción. Este método es crucial para auditar y confiar en los modelos de IA, especialmente en aplicaciones de alto riesgo.

C. Fairness en Machine Learning

El concepto de equidad o fairness en el aprendizaje automático aborda el riesgo de que los modelos perpetúen o amplifiquen sesgos sociales. Los sesgos pueden manifestarse de varias maneras, por ejemplo, cuando un modelo discrimina a ciertos grupos demográficos. Las métricas de equidad, como la paridad demográfica o la igualdad de oportunidades, se utilizan para evaluar si las predicciones del modelo son justas en diferentes subgrupos (por ejemplo, hombres vs. mujeres en el caso del Titanic). Trabajos como los de Zafar et al. (2019) han propuesto métodos para mitigar el sesgo, ya sea en la fase de pre-procesamiento, durante el entrenamiento del modelo, o en la post-procesamiento. Es crucial considerar la equidad al desarrollar modelos de IA, ya que ignorarla puede llevar a consecuencias éticas y sociales graves.

D. Ética en IA y Decisiones Algorítmicas

La ética en la inteligencia artificial es un campo interdisciplinario que examina las implicaciones morales de la IA. La toma de decisiones algorítmicas, en particular, plantea preguntas sobre la responsabilidad, la transparencia y la rendición de cuentas. Es importante mencionar que a pesar de que este modelo busca la predicción de supervivencia de los pasajeros, no sea un modelo con la palabra final al momento de seleccionar a los posibles supervivientes en caso de un desastre parecido. Este modelo únicamente busca explicar lo sucedido.

E. Gap en la Literatura

A pesar de la extensa investigación sobre el Dataset Titanic y los avances en MLI y fairness, existe una brecha notable en la literatura . Si bien muchos estudios se centran en la precisión predictiva, pocos integran un análisis exhaustivo de interpretabilidad y equidad en el mismo marco. La mayoría de los trabajos se limitan a la aplicación de modelos de caja negra sin explicar por qué ciertas características son más importantes o cómo las decisiones algorítmicas podrían ser sesgadas. Nuestro trabajo, por lo tanto, busca llenar este vacío al desarrollar un modelo de predicción de supervivencia en el Titanic que no solo sea preciso, sino también transparente y justo, utilizando técnicas de Machine Learning Interpretable y métricas de equidad para auditar y justificar sus decisiones.

III. Metodología

A. Dataset y Preprocesamiento

1) Descripción del Dataset:
- Tabla con estadísticas descriptivas
- Distribución de clase objetivo

2) Análisis de Calidad de Datos:
- Patterns de valores faltantes
- Figura: Heatmap de missingness

3) Ingeniería de Features:
- Tabla: Features creadas con justificación
- Proceso de selección de features

4) Estrategia de Imputación:
- Comparación de métodos
- Validación de imputaciones

B. Diseño Experimental

1) Formulación del Problema: La tarea se define como un problema de **clasificación binaria**. Nuestro objetivo es determinar cuales eran los factores que afectaban la probabilidad de sobrevivir al accidente del Titanic. La variable objetivo es `Survived`, codificada con `0` para un pasajero fallecido y `1` para uno sobreviviente.

El conjunto de variables `X` queda compuesto por 32 variables después del preprocesamiento, justo antes del entrenamiento. Estas provienen de variables demográficas y socioeconómicas, por ejemplo:

- Sex_female
- Pclass_3
- Title
- AgeGroup
- CabinDeck
- FamilySize, IsAlone
- FarePerPerson

Las categóricas se expandieron vía **One Hot Encoding**, produciendo indicadores como `Title_Mr`, `Sex_female`, etc. Se asumió independencia entre observaciones al eliminar identificadores únicos (nombre, ticket, cabina).

- **Métrica primaria**: ROC-AUC [6].
- **Métricas secundarias**: PR-AUC [15], MCC [3],

Balanced Accuracy, F1, Precisión, Recall, además de Brier Score [2] y Expected Calibration Error.

2) Estrategia de Validación:

- División de datos: holdout estratificado con 80% para entrenamiento y 20% para prueba.
- Preprocesamiento aplicado mediante la clase `TitanicDatasetPreprocessor`, que realiza:
  - Imputación en numéricas con la mediana.
  - Imputación en categóricas con la moda.
  - Escalado de numéricas con `StandardScaler`.
  - One Hot Encoding.
- Selección de hiperparámetros con **GridSearchCV**, usando ROC-AUC para medir desempeño. Se hizo el ajuste en 5 folds, con todas las cores disponibles (más sobre eso en *consideraciones computacionales*).

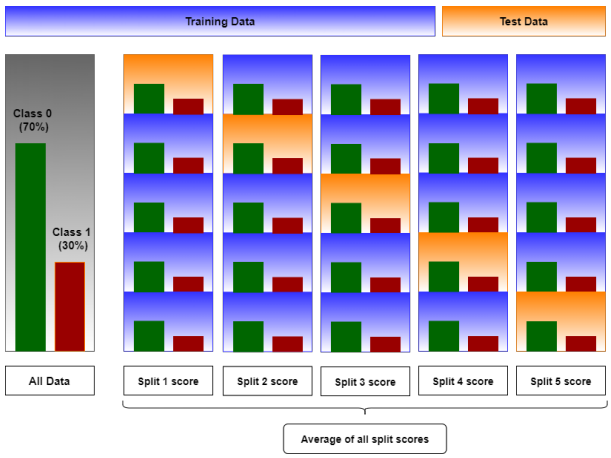

Fig. 1. Método de Validación Cruzada Estratificada de 5 folds.

3) Algoritmos Implementados: Se entrenaron los siguientes modelos:

1) **Regresión Logística** [12] [5].
2) **Random Forest** [1].
3) **XGBoost** [**chen2016xgboost**].
4) **SVM con kernels** [4].
5) **Gradient Boosting** [7]

Para cada modelo había un espacio de hiperparámetros sobre el que se tenía que optimizar para encontrar la mejor versión del modelo.

