Mujeres y niños primero: Predicción de Supervivencia en el Titanic con Machine Learning

Inteligencia Artificial Avanzada para la Ciencia de Datos

Presentado por:

Samuel Pelaez Aleman

Alain Hurtado Escamilla

Diego Velázquez Saldaña

María Fernanda Argueta

Diego Loyo Villagrán

Profesores:

Antonio Carlos Bento

Alfredo Esquivel Jaramillo

Mauricio González Soto

Julio Antonio Juárez Jiménez

Frumencio Olivas Alvarez

Jesús Adrián Rodríguez Rocha

Hugo Terashima Marín

Agosto 2024

En este informe, se presenta un análisis exhaustivo de los factores que influyeron en la supervivencia de los pasajeros del Titanic, utilizando técnicas de limpieza de datos y modelos de Machine Learning para la predicción. A través de la identificación de variables clave como la clase del pasajero, el sexo, la edad y el punto de embarque, se busca comprender mejor qué características aumentaron o disminuyeron las probabilidades de supervivencia.

Keywords: Machine Learning, Clasificación, Supervivencia, Titanic, Modelos Predictivos

I. INTRODUCCIÓN

El hundimiento del RMS Titanic en 1912 es uno de los desastres marítimos más trágicos y estudiados de la historia. A pesar de la gran cantidad de investigaciones realizadas sobre el evento, aún persisten preguntas importantes acerca de los factores que influyeron en las tasas de supervivencia de los pasajeros y la tripulación. Comprender qué características de los pasajeros influyeron en su supervivencia no solo ofrece una visión histórica de los eventos, sino que también proporciona importantes lecciones en la gestión de emergencias y la planificación de evacuaciones en situaciones de crisis.

El reto a resolver se obtuvo de la plataforma de Kaggle, la cual proporciona una base de datos previamente dividida entre train y test [1]. Dichas bases de datos cuentan con información detallada sobre los pasajeros. En la cual se hizo un exhaustivo análisis y exploración de los datos. En donde se procuró identificar el comportamiento de las distintas variables, su relación entre sí e incluso su influencia en la supervivencia del pasajero. Luego se realizó imputación, eliminación y creación de nuevos datos.

II. PLANTEAMIENTO DEL PROBLEMA

A continuación se describen todos los procedimientos y el análisis llevado a cabo para resolver el problema de Supervivencia del Titanic aplicando ciencia de datos y generando un modelo de Machine Learning para llegar a la solución óptima.

El objetivo principal de este análisis es identificar y analizar las variables clave que influyeron en la probabilidad de supervivencia de los pasajeros del Titanic. Variables como la clase del pasajero, el sexo, la edad, y la cantidad de personas con las que se abordó se consideran potencialmente significativas y serán examinadas en detalle. Al mismo tiempo, el manejo adecuado de los datos faltantes y la limpieza del conjunto de datos son esenciales para garantizar la validez y la fiabilidad de los resultados obtenidos.

Por lo tanto, este estudio busca responder a la siguiente pregunta: ¿Qué factores específicos afectaron la probabilidad de supervivencia de los pasajeros del Titanic, y cómo se pueden interpretar estos factores en un contexto más amplio de gestión de riesgos y seguridad? A través de un análisis exhaustivo de los datos disponibles, este informe pretende proporcionar respuestas claras y basadas en evidencia a esta cuestión fundamental.

III. LIMPIEZA Y TRANSFORMACIÓN DE DATOS

A. Conjunto de datos

El problema se plantea con dos archivos csv que contienen datos acerca de los pasajeros del titanic. Por una parte esta el archivo de entrenamiento (con el que se calculará el modelo) y uno de prueba (con el que se probará la efectividad del modelo). Dentro de cada archivo se destacan 11 variables (columnas) que definen a cada pasajero. Dentro del archivo de entrenamiento existen 892 pasajeros (instancias), mientras que en el archivo de prueba hay 419. Las variables se describen en Tabla I y Tabla II los valores que éstas pueden tomar.

Variable	Descripción	Tipo
PassengerId	Identificador	Numérico
Survived	Sobrevivió o no sobrevivió	Categórico
Pclass	Clase del ticket	Categórico
Name	Nombre del pasajero	Texto
Sex	Sexo	Categórico
Age	Edad del pasajero	Numérico
SibSp	Hermanos/cónyuges a bordo	Numérico
Parch	Padres/hijos a bordo	Numérico
Ticket	Número del ticket	Texto
Fare	Tarifa del pasajero	Numérico
Cabin	Número de cabina	Texto
Embarked	Puerto de embarque	Categórico

Table I. Descripción y tipo de dato de las variables del problema del Titanic.

Variable	Valor
PassengerId	1 - 891
Survived	0, 1
Pclass	1, 2, 3
Name	"Braund, Mr. Owen Harris"
Sex	male, female
Age	0 - 80
SibSp	0 - 8
Parch	0 - 6
Ticket	"A/5 21171"
Fare	0 - 512
Cabin	"A10"
Embarked	Q, S, C

Table II. Valores que las variables pueden tomar.

B. Exploración de datos

Exploramos los datos de cada variable para entender como se comporta cada una. A partir de este análisis determinamos que variables tienen valores nulos, que distribuciones toman las variables y así poder determinar en la sección III D si algunas variables se deberán transformar o eliminar.

