Escuela de Ingeniería y Ciencias, Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (TC3006C.102)

Momento de Retroalimentación: Reto Limpieza del Conjunto de Datos

Equipo 4:

Karla Andrea Palma Villanueva (A01754270)

Viviana Alanis Fraige (A01236316)

David Fernando Armendariz Torres (A01570813)

Alan Alberto Mota Yescas (A01753924)

Adrián Chávez Morales (A01568679)

Jose Manuel Armendáriz Mena (A01197583)

Docentes:

Alfredo Esquivel Jaramillo

Mauricio Gonzalez Soto

Frumencio Olivas Alvarez

Antonio Carlos Bento

Hugo Terashima Marín

Monterrey, Nuevo León, México. 31 de agosto de 2024

Índice

1. Abstract

Este documento presenta un análisis exhaustivo del conjunto de datos del Titanic con el objetivo de desarrollar un modelo de aprendizaje automático capaz de predecir la supervivencia de los pasajeros. Se llevó a cabo un proceso meticuloso de limpieza y preprocesamiento de datos, que incluyó la imputación de valores ausentes y la eliminación de variables irrelevantes. Se exploraron diversas características del conjunto de datos, como la clase social, el género y la edad, y se seleccionaron variables clave para el modelado. Se probaron tres modelos de aprendizaje automático: Support Vector Machine (SVM), Random Forest y Multilayer Perceptron (MLP), evaluando su rendimiento en función de métricas como precisión, recall y F1-score. Finalmente, se eligió el modelo Random Forest por su robustez y capacidad de manejar características irrelevantes, logrando un rendimiento óptimo en la clasificación de supervivientes.

2. Introducción

La inteligencia artifical se ha convertido en una herramienta fundamental en la industria, debido a su alto impacto en la toma de decisiones, principalmente en el ánalisis de datos. Entre sus ramas, el Aprendizaje Automático, también conocido como Machine Learning, destaca por su capacidad para mejorar continuamente su desempeño al aprender de los datos, en lugar de depender únicamente de la programación explícita, esta disciplina se basa en la aplicación de modelos estadísticos que buscan hacer predicciones precisas a partir de datos históricos.

El Aprendizaje Automático puede abordarse de diferentes maneras, y uno de sus enfoques más comunes es el Aprendizaje Supervisado. Se sabe que para realizar las predicciones se requieren de datos de entrada y de salida, por lo que para este tipo de modelos se requieren de datos etiquetados que sirven para el entrenamiento del mismo y así cumplir con su función de ejecutar las predicciones sobre los nuevos datos.

En este sentido, uno de los desafíos más comunes en la comunidad de las ciencia de datos es el "Titanic - Machine Learning from Disaster". Dicho desafío además de aplicar las técnicas de aprendizaje automático, posee información realista mediante un análisis de datos

del Titanic y permite a los algoritmos aprender de datos históricos para hacer predicciones sobre eventos futuros.

3. Objetivo

El objetivo planteado corresponde a la creación de un modelo de aprendizaje automático con la capacidad de predecir la supervivencia de un individuo ante la tragedia del hundimiento del Titanic. Se busca generar un modelo de buen desempeño dada una entrada de caracterísitcas de un pasajero hipotetico.

Para cumplir este objetivo, se entrenarán multiples modelos de aprendizaje supervisado para clasificar el estatus del pasajero (sobrevivió, pereció). Los modelos serán entrenados mediante un set de datos de características de pasajeros pasados como su clase social, edad, sexo y tarifa, entre otras cosas.

Ante este objetivo, primeramente se deben preparar los datos, este es el objetivo puntual de esta primera etapa. Algo sumamente importante para poder aplicar cualquier tipo de modelo son las actividades de limpieza y organización de datos, en las cuales, se trata de limpiar y suavizar cualquier inconsistencia que impida que el conjunto de datos sea coherente. El proceso de limpieza que se llevará a cabo se compone de manejo de datos nulos, caracteres problemáticos, inconsitencias en los datos, etcétera.

Este proceso es muy importante porque establecerá las bases para un análisis más extenso, permitiendo que los modelos de aprendizaje automático trabajen con un dataset ordenado y correctamente estructurado.de tal forma que el modelo sea más preciso y su rendimiento de resultados sea eficiente y viable.