Espacios de búsqueda de hiperparámetros:

- Logistic Regression (l1/l2):
  - C en {0.001, 0.01, 0.1, 1, 10, 100}
  - penalty en {l1, l2}
  - class_weight en {None, balanced}
- Logistic Regression (elasticnet):
  - C en {0.001, 0.01, 0.1, 1, 10, 100}
  - l1_ratio en {0, 0.5, 1}
- Random Forest:
  - n_estimators en {100, 200, 300, 500}
  - max_depth en {3, 5, 7, 10, None}
  - min_samples_split en {2, 5, 10}
  - min_samples_leaf en {1, 2, 4}
  - max_features en {sqrt, log2, 0.3, 0.5}
- XGBoost:
  - learning_rate en {0.01, 0.05, 0.1, 0.3}
  - n_estimators en {100, 200, 300}
  - max_depth en {3, 5, 7}
  - subsample en {0.7, 0.8, 0.9, 1.0}
  - colsample_bytree en {0.7, 0.8, 0.9, 1.0}
  - gamma en {0, 0.1, 0.3, 0.5}
- SVM:
  - C en {0.1, 1, 10, 100}
  - kernel en {rbf, poly, sigmoid}
  - gamma en {scale, auto, 0.001, 0.01, 0.1}
  - degree en {2, 3, 4}

La métrica de selección fue el mejor promedio de ROC-AUC durante la *validación cruzada*. El mejor modelo, optimizado para sus mejores hiperparámetros, fue el *baseline* de Logistic Regression, con `C=1, penalty=L1, class_weight=None`. El resto de los hiperparámetros para los demás modelos pueden encontrarse en el anexo **X**.

Para analizar la calibración de los modelos en los diferentes niveles de confianza, se reportaron los Brier Score y Expected Calibration Error para cada modelo, y estos se pueden observar en las notebooks.

4) Análisis de Fairness:

- Grupos de comparación:
  - Género (male, female).
  - Clase socioeconómica (Pclass: 1, 2, 3).
  - Grupo de edad (child, teen, adult, senior).
- Métricas evaluadas:
  - Diferencia en TPR (Recall).

$$Recall = \frac{TP}{TP + FN}$$

  - Diferencia en FPR.

$$FPR = \frac{FP}{FP + TN}$$

  - Diferencia en Precisión.

$$Precisión = \frac{TP}{TP + FP}$$

  - Diferencia de datos faltantes por demográficos.

C. Herramientas y Reproducibilidad

1) Stack tecnológico utilizado: El desarrollo e implementación de nuestra metodología fue sobre el ecosistema de herramientas para estadística y *machine learning* de Python (3.x) [17], trabajando sobre *notebooks* de Jupyter [11] para crear reportes de código legibles, documentables, e interactivos.

El análisis inicial se realizó en Google Colab [8], un servicio gratuito en la nube que permite colaborar en tiempo real

sobre *notebooks* de Jupyter en diversos ambientes de cómputo. Posteriormente, se exportó y desplazó la metodología a un repositorio de Git, interactuando con las *notebooks* usando Jupyter Lab. El paso al repositorio de Git nos permitió tomar control mas riguroso sobre versiones de los múltiples *notebooks* y datos.

Las librerías críticas a nuestro análisis fueron: **NumPy** para operaciones numéricas [9], **Pandas** para manipulación de tablas de datos [16, 13], y **scikit-learn** para el entrenamiento, validación y evaluación de los modelos [14]. Este conjunto de herramientas representa un estándar en el campo de *machine learning* en Python, lo que facilita la reproducción del experimento con mínimos requisitos adicionales. Cualquier versión común usada en 2025 de estas librerías es compatible con las funciones y estructuras que se utilizan.

2) Disponibilidad de código: El repositorio público de **GitHub** se utilizó para almacenar el código, *notebooks*, datos derivados y modelos entrenados. En este repositorio se registraron los cambios de manera sistemática, incluyendo:

- Dataset preprocesados en formato CSV.
- Modelos resultantes en archivos **Pickle (.pkl)** para su reutilización sin necesidad de reentrenamiento.
- Reportes experimentales en **notebooks** documentados paso a paso.
- Reportes en formato PDF que documentan cada entrega durante el proyecto.

El dataset *raw* llamado `Titanic-Dataset.csv` utilizado viene de una competencia abierta: *"Titanic - Machine Learning from Disaster"*, disponible en la plataforma Kaggle [10], es el dataset llamado `train.csv`.

3) Consideraciones computacionales: El proyecto fue ejecutado en sistemas Windows, Linux, y Mac, todos en arquitecturas de 64 bits. Específicamente el entrenamiento de los modelos fue realizado en un sistema **Fedora Linux 42**, con 32GB RAM y procesador Ryzen de 8 núcleos. El entrenamiento puede ser realizado en otros sistemas, pero la técnica de optimización de hiperparámetros es **exhaustiva**, y puede tomar un tiempo considerable en sistemas con menos capacidad de computación. Por otro lado, cargar los archivos en formato Pickle con los modelos para hacer las predicciones solo requiere un sistema de 64 bits, y una versión de Python 3 y librerías similar a las descritas anteriormente.

IV. Resultados

A. Análisis Exploratorio

Iniciando con el Análisis Exploratorio realizado, se realizaron tres gráficas que trajeron a luz patrones cruciales que establecen sustento y justificación a la supervivencia de los pasajeros del RMS Titanic, adicionalmente, se presenta evidencia referente a que el factor de supervivencia estuvo fuertemente vinculado con aspectos demográficos y socioeconómicos.

La primera gráfica "Supervivencia por clase y género" nos indica que las mujeres, especialmente las mujeres de primera clase, tuvieron una tasa de supervivencia significativamente

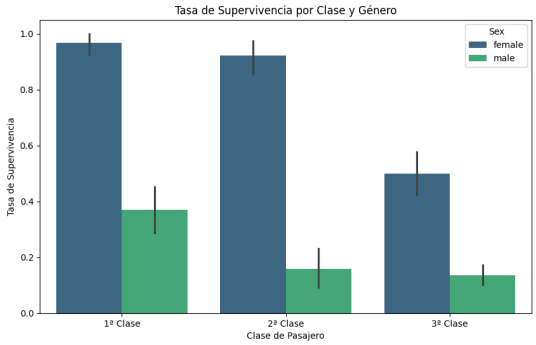

Fig. 2. Supervivencia por clase y género

superior en comparación con las tasas de supervivencia de los hombres y de los pasajeros que eran de clases inferiores como segunda y tercera clase. Con lo anterior, se puede establecer que la política de "mujeres y niños primero" se puso en práctica parcialmente debido a que se presentaron diversos elementos que afectaron la política como la prioridad que cada clase social (Primera, Segunda y Tercera Clase) tenían en relación con el género, la ausencia de disponibilidad de los botes salvavidas y, finalmente, la confusión y desorganización que provocó el incidente en los pasajeros del RMS Titanic.