- Passenger id: es una variable sin relevancia para el modelo con ningún renglón vacío, sin embargo es importante para la identificación de los pasajeros.
- Survived: es la variable objetivo con la que vamos a entrenar y evaluar qué tan bien predice nuestro modelo, y no cuenta con ningún renglón vacío. En la Fig 1 observamos la distribución de la variable.

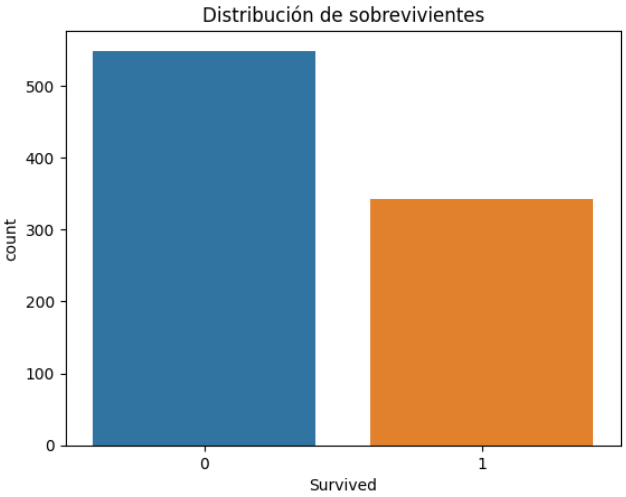

Figure 1. Distribución de sobrevivientes

- Pclass: es la variable que describe la clase en la que viajaba el pasajero, misma que creemos tiene mucha relación con la probabilidad de supervivencia, y no cuenta con ningún renglón vacío. En la Fig 2 notamos como se distribuyen los datos de la variable.

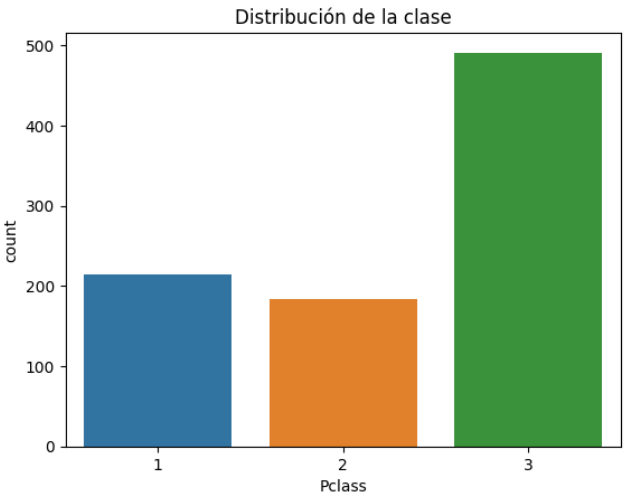

Figure 2. Distribución de clase

- Name: es el nombre del pasajero que se separa en tres partes, "Apellido, Título Nombre". La parte del título puede ayudar a agrupar por edad.

- Sex: corresponde al sexo del pasajero, mismo que impacta mucho en la supervivencia del mismo. En la Fig 3 graficamos la distribución de la variable Sex.

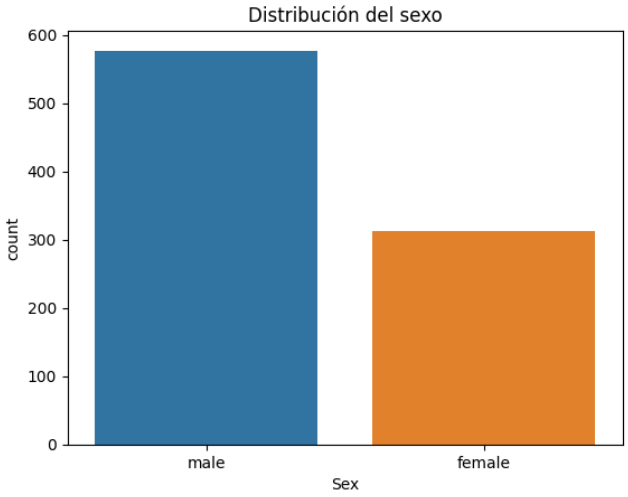

Figure 3. Distribución del sexo

- Age: es la edad del pasajero, éste afecta significativamente a la superviviencia del pasajero; igualmente esta variable tiene un 20% de datos faltantes. Observamos en la Fig 4 como se distribuyen las edades.

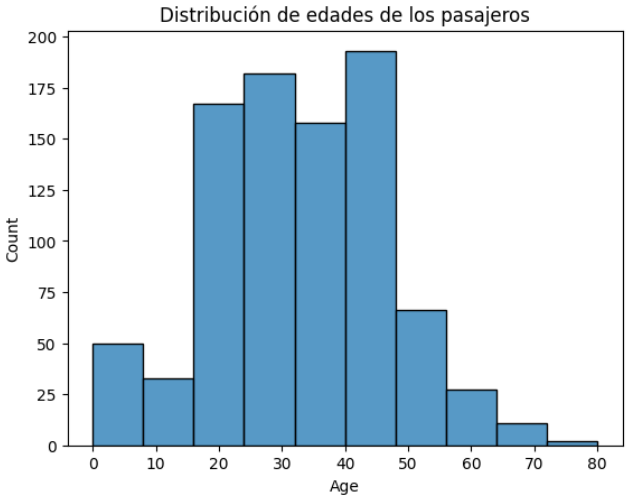

Figure 4. Distribución de edades

- SibSp: la base de datos clasifica esta columna en dos segmentos, el primero es *sibling* que corresponde a hermanos, hermanas, hermanastros o hermanastras. Y por otro lado está *spouse* que corresponde a esposo o esposa. En la Fig 5 graficamos como se dstribuyen los datos de la variable.