4. Limpieza del conjunto de datos

4.1. Estructura de Datos

Los datos de los pasajeros del Titanic están divididos en dos conjuntos distintos: un conjunto de entrenamiento y uno de prueba. El conjunto de entrenamiento se utiliza para ajustar el modelo de aprendizaje, mientras que el conjunto de prueba se emplea para evaluar

la precisión del modelo con datos desconocidos. El conjunto de entrenamiento contiene 891 instancias, mientras que el de prueba tiene 418. A continuación se presenta un diccionario de datos que describe las variables o características de ambos conjuntos, provenientes de la misma fuente en Kaggle (Cukierski, 2012).

Variable	**Tipo**	**Especificaciones**	**Descripción**
Survival	Categórica	0 = No, 1 = Sí	Supervivencia del pasajero
pclass	Categórica	1 = 1ª, 2 = 2ª y 3 = 3ª	Estatus socioeconómico
Name	Categórica	-	Nombre y título del pasajero
sex	Categórica	-	Sexo del pasajero
Age	Numérica	La edad puede ser fraccionaria si es menor de 1 año.	Edad del pasajero
sibsp	Numérica	-	Número de hermanos o cónyuges a bordo
parch	Numérica	Si los niños viajaron solo con una niñera el valor es cero en esos casos.	Número de padres o hijos a bordo
ticket	Categórica	-	Número de ticket del pasajero
fare	Numérica	-	Costo del ticket
cabin	Categórica	-	Número de cabina
embarked	Categórica	C=Cherbourg, Q=Queenstown, S=Southampton	Puerto de embarque

Cuadro 1: Diccionario de datos antes de limpieza

PassengerId	**Survived**	**pclass**	**Name**	**sex**	**Age**	**sibsp**	**parch**	**ticket**	**fare**	**cabin**	**embarked**
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Cuadro 2: Datos en crudo

4.2. Exploración de Datos

En esta sección, se hará una exploración de cada variable del dataset para definir su relevancia para la predicción del modelo.

- **PassengerId:** Debido a su naturaleza como identificador, no es relevante para determinar la supervivencia del pasajero. Esta variable no se incluirá en el dataset de entrenameniento, en cambio se utilizará como indice.

- **Survived:** Esta variable es el 'target', será aislada en otro dataset. Tiene 549 registros con '0' (difunto) y 342 con '1' (sobreviviente).

- **Pclass:** Esta variable describe el estatus socioeconómico del pasajero. Tiene 216 registros con '1', 184 con '2', y 491 con '3'. Esto se relaciona con la calidad de la estancia (comodidad y seguridad) de los pasajeros abordo del titanic, por lo que se considera releveante para el modelo.

- **Name:** Además de los nombres de los pasajeros, estos vienen acompañados con títulos representando su ocupación o estado civil. Entre estos títulos están: Mr, Miss, Mrs, Master, Dr, Rev, Mlle, Major, Col, The Countess, Capt, Ms Sir, Lady, Mme, Don, Jonkheer. Mientras la variable 'Name' se va a eliminar del dataset, cierta información relevante se mantendrá por medio de la agrupación de estos títulos y una una codificación one-hot encoding.

- **sex:** El género del pasajero es relevante para futuros analisis sobre hombres y mujeres que abordaron. Tiene 577 registros con 'male' y 314 con 'female'. Historicamente se sabe que las mujeres tuvieron una mayor taza de supervivencia ante la tragedia.

- **Age:** Esta variable es la que presenta más datos ausentes (177 registros son NaN) pero también es de las que son relevantes para el modelo. Para evitar reducira la cantidad de datos de entrenamiento se imputaran los datos ausentes, las tecnicas utilizadas para dicho proceso se presentan mas adelante.

- **Sibsp:** El número de hermanos del pasajero no se considera significativo para la predicción del modelo. Esta variable se descarta.

- **Parch:** Esta variable se consdiera de utilidad para el modelo, ya que indica el número de padres y/o hijos abordo.

- **ticket:** El número de ticket no es relevante para la predicción del modelo. No proporciona información como ubicación del pasajero y solo sirve como una etiqueta. Esta variable se descarta.

- **fare:** El costo del ticket también se considera de poca importancia para la predicción del modelo. Al ser 'Pclass' un proxy del estatus socioeconómico del pasajero, existe una relación con 'fare' y por ende se espera la información relevante de esta variable sea contenida por 'Pclass'-

- **cabin:** 'Cabin' se encuentra en una situación similar a 'Fare', adicionalmente, cuenta con una exuberante cantidad de datos austenes por lo cual se descartará.

- **embarked:** El punto de embarcación del pasajero se descarta al no poseer una relación directa lógica con la supervivencia de los pasajeros.