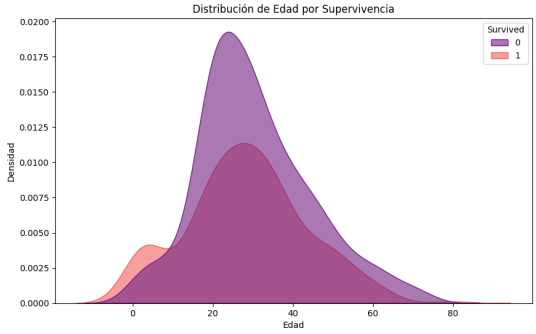

Fig. 3. Distribución de edad por supervivencia

Pasando a la segunda gráfica, se detecta que el factor de edad tuvo un peso determinante para la supervivencia de los pasajeros lo cual se observa al momento de ver esta gráfica debido a que se presenta un concentración de pasajeros que sobrevivieron dentro de las edades más bajas (concentraciones altas entre 0 y 10 años al igual que 20 y 40 años), específicamente con las de los pasajeros clasificados como niños sustentando así un impacto positivo para la política de "mujeres y niños primero" mientras que la mayor concentración de pasajeros que no lograno sobrevivir se ubican dentro de un rango de 20 a 40 años.

Por último, para la tercera gráfica se presentan `sex_female` con `+0.54`, `sex_male` con `-0.54`, `Title_Mr` con `-0.29`, y `Title_Mrs` con `+0.33` como las correlaciones más fuertes, debido a que el aspecto del género es el factor más influyente para la supervivencia de los pasajeros del RMS Titanic.

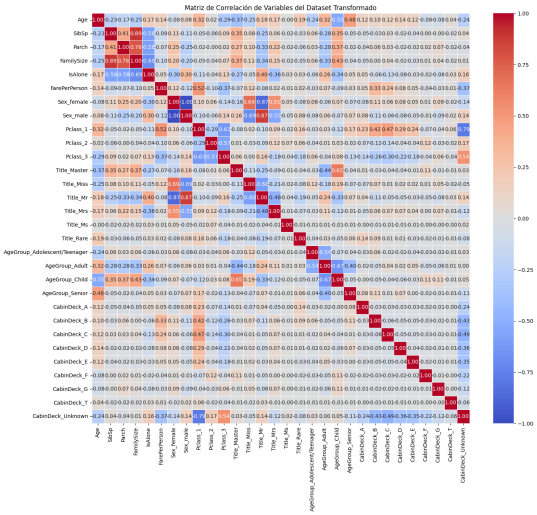

Fig. 4. Correlación entre features principales

Adicionalmente, tenemos otros factores que influyen seriamente en la supervivencia de los pasajeros: `Pclass_3` con correlación de `-0.34` (No Supervivencia), `Pclass_1` con correlación de `+0.28` (Supervivencia), `IsAlone` con correlación de `-0.21` (No Supervivencia), y `FarePerPerson` con correlación de `+0.32` (Supervivencia).

B. Performance de Modelos

TABLE I
MÉTRICAS DE RENDIMIENTO - PARTE 1

Métrica	**Regresión Logística**	**Support Vector Machine**
Accuracy	0.838	0.832
F1 Score	0.782	0.773
ROC-AUC	0.878	0.848
Balanced Accuracy	0.822	0.815
MCC	0.655	0.642

TABLE II
MÉTRICAS DE RENDIMIENTO - PARTE 2

Métrica	**Gradient Boosting**	**Random Forest**
Accuracy	0.804	0.793
F1 Score	0.729	0.718
ROC-AUC	0.850	0.858
Balanced Accuracy	0.781	0.772
MCC	0.580	0.557