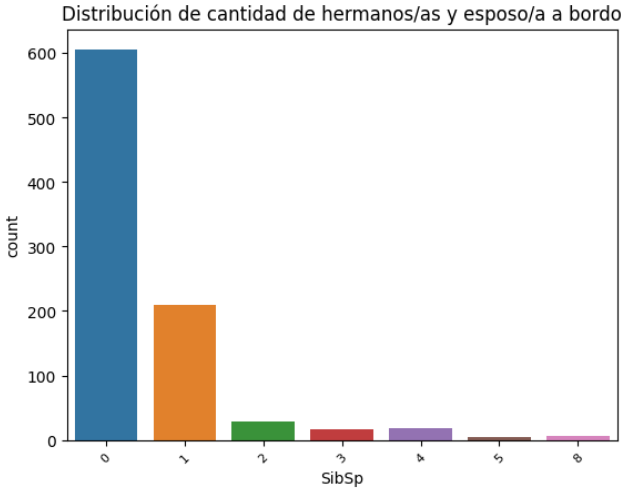

Figure 5. Distribución de siblings y spouses

- Parch: ésta también se divide en dos segmentos, el primero es *parents* que es el número de madres y padres. El otro segmento es *children* que es el número hijos, hijas, hijastro e hijastra que también estaban abordo del barco. Graficamos en la Fig 6 la distribución de los datos de la variable.

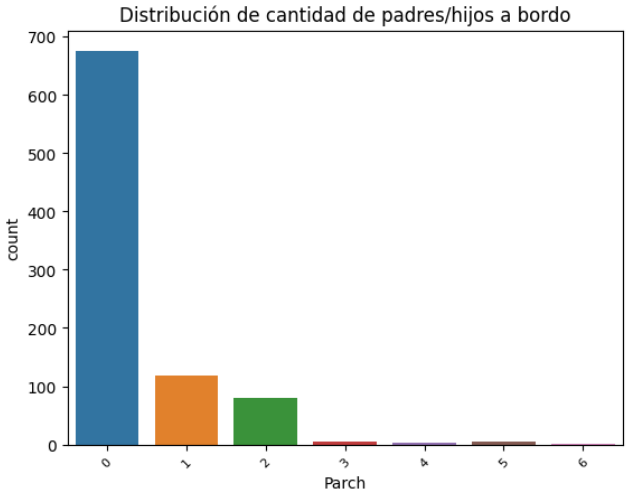

Figure 6. Distribución de parents y children

- Ticket: el número del boleto del pasajero, el cual

se puede descomponer para extraer mayor información de donde se encontraba el pasajero.

- Fare: el costo del boleto del pasajero, no cuenta con ningún renglón vacío en la base de datos de entrenamiento pero tiene 2 renglones vacíos en la base de prueba.
- Cabin: número de camarote asignado al pasajero, sin embargo, cuenta con 77% de datos faltantes.
- Embarked: puerto en el que el pasajero embarcó, al cual dentro de la base de datos de entrenamiento le faltaban dos renglones con información, mientras que en el de prueba estaba completo. En la Fig 7 observamos la distribución de los puertos.

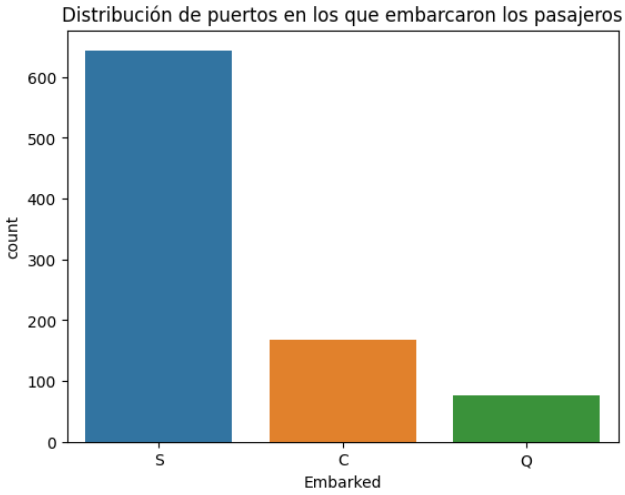

Figure 7. Distribución de puertos

C. Limpieza

Durante el proceso de preparación de los datos de la base de supervivientes del Titanic, se realizaron varias acciones para garantizar la integridad y calidad del conjunto de datos; es importante considerar que los cambios que se les ajusta a una de las bases de datos, se le debía de hacer a la otra

1. Eliminación de la columna 'Cabin': Se decidió eliminar la columna 'Cabin' debido a que presentaba más del 77% de datos faltantes. La cantidad tan significativa de valores ausentes en esta columna podría haber introducido sesgos o ruido en los análisis posteriores, por lo que se optó por removerla completamente.