La relación, y por ende relevancia, de las 4 variables originales, planteadas como las variables de entrenamiento del modelo, y la taza de superviencia se puede observar en la figura 1.

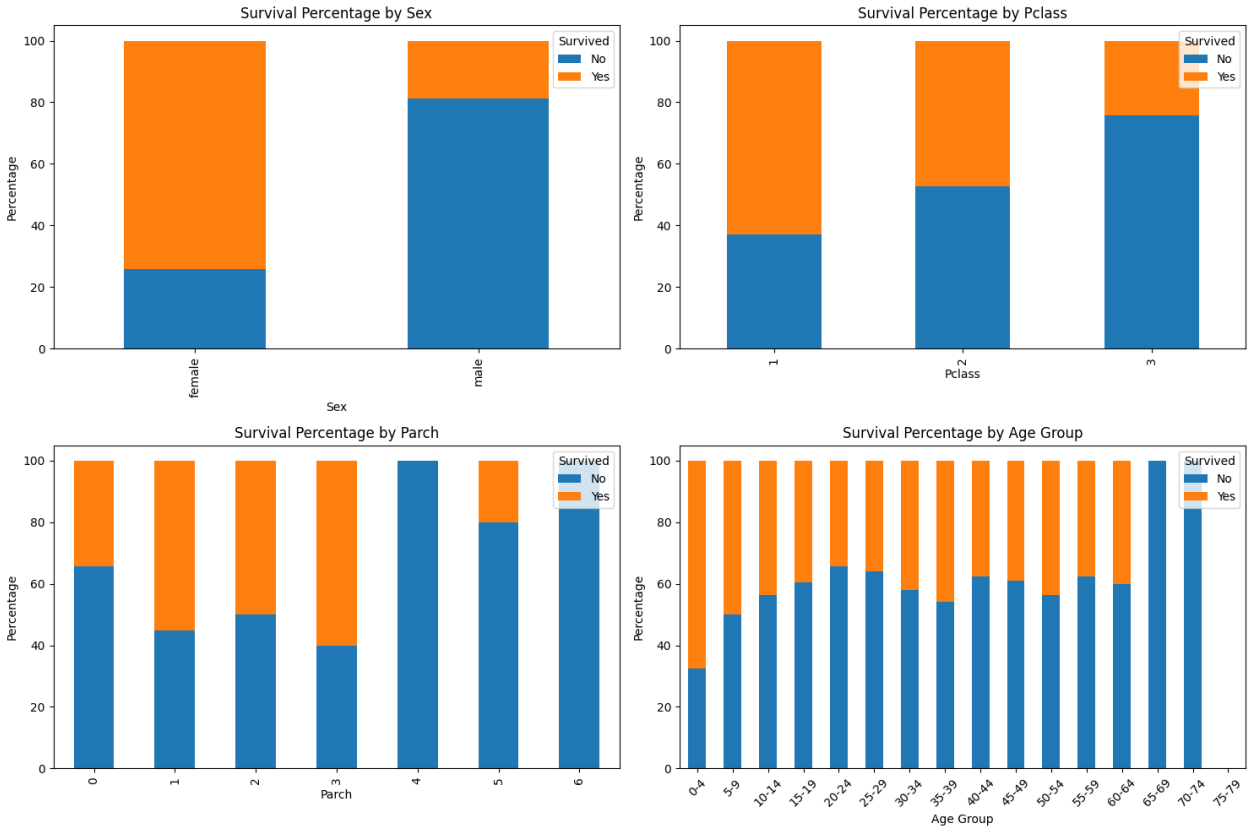

Figura 1: Tasa de supervivencia por variables (Sex, Pclass, Parch, Age)

4.3. Limpieza de datos

Primeramente se llevó a cabo un analisis exploratorio de los datos, mediante el cual se realizó un descarte preeliminar de las variables o features que aparentan no ser significativas para la predicción del modelo (sibsp, ticket, fare, cabin, embarked). Mas adelante se habla nuevamente sobre el motivo tras la decisión de trabajar sobre los otros features.

Una vez completado esto, se buscó lidíar con los valores ausentes del conjunto de datos de entrenamiento, en este caso, "Age" es la unica variable seleccionada que presenta datos ausentes. Para lidiar con ellos, se decidió generar un feature adicional "Titles", correspondiente a los titulos de cada pasajero extraidos de "Name". La principal intención de dicha generación es la de utilizar información generada de subgrupos de datos a base de los titulos para completar los valores faltantes de "Age". A continuación se muestran los títulos, Cuadro 3, y el proceso de manejo de nulos.