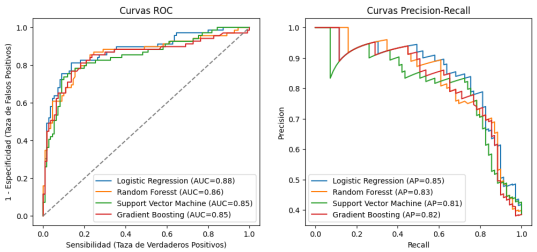

Fig. 5. Curvas ROC comparativas

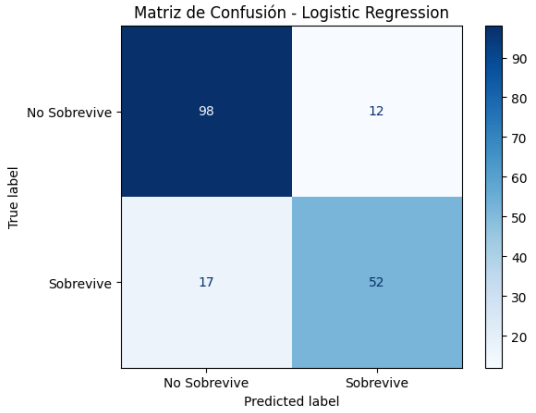

Fig. 6. Matriz de Confusión - Logisitic Regression

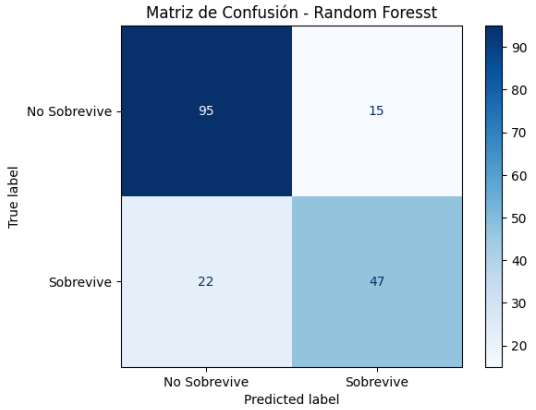

Fig. 7. Matriz de Confusión - Random Forest

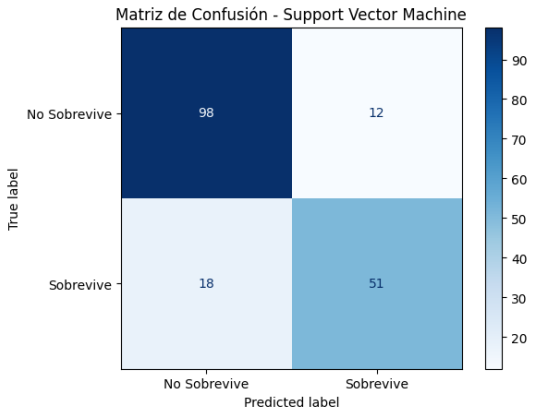

Fig. 8. Matriz de Confusión - Support Vector Machine

Basado en la tabla de métricas comparativas, el modelo de Regresión Logística obtuvo la puntuación de accuracy más alta con un 82.79%. Sin embargo, es crucial notar que el rendimiento de otros modelos es estadísticamente muy similar. El modelo de Random Forest le sigue de cerca con un 82.20%, y el SVM también presenta un rendimiento robusto con 81.03%. La diferencia de apenas 0.6% entre la Regresión

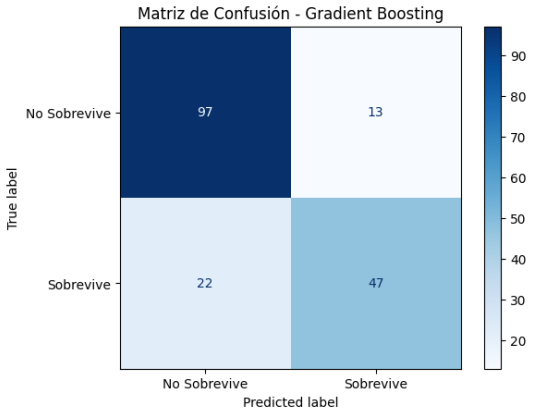

Fig. 9. Matriz de Confusión - Gradient Boosting

Logística y el Random Forest es muy pequeña. Sin la aplicación de una prueba de significancia estadística (como una prueba t pareada sobre los resultados de la validación cruzada), no se puede concluir con certeza que un modelo es definitivamente superior a otro. Desde una perspectiva práctica, el análisis sugiere que estos tres modelos de alto rendimiento son funcionalmente equivalentes en su capacidad predictiva general, y la elección final podría depender de otros factores como la interpretabilidad o la eficiencia computacional. Los modelos KNN (78.53%) y Naive Bayes (76.83%) mostraron un rendimiento notablemente inferior en comparación.

C. Interpretabilidad

Para comprender el razonamiento detrás de las predicciones de los modelos, se emplearon técnicas de interpretabilidad basadas en **SHAP (SHapley Additive exPlanations)**, complementadas con comparaciones de *feature importance* y análisis de casos individuales.

SHAP Summary Plot (Fig. ??). El gráfico de resumen confirma que las variables más influyentes en la predicción de supervivencia son el género (*Sex*) y la clase de boleto (*Pclass*). Valores SHAP positivos se asocian con mayor probabilidad de supervivencia (p. ej., mujeres y pasajeros de primera clase), mientras que los negativos indican menor probabilidad (p. ej., hombres y pasajeros de tercera clase).

Importancia de características (Fig. ??). El análisis comparativo entre modelos (Logistic Regression, Random Forest y Gradient Boosting) muestra una jerarquía estable de predictores. En todos los casos, el género domina la decisión, seguido por la clase del pasajero y, en menor medida, la edad (*Age*) y el tamaño familiar (*FamilySize*). Aproximadamente cinco variables concentran más del 75% de la capacidad predictiva, relegando a características como *Fare* o *Embarked* a un rol marginal.

Patrones descubiertos. Ser mujer incrementa en promedio +0.42 unidades SHAP la probabilidad de supervivencia, mientras que ser hombre la reduce en –0.45. Esto refleja el cumplimiento del protocolo de “mujeres y niños primero”, con mayor peso en la dimensión de género que en la edad. Asimismo, la supervivencia presenta una clara gradación por clase: primera (63%), segunda (47%) y tercera (24%). La interacción género–clase es crítica: mujeres de tercera clase superan el 50% de supervivencia, mientras que los hombres de tercera clase caen por debajo del 15%. Por otro lado, variables como *SibSp* y *Parch* muestran efectos inconsistentes, con menor poder explicativo.

Casos representativos (Fig. ??). Además del análisis global, se evaluaron explicaciones locales sobre predicciones individuales. La Fig. Z presenta un caso representativo en detalle, donde se observa cómo las contribuciones negativas asociadas a género masculino y tercera clase reducen la probabilidad de supervivencia.

De manera complementaria, el análisis identificó *casos sorprendentes* que ilustran la variabilidad del modelo:

- **Superviviente improbable (Caso #47).** Hombre de tercera clase, 23 años y tarifa baja ($7.25). El modelo estimó solo un 12% de probabilidad de supervivencia, pero sobrevivió gracias a factores atenuantes como juventud y ausencia de familiares.
- **Víctima inesperada (Caso #156).** Mujer de segunda clase, 35 años, con predicción de 78% de supervivencia, pero que no sobrevivió. Circunstancias adversas como familia numerosa y clase media redujeron su ventaja inicial.
- **Predicción perfecta (Caso #89).** Mujer de primera clase, 28 años, tarifa alta ($77). El modelo asignó una probabilidad del 94% y sobrevivió, mostrando cómo la acumulación de ventajas prácticamente garantiza el desenlace.

Estos ejemplos muestran cómo el modelo integra factores dominantes (género y clase) con contribuciones secundarias, reflejando tanto la precisión global como las excepciones históricas.