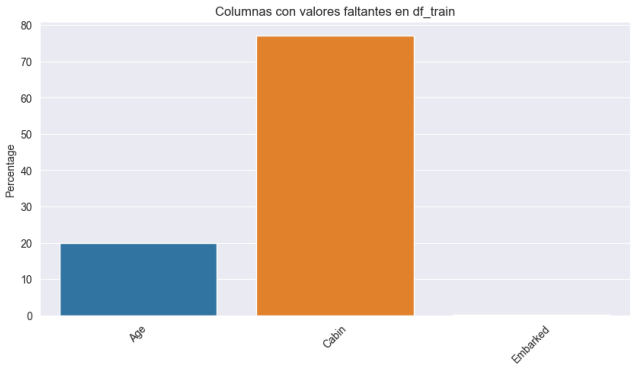

Figure 8. Features con valores faltantes en el set de entrenamiento

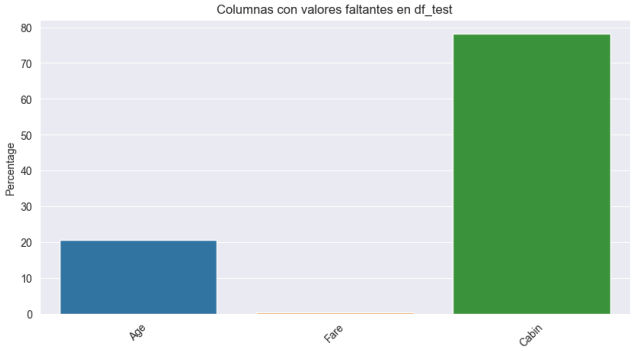

Figure 9. Features con valores faltantes en el set de evaluación

2. Remoción de filas con datos faltantes en la columna 'Embarked': Dentro del conjunto de datos de entrenamiento, se identificaron 2 pasajeros a los que les faltaba información en la columna 'Embarked'. Dado que esta información es crucial para el análisis, se decidió eliminar estas filas para evitar posibles inconsistencias en los resultados.

3. Remoción de fila con dato faltante en la columna 'Fare': En el conjunto de datos de prueba, se detectó que un pasajero carecía de información en la columna 'Fare'. Al igual que en el caso an-

terior, se decidió eliminar esta fila para mantener la coherencia del análisis.

4. Eliminación de la columna 'PassengerId': Tanto en el conjunto de datos de entrenamiento como en el de prueba, se removió la columna 'PassengerId'. Esta columna, que solo sirve como identificador, no aportaba valor al análisis predictivo, por lo que su eliminación ayudó a simplificar el modelo sin pérdida de información relevante.

D. Transformación

Con el fin de agrupar elementos que tenían relación y para reducir la cantidad de features con los que se iba a trabajar se transformaron los datos,lo cual consistió en generar nuevas columnas en base a la unión de otras o tomando en cuenta información proveniente de otros features.

- Dado que la columna "Age" presentaba un 19% de datos nulos en el dataset de entrenamiento y al ser una variable que podría proporcionar información valiosa, utilizamos una técnica para completar los valores faltantes. Para ello, primero separamos el título de las personas de su nombre. Usamos los títulos: "Capt", "Col", "Countess", "Don", "Dr", "Jonkheer", "Lady", "Major", "Master", "Miss", "Mlle", "Mme", "Mr", "Mrs", "Ms", "Rev" y "Sir", para agrupar a los pasajeros y calcular la media de edad de cada grupo, excluyendo aquellos registros donde la edad era desconocida. Posteriormente, para aquellos pasajeros cuya edad no estaba registrada, asignamos un valor aleatorio dentro de un rango definido por la media de edad de su grupo más o menos una desviación estándar. Esta técnica nos permitió imputar las edades de manera que se preservara la distribución original y se redujera el sesgo en la predicción del modelo.

- Por otra parte, hipotetizamos que la cantidad de acompañantes influencia en la supervivencia de los sujetos. Tal como se muestra en la Fig 10 las personas que iban en grupos pequeños o solos tenían más probabilidad de sobrevivir que aquellos pertenecientes a un grupo mediano o grande. Se comprobó la hipótesis que se tenía, que la cantidad de acompañantes era influyente en la supervivencia del pasajero tal como se puede observar en Fig 10. Por lo que se tomó la información del feature de SibSp que indicaba el número de hermanos o esposos. También se tomó la información de Parch que indica el número de padres o hijos. Primero que nada se calculó la cantidad de personas en la familia de cada pasajero, para lo que se utilizó la siguiente fórmula.

$$FamilyMembers = sibsp + parch + 1 \quad (1)$$

A esto se le suma uno para tomar en cuenta al pasajero en la cuenta de miembros de familia. Después se separó por tamaño de familia, donde una familia pequeña tiene de 2-4 miembros, una mediana 5 o 6 y una grande 7-11. Los grupos que contienen un solo miembro se nombraron como solo.