Titles	**Count**
Mr	517
Miss	182
Mrs	125
Master	40
Dr	7
Rev	6
Mlle	2
Major	2
Col	2
the Countess	1
Capt	1
Ms	1
Sir	1
Lady	1
Mme	1
Don	1
Jonkheer	1

Cuadro 3: Conteo de registros por título

Se identificaron los "Titles" con una plresencia de valores nulos en la variable de "Age". Los títulos que presentaron un número significativo de datos faltantes en la edad fueron: "Master", "Mr", "Miss" y "Mrs". De dichos títulos, se graficó la distribución de edades en gráficos de

barras y caja y bigote, como se muestran en las figuras 2, 3, 4, 5, 6, 7, 8 y 9.

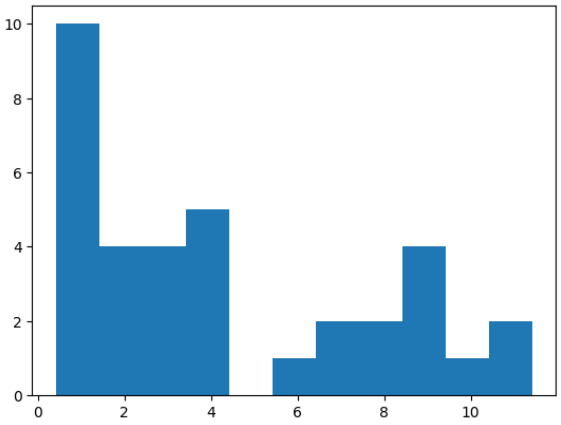

Figura 2: Distribución edades, título: Master

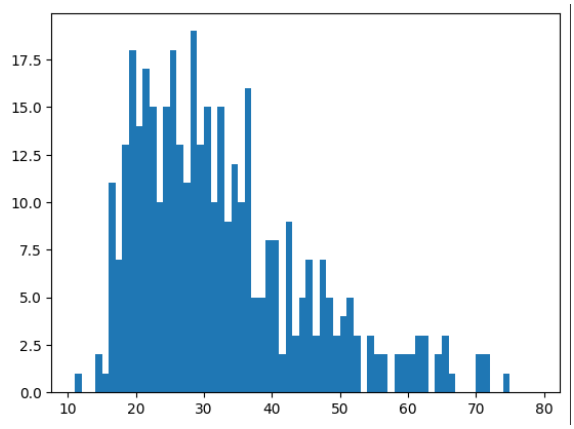

Figura 3: Distribución edades, título: Mr

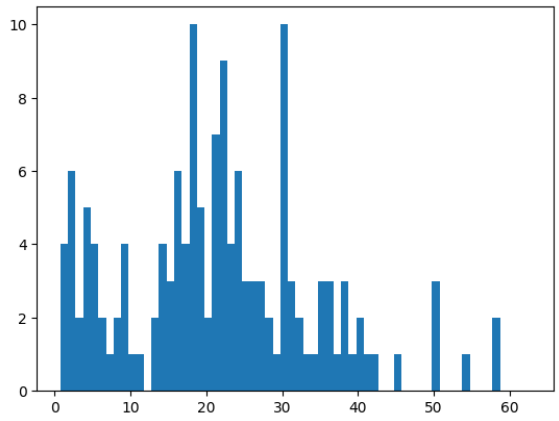

Figura 4: Distribución edades, título: Miss

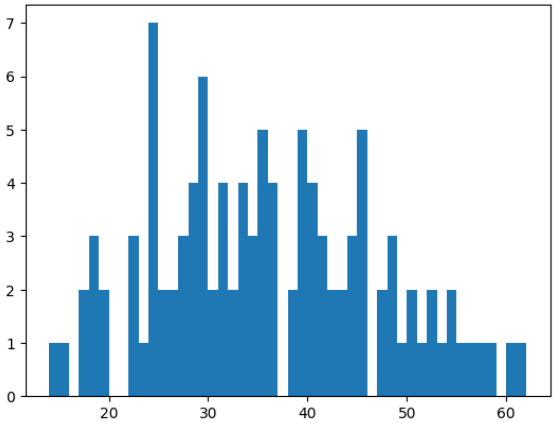

Figura 5: Distribución edades, título: Mrs

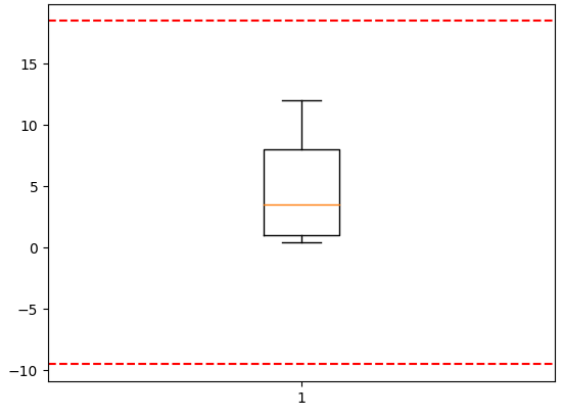

Figura 6: Boxplot edades, título: Master

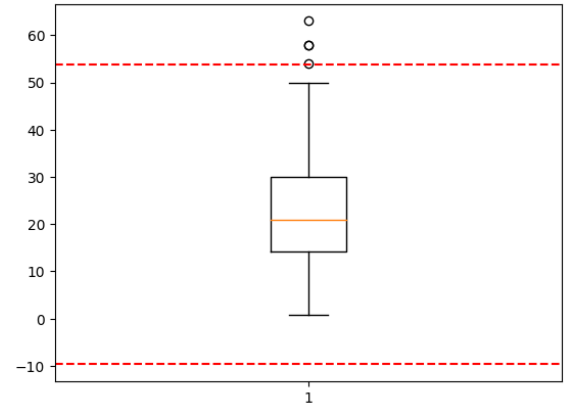

Figura 7: Boxplot edades, título: Mr

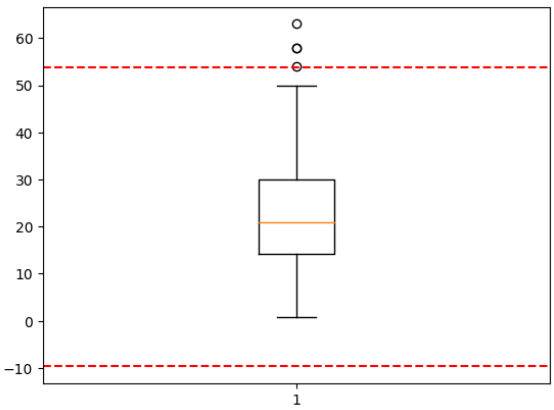

Figura 8: Boxplot edades, título: Miss

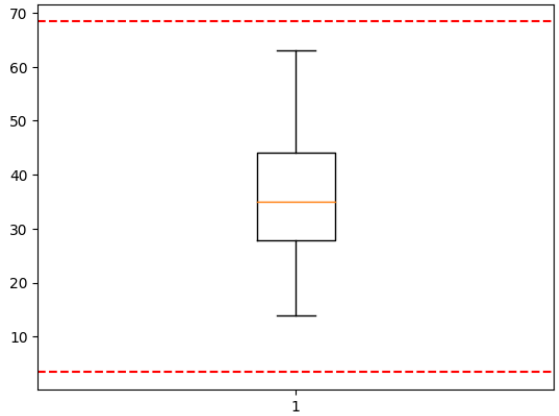

Figura 9: Boxplot edades, título: Mrs