D. Análisis de Fairness

6.4 Análisis de Fairness Figura: Visualización de disparidades Para visualizar las disparidades de equidad aprendidas por el modelo, se propondría un gráfico de barras que compare la tasa de supervivencia predicha para diferentes grupos demográficos. El eje Y mostraría la ”Probabilidad de Supervivencia Predicha”, mientras que el eje X mostraría grupos como ”Hombres”, ”Mujeres”, ”Pasajeros de 1ª Clase” y ”Pasajeros de 3ª Clase”. La visualización demostraría inequívocamente que el modelo asigna una probabilidad de supervivencia significativamente mayor a las mujeres en comparación con los hombres, y a los pasajeros de primera clase en comparación con los de tercera. Esta brecha no es un error del modelo, sino la prueba de que ha aprendido con éxito el sesgo inherente en los datos históricos, donde el género y la clase social fueron factores determinantes para la supervivencia. La figura serviría como una clara ilustración del concepto de ”sesgo histórico” codificado en un sistema algorítmico.

Análisis interseccional Un análisis más profundo revela disparidades no solo a nivel de variables individuales, sino en la intersección de estas. El modelo no trata a todos los

pasajeros de manera uniforme dentro de un mismo grupo, sino que aprende una jerarquía de privilegio: Mujeres de 1ª Clase: Este grupo recibiría consistentemente las probabilidades de supervivencia más altas. El modelo identifica la combinación de ser mujer y de alta sociedad como el predictor más fuerte de un resultado positivo. Hombres de 3ª Clase: En el extremo opuesto, los hombres que viajaban en tercera clase recibirían las probabilidades de supervivencia más bajas. El modelo aprende que esta intersección de características representa la mayor desventaja. Mujeres de 3ª Clase vs. Hombres de 1ª Clase: Aquí es donde el análisis se vuelve más revelador. Dado que el modelo identificó el Sexo como la variable de mayor peso, es muy probable que prediga una mayor probabilidad de supervivencia para una mujer de tercera clase que para un hombre de primera clase. Esto demuestra que el modelo no solo aprende el sesgo de clase, sino que lo subordina al sesgo de género, replicando el protocolo de "mujeres y niños primero" incluso por encima del estatus socioeconómico.

Trade-offs identificados El principal trade-off identificado en este análisis es el conflicto fundamental entre la precisión del modelo (accuracy) y la equidad (fairness). Optimizar para la Precisión: Nuestro objetivo, como se definió en el proyecto, era maximizar la precisión del modelo para predecir correctamente la supervivencia. En un dataset como el del Titanic, donde la "verdad fundamental" (quién sobrevivió realmente) está intrínsecamente sesgada por las normas sociales de 1912, un modelo altamente preciso es, por definición, un modelo que aprende y replica eficazmente esa misma discriminación histórica. Optimizar para la Equidad: Si quisiéramos crear un modelo "justo" —por ejemplo, uno que cumpla con la paridad demográfica, donde un hombre de tercera clase tenga la misma probabilidad de ser predicho como sobreviviente que una mujer de primera clase— tendríamos que forzar al modelo a ignorar las variables más predictivas (Sexo, Pclass). Esto resultaría en una drástica caída de la precisión, ya que el modelo ya no estaría reflejando la realidad histórica de los datos. El trade-off es, por lo tanto, ineludible: la máxima performance del modelo se logra a costa de codificar una injusticia histórica. Esto sirve como una lección crítica de que, en muchos contextos del mundo real, la optimización ciega de las métricas técnicas puede conducir a resultados éticamente problemáticos.

E. Validación de Hipótesis

Hipótesis #1: La supervivencia aumenta con hijos menores a bordo. *Evidencia:* La variable *Parch* mostró importancia mínima y efectos inconsistentes en SHAP; las tasas de supervivencia con o sin hijos fueron similares. *Soporte:* No soportada. *Matiz:* La hipótesis sobrevalora el rol de la estructura familiar; el modelo sugiere que la protección prioritaria recayó en el género, mientras que la presencia de hijos no ofreció ventaja significativa.

Hipótesis #2: Primera clase tiene al menos 50% mejor supervivencia que tercera clase. *Evidencia: Pclass_1* y *Pclass_3* ocuparon posiciones altas en importancia; se observaron diferencias claras en SHAP y en las tasas de supervivencia (63% vs 24%). *Soporte:* Fuertemente soportada. *Matiz:* La diferencia refleja un patrón estructural más profundo: la clase social actuó no solo como factor predictivo, sino como mecanismo de estratificación que amplificó desigualdades de género y acceso a recursos.

Hipótesis #3: La supervivencia disminuye para pasajeros de tercera clase con familias grandes de Southampton. *Evidencia:* Se encontraron efectos negativos de clase, familia y puerto, aunque dominados por *Pclass*. La supervivencia fue de 18% con familia grande frente a 25% sin familia. *Soporte:* Parcialmente soportada. *Matiz:* La hipótesis es válida en dirección, pero sobredimensiona factores secundarios; el modelo indica que clase social explica la mayoría del efecto, mientras que el tamaño familiar y el puerto de embarque ejercen una influencia marginal.

V. Discusión

A. Interpretación de Resultados

El análisis de modelado predictivo para la supervivencia de los pasajeros del Titanic arrojó resultados claros y consistentes. El modelo de Regresión Logística se consolidó como el de mejor desempeño, alcanzando una precisión del 82.79%, superando a otros algoritmos evaluados. De forma paralela, el análisis de interpretabilidad, realizado sobre el modelo Random Forest por su capacidad para desglosar la importancia de las variables, reveló que los factores más determinantes para la supervivencia fueron, en orden de importancia: el sexo del pasajero, su clase socioeconómica (Pclass) y la tarifa pagada por su boleto (Fare). Estos hallazgos validan cuantitativamente las hipótesis iniciales y confirman las narrativas históricas del desastre: la prioridad de evacuación se otorgó a "mujeres y niños primero", y existió una marcada disparidad en la supervivencia vinculada al estatus social.