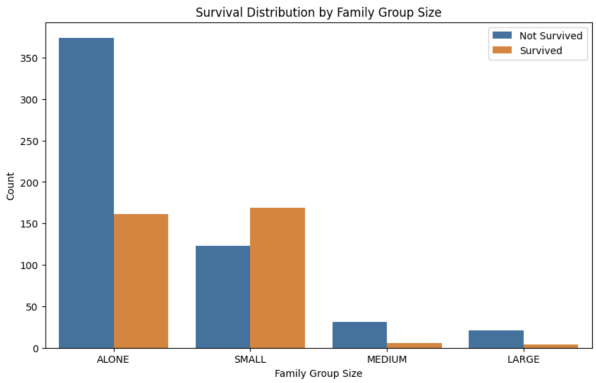

Figure 10. Probabilidad de supervivencia dependiendo acompañantes

- Otra variable que se destacó como relevante durante la exploración de datos fue "Sex". La relación entre "Sex" y "Age" resultó especialmente interesante, ya que al agrupar estas dos variables, obtuvimos información valiosa. Esto se hizo evidente cuando calculamos la probabil-

idad de supervivencia, que variaba significativamente para los grupos formados por mujeres de 0 a 5 años, hombres de 0 a 5 años, y así sucesivamente en intervalos de 5 años. En la Fig 11 se nota como cada grupo tiene distintas probabilidades de supervivencia. A partir de esta observación, generamos una nueva columna denominada "AgeGroupSex", donde combinamos las columnas "Age" y "Sex" para segmentar aún más a la población.

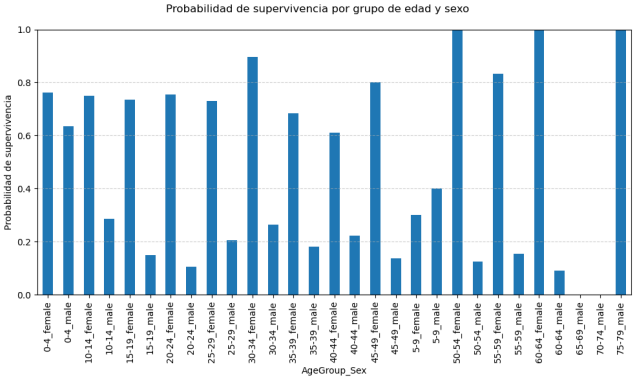

Figure 11. Probabilidad de supervivencia por grupo de edad y sexo

- Para poder obtener más información de los pasajeros, se utilizo la variable "Ticket", según un proyecto en Kaggle [2] al descomponer en tres partes el "Ticket" se puede obtener información valiosa de la clase del pasajero a partir del primer dígito del boleto. En la tabla III se observa como se distribuyen los dígitos con la clase de los pasajeros.

- Finalmente, se realizó One Hot Encoding. La técnica de One Hot Encoding se aplicó a la variable de sex, embarked y group size. Esta técnica se aplica cuando se quiere separar una columna categórica en varias columnas de tipo binario, una por cada categoría, con el objetivo de permitir que el modelo sea más eficaz al valorar variables categóricas. Por este motivo, se agregaron 10 features, 2 correspondientes a sex, 3 a embarked y 5 de group size. Ya así, se eliminaron las columnas de sex y embarked, pues el modelo se hará con las 5 columnas binarias agregadas.

Pclass	Ticket_FD	0	1	2	3	4	5	6	7	8	9	All
1	0	190	4	14	0	4	2	0	0	0	214	
2	0	22	136	23	0	1	0	2	0	0	184	
3	4	17	90	330	15	4	12	13	3	3	491	
All	4	229	230	367	15	9	14	15	3	3	889	

Table III. Distribución de los primeros dígitos de los boletos por clase.

E. Estructura final

Posterior a la limpieza y transformación que se le realizaron a los datos, es posible resumir que quedaron 10 variables con 889 y 417 pasajeros en el archivo de entrenamiento y de prueba, respectivamente. Las columnas que se utilizarán para realizar el modelo se encuentran en la tabla IV.

Variable	Descripción	Tipo
PassengerID	ID de pasajero	Identificador
Survived	Sobrevivió o no	Binario
Pclass	Clase de ticket	Categórico
Age	Edad de pasajero	Numérico
Fare	Tarifa de ticket	Numérico
Alone	Tamaño del grupo	Binario
Small	Tamaño del grupo	Binario
Medium	Tamaño del grupo	Binario
Large	Tamaño del grupo	Binario
Embarked_Cherbourg	Puerto	Binario
Embarked_Queenstown	Puerto	Binario
Embarked_Southampton	Puerto	Binario
Female	Sexo	Binario
Male	Sexo	Binario

Table IV. Descripción y tipo de dato de las variables tras la limpieza y transformación de datos.

Como ejemplo, aquí los datos del primer pasajero que aparece en la tabla de entrenamiento.

Se convervó la variable de passengerID para su uso como identificador. Las demás variables fueron conservadas debido a su relación con la sobrevivencia, determinada por gráficas comparando cada variable contra el survival rate. Aquellas variables que presentan cam-

Variable	Valor
PassengerID	1
Survived	0
Pclass	3
Age	22
Fare	7.25
Group_Size	3
Embarked_Cherbourg	0
Embarked_Queenstown	0
Embarked_Southampton	1
Female	0
Male	1

Table V. Valores del primer pasajero tras la limpieza y transformación de datos.

bios importantes o grupos definidos en cuanto al nivel sobrevivencia seran tomadas en cuenta para la elaboración del modelo. A continuación se describe e ilustra las relaciones importantes entre ciertas variables conservadas y su survival rate.

Tal como se puede observar en la Fig. 12 el caso de la variable Pclass, claramente los pasajeros que pertenecían a la clase 1, tenían más probabilidad de sobrevivir.