Tras realizar la visualización de los datos correspondientes a cada título, se decidió rellenar los valores faltantes en la variable de edad utilizando el una medida de tencencia central de la misma asociado a cada título respectivo. Los registros nulos en edad de los titulos con presencia de edades atípicas (Mr, Miss) se remplazaron con la mediana de los datos respectivos al titulo, en cambio los titulos sin edades atípicas (Master, Mrs) se imputaron con la media de los datos. Estas tecnicas se conocen como imputación media incondicional y imputación mediana (Arteaga y Ferrer-Riquelme, 2009). Ejemplificando lo previamente planteado, el registro con “PassengerID": 20 cuenta con un registro nulo en “Age", dado que ese pasajero cuenta con el título de “Mrs", el valor de la edad se imputa con la media de las edades de “Mrs "que equivale a 35,90.

Para el caso especifico del titulo de Dr., solo hubo un dato nulo de edad presente, por lo que se optó buscar el nombre de la persona en una base de datos que contenía los nombres y las edades de algunos tripulantes. Una vez encontrada la edad, se procedió a añadirla en nuestra base de datos (Encyclopedia Titanica, 2016).

El objetivo de estas imputaciones fue el de no eliminar una cantidad significativa de datos de entrenamiento y de afectar lo mínimo posible la distribución inata de las edades por título.