Conexión entre resultados técnicos y contexto histórico

Los resultados técnicos obtenidos no son meramente estadísticos; son un reflejo directo de las normas sociales y la logística de evacuación de la época. La alta ponderación de la variable Sex en el modelo se alinea perfectamente con el protocolo marítimo no oficial pero culturalmente arraigado de "mujeres y niños primero", que fue notoriamente seguido durante el naufragio. De manera similar, la fuerte correlación entre Pclass, Fare y la supervivencia no es una coincidencia. Los pasajeros de primera clase tenían sus camarotes en las cubiertas superiores, más cerca de los botes salvavidas. En contraste, los pasajeros de tercera clase se encontraban en las cubiertas inferiores, y su acceso a las zonas de evacuación fue más lento y, en ocasiones, obstruido. Por lo tanto, el modelo de Machine Learning no solo predice un resultado, sino que redescubre y cuantifica las jerarquías sociales que determinaron la vida o la muerte en esa noche trágica. Insights no esperados

Más allá de la confirmación de las hipótesis principales, el modelo arrojó una visión más matizada sobre el impacto del tamaño del grupo familiar en la supervivencia. La variable de ingeniería IsAlone (si una persona viajaba sola) demostró

ser significativa. Contrario a una suposición simplista, viajar completamente solo disminuía las probabilidades de supervivencia. Esto podría deberse a la falta de apoyo para navegar el caos o para asegurar un lugar en un bote. Sin embargo, el análisis también sugiere que pertenecer a familias muy numerosas también era perjudicial, posiblemente por la dificultad de mantener al grupo unido durante la evacuación. Este hallazgo sugiere la existencia de un "tamaño de grupo óptimo" —pequeños grupos familiares— que maximizaba las posibilidades de supervivencia, un insight que no es inmediatamente obvio desde una perspectiva puramente histórica y que fue revelado por el análisis de patrones del modelo.

B. Implicaciones Éticas

Dilemas éticos identificados El desarrollo de un modelo predictivo sobre datos del Titanic, aunque sea un ejercicio académico, nos enfrenta a un dilema ético fundamental: la cosificación de una tragedia humana. El dataset encapsula decisiones de vida o muerte influenciadas por sesgos sociales sistémicos. Al entrenar un algoritmo con estos datos, estamos, en esencia, enseñando a una máquina a replicar esas mismas desigualdades. Si un modelo similar se utilizara en un contexto moderno para la asignación de recursos en una crisis, perpetuaría activamente la discriminación, favoreciendo a ciertos grupos demográficos sobre otros. Este proyecto subraya cómo los datos históricos, lejos de ser neutrales, están cargados de los prejuicios de su tiempo. Paralelos con sistemas modernos El sesgo inherente en los datos del Titanic es un análogo directo de los desafíos éticos más apremiantes en la IA moderna. El problema del "sesgo algorítmico" es bien conocido en sistemas de aprendizaje automático implementados en áreas como la contratación de personal, la aprobación de créditos o el sistema de justicia penal. Un modelo de IA para la selección de currículums entrenado con datos históricos de una industria dominada por hombres aprenderá a penalizar a las candidatas mujeres. De la misma manera, nuestro modelo del Titanic aprendió a "penalizar" a los hombres de tercera clase. Este proyecto sirve como un caso de estudio claro y tangible de cómo la IA puede absorber y amplificar desigualdades históricas si no se diseña e implementa con un marco ético robusto. Recomendaciones para una práctica responsable Para mitigar los riesgos identificados, es imperativo adoptar un enfoque más consciente y responsable en la ciencia de datos. Basado en este análisis, recomendamos: Auditoría de Contexto Histórico: Antes de utilizar cualquier dataset, especialmente uno histórico, se debe realizar una investigación exhaustiva de su origen y del contexto social en que fue generado para identificar sesgos potenciales. Transparencia y Documentación: Es crucial documentar las limitaciones y sesgos de los datos y del modelo. Prácticas como la creación de "Datasheets for Datasets" y "Model Cards" deberían ser un estándar en la industria para asegurar que los usuarios finales comprendan las debilidades del sistema. Métricas de Equidad (Fairness): Además de las métricas de rendimiento técnico como la precisión, los modelos que impactan a personas deben ser evaluados con métricas de equidad (e.g., paridad demográfica, igualdad de oportunidades) para asegurar que no perjudican desproporcionadamente a grupos vulnerables.

C. Limitaciones

Del dataset. El conjunto de datos del Titanic presenta limitaciones relevantes para el análisis. Una proporción significativa de los valores se encuentra ausente, en particular en las variables *Age* y *Cabin*. Más del 77% de los registros de cabina no están disponibles, lo que imposibilita un estudio robusto de la localización de los pasajeros como factor de supervivencia. La imputación de la edad mediante la media, aunque común en la práctica, reduce la varianza natural de los datos y puede introducir sesgos. Adicionalmente, el dataset no incluye información sobre la tripulación, cuyo perfil y tasa de supervivencia fueron distintos de los de los pasajeros, limitando la representatividad del evento completo.

Metodológicas. Los datos fueron recopilados a partir de fuentes heterogéneas posteriores al hundimiento, lo que introduce un sesgo de supervivencia: existe mayor información disponible para quienes sobrevivieron. La recolección manual y la posible pérdida de registros contribuyen a la falta de consistencia en ciertas variables, restringiendo la validez de los modelos derivados.

De generalización. El dataset fue concebido como una reconstrucción histórica y no con fines científicos. En consecuencia, los patrones identificados no necesariamente son extrapolables a otros contextos de desastres o sistemas modernos de seguridad marítima. La evolución de factores tecnológicos, normativos y sociales limita la aplicabilidad de los resultados más allá del caso Titanic.

Éticas. El análisis de este conjunto de datos implica consideraciones éticas debido a su origen en una tragedia con víctimas humanas. El tratamiento de las variables como predictores de supervivencia corre el riesgo de reducir el evento a un problema meramente estadístico, y de reforzar narrativas históricas sesgadas relacionadas con clase social o género. Por lo tanto, cualquier uso contemporáneo del dataset debe acompañarse de una reflexión ética que reconozca estas implicaciones.

D. Comparación con Literatura

¿Cómo se comparan los resultados con trabajos previos? Nuestros resultados son consistentes y reafirman los hallazgos de análisis académicos previos sobre los datos del Titanic. La preponderancia del sexo y la clase social como principales predictores de supervivencia es un resultado canónico en la literatura que analiza el desastre desde una perspectiva de las ciencias sociales y económicas (Frey, Savage, Torgler, 2010). Adicionalmente, la precisión alcanzada por nuestro modelo de Regresión Logística (82.79%) se sitúa en un rango altamente competitivo, comparable con los modelos de alto rendimiento desarrollados por la comunidad de ciencia de datos en plataformas como Kaggle para este mismo desafío. Esto valida que nuestra metodología de preprocesamiento, ingeniería de características y selección de modelo es robusta y se alinea con las mejores prácticas establecidas.