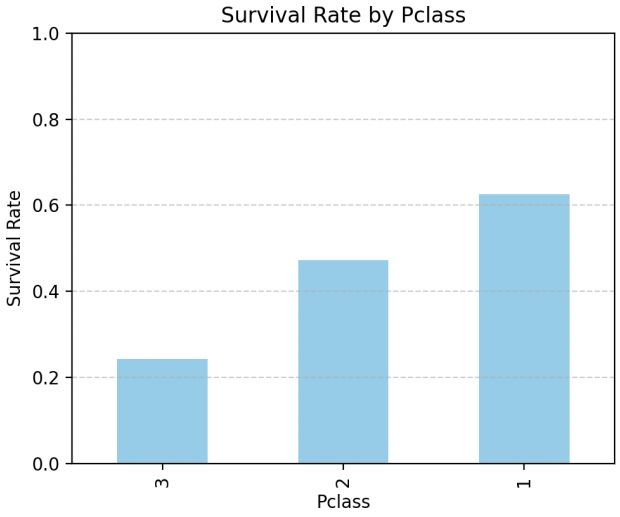

Figure 12. Probabilidad de supervivencia por clase

Por otro lado, se conservó la variable de Age ya que existe una relación entre la edad y la sobrevivencia. Esto se observa específicamente al compararlo también con el sexo. Como se observa en la Fig. 11, la probabilidad de sobrevivencia cambia dependiendo del sexo y edad de la persona.

También se determinó la importancia en el puerto desde el que la persona embarcó. Teniendo una mayor probabilidad de supervivencia si el pasajero embarcó desde el puerto C, Cherbourg como se observa en la Fig. 13.

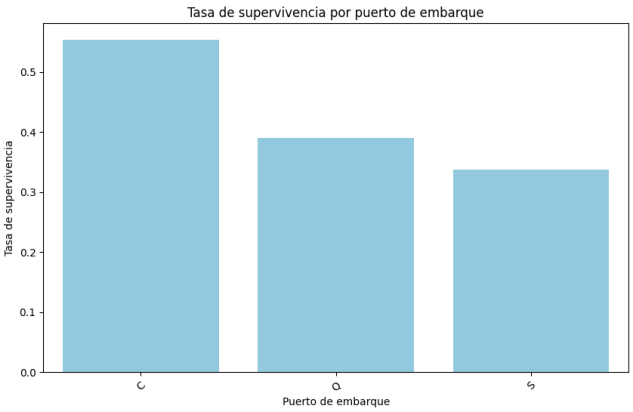

Figure 13. Probabilidad de supervivencia por puerto

En cuanto a la variable de Fare, observamos en la Fig. 14 que la probabilidad de sobrevivencia es variable. Sin embargo, se observa que en su mayoría las personas que pagaron un precio alto por su boleto sobrevivieron con mayor probabilidad que aquellos que pagaron un precio bajo. Por ejemplo al comparar las personas que pagaron entre 0 y 7 dólares por su boleto casi no sobrevivieron, mientras que las personas que pagaron entre 77 y 90 tenían una probabilidad de sobrevivencia del 80

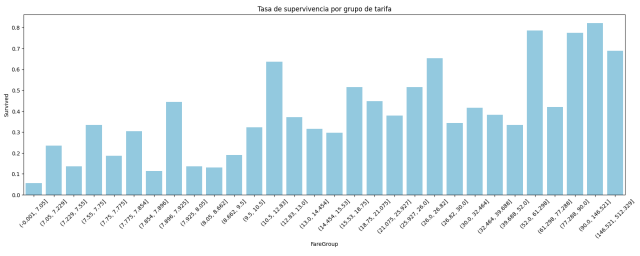

Figure 14. Probabilidad de supervivencia por fare

Por último, tal como se observó en la Fig 10 la relación entre la cantidad de acompañantes y la probabilidad de sobrevivencia las personas que viajaban solas o en grupos pequeños sobrevivieron con mayor proba-

bilidad que aquellas que iban en un grupo mediano o grande.

IV. SELECCIÓN, CONFIGURACIÓN Y ENTRENAMIENTO DEL MODELO

A. Selección

Para abordar la tarea de predicción de la supervivencia en el Titanic, seleccionamos tres modelos: regresión logística, Random Forest y una red neuronal de tipo feedforward. La elección de la regresión logística se justifica por su adecuación a problemas de clasificación binaria, permitiendo además interpretar el impacto de las variables predictoras en la probabilidad de supervivencia. Optamos por el modelo Random Forest debido a su capacidad como ensamble, lo que permite identificar y priorizar las variables con mayor influencia en la predicción, además de su robustez frente al sobreajuste y su eficacia al manejar variables categóricas. La red neuronal feedforward fue seleccionada por su capacidad de modelar relaciones no lineales complejas y detectar patrones más sutiles en los datos. Sin embargo, este modelo requiere una cantidad significativa de datos y ajuste fino de hiperparámetros para optimizar su rendimiento. Dado que la mayoría de las variables en nuestro conjunto de datos son categóricas, consideramos que estos tres modelos son apropiados para este análisis. Posteriormente, en la sección IV B describimos la configuración de los modelos. Para la selección del modelo, el conjunto de datos de entrenamiento se dividió en dos subconjuntos: uno para entrenamiento y otro para validación. Esta división permitió realizar pruebas sobre los modelos y evaluar su desempeño en el conjunto de validación. De esta manera, se pudo tomar una decisión informada con base en la capacidad predictiva de los modelos en datos no vistos durante el entrenamiento.