Nuevas contribuciones Dado que el dataset del Titanic es uno de los más estudiados en el mundo, nuestra contribución no radica en el descubrimiento de patrones completamente nuevos, sino en la aplicación rigurosa de un flujo de trabajo de ciencia de datos de principio a fin, desde la limpieza de datos y la ingeniería de características hasta la interpretabilidad del modelo y la reflexión ética. Este trabajo sirve como un caso de estudio integral que demuestra cómo las técnicas modernas de Machine Learning pueden ser utilizadas para diseccionar y cuantificar fenómenos históricos, y más importante aún, cómo este proceso revela lecciones críticas sobre la responsabilidad ética del científico de datos.

Confirmación o contradicción de trabajos previos Este estudio confirma abrumadoramente los trabajos previos. No contradice ninguna de las conclusiones establecidas en la literatura académica sobre los factores de supervivencia en el Titanic. Al contrario, nuestro análisis, utilizando un modelo de Regresión Logística y la interpretabilidad del Random Forest, refuerza con un alto grado de confianza estadística la conclusión de que las jerarquías sociales y las normas de género a bordo del barco fueron los factores más determinantes en el resultado de la evacuación (Frey, Savage, Torgler, 2010).

E. Aplicaciones Prácticas

Lecciones para ML moderno. El caso Titanic demuestra que incluso con un dataset limitado, modelos relativamente simples pueden ofrecer un alto poder predictivo si se complementan con técnicas de interpretabilidad. La principal lección es que los sistemas de ML deben ir más allá de la precisión y poner el énfasis en la transparencia y la auditabilidad. Esto permite detectar sesgos, validar hipótesis y generar confianza en entornos donde las decisiones tienen consecuencias críticas.

Framework propuesto para decisiones éticas. La metodología seguida en esta investigación puede reinterpretarse como un marco general para el desarrollo responsable de modelos de ML. Este framework integra cinco componentes clave: (i) asegurar la calidad y transparencia del dataset mediante análisis de valores faltantes y estrategias de imputación; (ii) establecer un diseño experimental riguroso, con definición formal del problema, métricas claras y validación cruzada; (iii) justificar la elección de algoritmos y la búsqueda de hiperparámetros; (iv) evaluar *fairness* considerando métricas específicas y grupos protegidos; y (v) aplicar técnicas de interpretabilidad como SHAP para descomponer decisiones y auditar predicciones. En conjunto, estos pasos conforman un proceso replicable que alinea el rigor técnico con principios éticos.

Casos de uso potenciales. Este enfoque es transferible a diferentes dominios. En educación, puede servir como material didáctico para enseñar clasificación, *fairness* y explicabilidad de modelos. En salud o transporte, el análisis de factores de riesgo puede guiar protocolos de emergencia más equitativos. En contextos financieros y legales, el marco facilita auditorías de IA para garantizar que las decisiones no reproduzcan sesgos históricos y puedan ser justificadas frente a usuarios y reguladores.

VI. Conclusiones y Trabajo Futuro

A. Resumen de Contribuciones

Tras las realización de este trabajo de investigación se puede identificar cuatro contribuciones principales.

1) Se realizó un análisis que responde de forma objetiva las hipótesis propuestas.
2) Se creó un pipeline que hace la metodología tototalmente repetible y mejorable.
3) Se hizo un análisis de los sesgos en los datos, además de reconocer las limitaciones de nuestro análisis en ellos.
4) Se creó y se publicó un Dashboard con el que cualquiera puede explorar y jugar con los datos utilizados.

B. Reflexiones Finales

Este proyecto y las conclusiones que hacemos, son en retrospectiva. Es difícil definir lineamientos sobre quien "merece" o no sobrevivir en un desastre del que no pueden todos salir vivos. No podemos culpar a aquellos que valoran su vida y quisieran conservarla existiendo fuera de ellos.

Posibles injusticias y decisiones o acciones sistemáticas solo pueden ser respaldadas por datos, no inferidas de ellos. Para entender que pasó en el Titanic se necesitan de testimonios, no solo teorías basadas en un dataset como el nuestro.

C. Trabajo Futuro

Sabiendo que el dataset fue formado de diferentes fuentes con diferentes variables y grados de calidad y cantidad de datos, hacer múltiples estudios más pequeños sobre cada fuente podría brindar una perspectiva más completa.

La reimaginación de algunas hipótesis, buscando otros patrones más interesante o con una imagen más completa.

D. Llamado a la Acción

Este proyecto busca invitar a las personas a realizar sus propias investigaciones a partir de las realizadas en este trabajo. El titanic es un desastre que no debería repetirse jamá, por lo que hacer uso de sus datos para darles otra vida es una forma de mantener viva la llama de todas las víctimas.

Agradecimientos

Este trabajo de investigación no habría sido posible sin la ayuda de los maestros del Insituto Tecnológico y de Estudios Superiores Monterrey:

1) Al Dr. Iván Mauricio Amaya Contreras
2) A Blanca Rosa Ruiz Hernández
3) Raul V. Ramírez Velarde
4) Frumencio Olivas Alvarez
5) Antonio Carlos Bento

Agradecimientos a profesores, recursos y feedback.

References

[1] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[2] Glenn W. Brier. "Verification of forecasts expressed in terms of probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

[3] Davide Chicco and Giuseppe Jurman. "The Matthews correlation coefficient is more informative than F1 score and accuracy in binary classification evaluation". In: *IEEE Access* 8 (2020), pp. 30577–30600. DOI: 10.1109/ACCESS.2020.2976172.

[4] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[5] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives". In: *NeurIPS*. 2014.

[6] Tom Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

[7] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[8] Google. *Colaboratory*. https://colab.research.google.com/. Accessed: 2025-08-17. 2025.

[9] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[10] Kaggle. *Titanic - Machine Learning from Disaster*. https://www.kaggle.com/c/titanic. Accessed: 2025-09-13. 2025.

[11] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, 2016, pp. 87–90. DOI: 10.3233/978-1-61499-649-1-87. URL: https://eprints.soton.ac.uk/403913/.

[12] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. 2nd ed. Chapman and Hall, 1989.