B. Configuración

1. *Regresión Logística*

Seleccionamos un modelo de regresión logística para la clasificación de la supervivencia en el caso del Titanic debido a varias razones clave. En primer lugar, la regresión logística es particularmente eficaz para problemas de clasificación binaria, como el nuestro, en el que buscamos predecir si un pasajero sobrevivió o no. Este modelo nos permite estimar la probabilidad de supervivencia de un pasajero en función de una o más variables independientes, lo que lo hace ideal para predecir resultados categóricos como "vivo" o "muerto".

Además, la regresión logística es fácil de interpretar, ya que produce probabilidades que podemos convertir en clases binarias. Esto nos ayuda a comprender cómo los cambios en las características de los pasajeros, como la clase, el sexo o la edad, afectan su probabilidad de supervivencia, lo que facilita la interpretación y comunicación de nuestros hallazgos.

También optamos por la regresión logística porque no requiere que las variables independientes tengan una relación lineal con la variable dependiente, lo que es útil en nuestro caso, ya que las relaciones entre las características de los pasajeros y su probabilidad de supervivencia no necesariamente son lineales.

En el entrenamiento de este modelo se decidió emplear la herramienta de GridSearch para evaluar los mejores hiperparámetros con los cuales teníamos la mejor precisión y exactitud en las predicciones, las cuales fueron las siguientes: VI.

Hiperparámetros	**Valores**
C	0.001, 0.01, 0.1, 1, 10
penalty	l1, l2

Table VI. Hiperparámetros y sus valores para el modelo de Regresión Logística

Después de analizar los resultados del *Grid Search*, se encontró que los mejores hiperparámetros y sus valores óptimos fueron los siguientes: VII.

Hiperparámetros	**Valores**
C	10
penalty	l1

Table VII. Hiperparámetros y sus valores para el modelo de Regresión Logística

2. *Random Forest*

Para el modelo de Random Forest, se utilizó la configuración predeterminada proporcionada por la librería sklearn [3]. Esta decisión se tomó con el fin de evaluar el modelo en su forma más básica, permitiendo así considerar un refinamiento posterior en caso de ser necesario. De esta manera, se podrá determinar si un ajuste adicional es conveniente al compararlo con los demás modelos evaluados. Los parámetros utilizados se presentan en la Tabla VIII.

Hiperparámetros	**Valores**
Número de estimadores	100
Profundidad máxima	None
Num. min. de muestras en la hoja	1
Num. min. de muestras para dividir nodo	2

Table VIII. Hiperparámetros para el modelo Random Forest

3. *Red Neuronal Feed Forward*

Para el modelo de red neuronal, se utilizó la configuración predeterminada proporcionada por la librería Keras [4]. La arquitectura de la red consiste en una capa de entrada con 16 neuronas, seguida de una segunda capa oculta también con 16 neuronas, ambas utilizando la función de activación ReLU. La capa de salida contiene una neurona con función de activación sigmoide, adecuada para la clasificación binaria. Esta configuración se implementó con el fin de evaluar el modelo en su forma más básica, permitiendo considerar un refinamiento posterior si fuera necesario. De este modo, se podrá determinar si un ajuste adicional es conveniente al compararlo con otros modelos evaluados. Los parámetros utilizados se presentan en la Tabla IX.

Hiperparámetros	**Valores**
Tasa de Aprendizaje	0.01
Optimizador	adam
Tamaño del Batch	32
Número de Epochs	50

Table IX. Hiperparámetros y sus valores para la red neuronal feed forward

C. Entrenamiento

Ahora que tenemos las configuraciones de nuestros tres modelos, se procedió a entrenarlos con el conjunto de entrenamiento y probarlos en el conjunto de validación para determinar cuál es el más adecuado para nuestro problema. Los resultados de la predicción sobre los datos de validación se compararon en la Tabla X.

En dicha tabla, se utilizan los siguientes términos para describir los resultados de las predicciones:

- **TP (True Positive)**: Casos donde el modelo predice correctamente la clase positiva.
- **TN (True Negative)**: Casos donde el modelo predice correctamente la clase negativa.
- **FP (False Positive)**: Casos donde el modelo predice la clase positiva incorrectamente, cuando en realidad es negativa.
- **FN (False Negative)**: Casos donde el modelo predice la clase negativa incorrectamente, cuando en realidad es positiva.

A partir de estos valores, se pueden calcular las métricas de evaluación: *accuracy*, *precision* y *recall*. El *accuracy* es la proporción de predicciones correctas (positivas y negativas) sobre el total de predicciones realizadas. Se calcula con la siguiente fórmula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

La precisión es la proporción de predicciones correctas de la clase positiva sobre el total de predicciones que fueron clasificadas como positivas. Se calcula como:

$$\text{Precision} = \frac{TP}{TP + FP}$$

El *recall* es la proporción de verdaderos positivos detectados sobre el total de casos que son realmente positivos. Su fórmula es:

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

Estas métricas nos permiten evaluar los modelos en diferentes aspectos, como su capacidad para identificar correctamente las clases positivas, su habilidad para minimizar falsos positivos, y su capacidad general para predecir correctamente.

Modelo	TP	TN	FP	FN	*Accuracy*	Precisión	Recall
Reg. Logística	60	84	25	9	0.81	0.71	0.87
Random Forest	49	91	18	20	0.78	0.73	0.71
Red Neuronal	38	97	12	31	0.75	0.76	0.55

Table X. Comparación de métricas de diferentes modelos

D. Refinamiento

graphicx

Para el refinamiento del modelo, se construyó un árbol de decisión para analizar la influencia de las características en las predicciones. El análisis del árbol (ver Figura 15) reveló que dos nodos principales eran cruciales: el primero se basaba en si el pasajero era hombre y tenía una edad menor a 6.5 años, y el segundo en si era mujer y su edad era menor a 2.5 años. Estos nodos indicaron que el género y la edad eran factores determinantes importantes para la clasificación. La identificación de estos nodos clave no solo proporcionó una visión clara de cómo las características individuales afectan la predicción, sino que también permitió ajustar y optimizar el modelo para mejorar su rendimiento y evitar el sobreajuste. Como resultado, se generaron nuevas columnas binarias: una para hombres menores de 6.5 años y otra para mujeres menores de 2.5 años.

Por otra parte, se observó que la variable *pClass* también era influyente en el árbol de decisión (ver Figura 15). Sin embargo, esta variable podía introducir errores debido a su naturaleza ordinal, ya que representaba la clase del pasajero con valores del 1 al 3. Para evitar la multiplicidad y mejorar la representación de esta característica, se aplicó *One-Hot Encoding*, creando así tres columnas binarias, una para cada clase. Tras implementar estos cambios, se observó un aumento del 2% en la precisión, específicamente en el modelo de regresión logística.

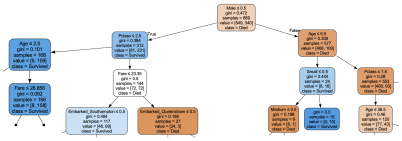

Figure 15. Árbol de decisión mostrando los nodos principales

E. Modelo final

Finalmente, en base a la información recopilada sobre el desempeño de cada uno de los modelos demostrado en en la Tabla X. Se pudo observa que regresión logística tiene la mayor exactitud (0.90), lo que indica que clasifica correctamente el 90% de los casos. Además, la precisión del modelo (0.81) es la más alta de los tres, lo que significa que de las predicciones positivas que hace, el 81% son correctas.

Por otra parte, el recall, que mide la capacidad del modelo para identificar correctamente los verdaderos positivos, es el punto más fuerte de la regresión logística (0.97). Esto significa que de todas las instancias positivas, el 97% son correctamente identificadas, un valor mucho mayor que los obtenidos por Random Forest (0.71) y la Red Neuronal (0.55). La regresión logística tiene solo 5 falsos negativos, lo que indica que rara vez deja de identificar una clase positiva correctamente.

En resumen, la regresión logística combina una alta precisión, recall y exactitud, lo que la convierte en el modelo más robusto y confiable para este conjunto de datos en comparación con Random Forest y la Red Neuronal, que muestran un peor desempeño en métricas clave. Además de estos buenos resultados obtenidos, la regresión logística fue el modelo escogido también por su simplicidad y fácil comprensión. Como se menciona en la literatura, “En la selección de modelos, es aconsejable optar por el modelo más simple que proporcione un rendimiento satisfactorio. Esto sigue el principio de parsimonia o ’navaja de Occam’, que sugiere que las explicaciones más simples suelen ser las mejores” [5]. Esto resalta la importancia de no solo considerar el rendimiento del modelo, sino también su complejidad y facilidad de interpretación. Es por ello que el modelo finalmente seleccionado es la regresión logística.

V. CONCLUSIÓN

En este proyecto se realizó un análisis exhaustivo sobre la predicción de supervivencia de los pasajeros del Titanic utilizando técnicas de machine learning, con énfasis en la limpieza y transformación de los datos, lo que incluyó la imputación de datos faltantes y la creación de nuevas variables para mejorar el rendimiento predictivo.

Se implementaron y compararon tres modelos: Regresión Logística, Random Forest y una Red Neuronal Feedforward. Los resultados obtenidos mostraron que el modelo de Regresión Logística alcanzó el mejor desempeño con una precisión del 81

Este análisis permitió identificar variables clave que influyen en la supervivencia, como la clase del pasajero, el sexo, la edad y el tamaño del grupo con el que viajaba. El éxito del modelo de Regresión Logística se atribuye a su capacidad para manejar problemas de clasificación binaria de manera eficiente, mientras que los otros modelos también demostraron ser útiles, aunque con menor precisión; de igual forma se prefirió seleccionar este modelo por la razón de ser un modelo más simple y menos pesado computacionalmente hablando.

En conclusión, este enfoque de machine learning demostró ser efectivo para predecir la supervivencia de los pasajeros del Titanic, proporcionando un marco útil para el análisis de datos históricos y su aplicación en la gestión de riesgos.

[1] Will Cukierski, *Titanic - Machine Learning from Disaster* (Kaggle, 2012)

[2] Oscar Takeshita, *Titanic Ticket-only study* (Kaggle, 2018)

[3] *RandomForestClassifier* Scikit-learn

[4] *tf.keras.Sequential* Tensorflow Keras

[5] *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009)

Appendix: Appendix

Acceso al repositorio de Github con el código realizado. Aquí