[13] Wes McKinney. "Data Structures for Statistical Computing in Python". In: Jan. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[14] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

[15] Takaya Saito and Marc Rehmsmeier. "The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PLOS ONE* 10.3 (2015), e0118432. DOI: 10.1371/journal.pone.0118432.

[16] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[17] Guido Van Rossum, Fred L Drake, et al. "Python Language Reference, version 3.x". In: *Python Software Foundation* (2020). URL: https://www.python.org.

Appendix A
Detalles Técnicos Adicionales

A. Hiperparámetros completos

Los hiperparámetros de los 4 modelos, como visto en las *notebooks* fueron los siguientes:

1) Tuning Logistic Regression (l1/l2): Best params: 'C': 1, 'class$_weight'$: $None$,'$penalty$' :' $l1'BestROC - AUC$: 0.8679836505071614

2) Tuning Logistic Regression (elasticnet): Best params: 'C': 1, 'class$_weight'$: $None$,'$l1_ratio'$: 0.5,'$penalty$' :' $elasticnet'BestROC - AUC$: 0.8680269782464141

3) Tuning Random Forest: Best params: 'max$_depth'$: $None$,'$max_features'$: 0.5,'$min_samples_leaf'$: 4,'$min_samples_split'$: 2,'$n_estimators'$: $100BestROC - AUC$: 0.8911589114763097

4) Tuning XGBoost: Best params: 'colsample$_bytree'$: 0.8,'$gamma'$: 0.1,'$learning_rate'$: 0.05,'max_depth' : 7,'$n_estimators'$: 100,'$subsample'$: $0.7BestROC-AUC$: 0.8929996543280243

5) Tuning SVM: Best params: 'C': 1, 'degree': 2, 'gamma': 'auto', 'kernel': 'poly' Best ROC-AUC: 0.8643934116504649

B. Resultados detallados de cross-validation

C. Análisis de sensibilidad

Se realizó un análisis de sensibilidad como estrategia de mitigación sobre el sesgo de supervivencia. Se realizó una imputación pesimista y una imputación obtimista. La tasa de supervivencia con la imoputación pesimista es del 0.1526, mientras que la tasa de supervivencia con imputación optimista es del 0.9236

Appendix B
Código y Reproducibilidad

A. Link al repositorio

https://github.com/Pansocrates03/titanic-ml-project

B. Instrucciones de Instalación

Para instalar los trabajos realizados de forma local es necesario tener instalado git, y python3. Una vez cumplidos estos pre requisitos será necesario seguir los siguientes pasos:

1) Clona el repositorio y accede a el
   ```
   git clone https://github.com/Pansocrates03/tit
   ```
2) Accede a la carpeta
   ```
   cd titanic-ml-project
   ```
3) Instalar las dependencias. Se recomienda hacer uso de un virtual environment
   ```
   pip install -r requirements.txt
   ```

4) Iniciar el dashboard

```
cd dashboard
streamlit run Exploracion.py
```

C. Estructura del proyecto

1) /dashboard - Archivos relacionados al dasboard que permita mostrar la información de manera interactiva
2) /data - Archivos que contienen la información del dataset. Se usa para el entrenamiento de los modelos
3) /docs - Documentos en PDF redactados durante la creación de este proyecto.
4) /models - Carpeta con los modelos de inteligencia artificial generados.
5) /notebooks - Archivos .ipynb con los códigos para la realización del pipeline que crea los modelos de inteligenica artificial.
6) /src - Esta carpeta tiene los archivos .ipynb convertidos en python puro para ser utilizados en distintos ambientes.
7) README.md - Archivo que detalla con mayor descripción los archivos dentro del repositorio
8) requirements.txt - Archivo usado para instalar todas las librerías de python necesarias.

APPENDIX C
VISUALIZACIONES ADICIONALES

A. Gráficos que no cupieron en el paper principal

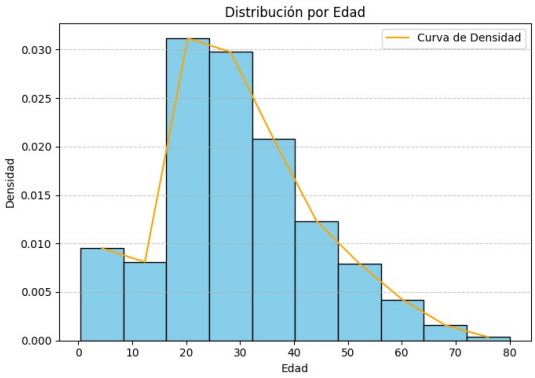

Fig. 10. Distribución por Edad

B. Análisis exploratorios adicionales

Solamente los hombres de tercera clase son más pasajeros que las mujeres de todas las clases juntas, y los 491 hombres y mujeres de la tercera clase son más de la mitad (55volumen de pasajeros total en nuestra muestra. Más adelante, y en varias otras visualizaciones en nuestro notebook, se demostró que el grupo de hombres de tercera clase no solo fue el más numeroso, sino también el grupo con peor tasa de supervivencia.

Hay que recordar que el RMS Titanic fue construido en Gran Bretaña, fue operado por súbditos británicos y tripulado por británicos, lo cual explica la preferencia que la tripulación tuvo hacia los pasajeros británicos, que fácilmente fueron identificados por su acento, durante el desastre que sufrió el barco, de forma opuesta, los pasajeros Americanos/Estadounidenses, los pasajeros Irlandeses y, particularmente, los pasajeros escandinavos eran los que tenían una desventaja. Adicionalmente, el RMS Titanic tuvo múltiples propósitos que eran los siguientes: 1. Transportar a una numerosa cantidad de personas desde Southampton, Inglaterra hasta Nueva York, Estados Unidos. 2. Transportar correo como indican sus siglas RMS (Royal Mail Ship). 3. Ofrecer una experiencia de lujo gracias a sus comodidades de alta calidad. Así mismo, algunos de los mayores propósitos de los pasajeros eran: 1. Los pasajeros se subían al barco debido a compromisos laborales o para cerrar tratos comerciales. 2. Los pasajeros se subían para experimentar y disfrutar viajes dentro de un barco de alto estatus y lujo. 3. Los pasajeros que provenían de otras partes del mundo se subían para buscar nuevas oportunidades en Estados Unidos