4.4. Pre-procesamiento de datos

Para la variable “sex”, se realizó una codificación binaria, asignando el valor 0 a “male” y el valor 1 a “female”.

En el análisis de los títulos, se decidió agrupar aquellos que representaban la misma clase social en una sola categoría. Esto se hizo con el propósito de reducir la cantidad de datos únicos y facilitar el análisis, manteniendo asi la relevancia de la información sobre el estatus social de los pasajeros. La agrupación se realizó de la siguiente manera:

- **Mr**: Se mantuvo como una categoría individual.
- **Miss**: Agrupó títulos equivalentes como *Miss*, *Mlle* (Mademoiselle) y *Ms*.
- **Mrs**: Incluye únicamente a las pasajeras con el título *Mrs*.
- **Master**: Corresponde al título *Master*, utilizado para varones jóvenes.
- **Social Workers**: Agrupamos a pasajeros con títulos como *Dr* y *Rev* (Reverendo), aunque no sean trabajadores sociales en un sentido moderno.
- **Military**: Incluye a pasajeros con títulos militares como *Major*, *Col* (Coronel) y *Capt* (Capitán).
- **Nobility**: Los títulos de nobleza como *Sir*, *Lady*, *Mme* (Madame), *Don* y *Jonkheer* fueron agrupados en una sola categoría.

Después de realizar esta agrupación, se aplicó una codificación *one-hot*, lo cual generó columnas binarias para cada grupo de títulos. De esta forma, un valor de `True` indica que el pasajero pertenece a esa categoría, tal y como se ve en el cuadro 4. Finalmente, se eliminó la columna original de *Titles* para evitar redundancias y trabajar únicamente con las nuevas columnas generadas. Esto permitirá analizar los datos de manera más eficiente y estructurada.

PassengerId	**Pclass**	**Sex**	**Age**	**Parch**	**Mr**	**Miss**	**Mrs**	**Master**	**Social Worker**	**Military**	**Nobility**
1	3	0	22.0	0	True	False	False	False	False	False	False
2	1	1	38.0	0	False	False	True	False	False	False	False
3	3	1	26.0	0	False	True	False	False	False	False	False
4	1	1	35.0	0	False	False	True	False	False	False	False
5	3	0	35.0	0	True	False	False	False	False	False	False

Cuadro 4: Clasificación de títulos por clase social

4.5. Estructura final del dataset

Como resultado del análisis, se seleccionaron cinco características que se considera son clave para realizar un estudio profundo de las personas que viajaron y sobrevivieron en el Titanic. El dataset final contiene con 891 registros. Además, se centró el análisis en el objetivo principal: determinar si la persona sobrevivió o no. Las variables seleccionadas son las siguientes, el ultimo elemento de la lista corresponde a la variable objetivo:

- **Pclass**: Es relevante para saber la clase social o estatus económico del pasajero.
- **Sex**: El género del pasajero es relevante para futuros analisis sobre hombres y mujeres que abordaron.
- **Age**: La edad del pasajero, al igual que el sexo, es de importacia para analisis sobre los diferentes rangos de edades que abordaron y sobrevivieron.
- **Parch**: Indica el número de padres y/o hijos en el Titanic. De esta manera se asigna un valor y vemos si la persona viajo con familiares o, sin ellos.
- **Survived**: Es el indicador de supervivencia y la variable objetivo. Hay 549 registros con "1", y 342 registros con "0".
- **Mr**: Se refiere a hombres adultos que no tienen un título nobiliario.
- **Miss**: Se refiere a mujeres jóvenes o solteras.
- **Mrs**: Se refiere a mujeres casadas o viudas.
- **Master**: Se refiere a niños varones.
- **Social Worker**: Este título se utilizaría para referirse a cualquier pasajero que trabajara en el ámbito de servicios sociales.
- **Military**: Se refiere a pasajeros con un título militar, como oficiales o soldados.
- **Nobility**: Se refiere a pasajeros con títulos nobiliarios, como “Sir” o “Lady”.

5. Selección, configuración y entrenamiento del modelo

5.1. Identificación del modelo

Para abordar el problema de predecir quién sobrevivirá al naufragio del Titanic, se implementarán tres modelos de Aprendizaje Automático: MLP, SVM y Random Forest. Se emplearon estos tres moedelos debido a su capacidad para generar clasificación binaria y sus variantes grados de complejidad asi como diversas fortalezas y debildidades. Esto con la intención de generar un variado repertorio de modelos de los cuales poder seleccionar el mejor.

5.2. Modelos a analizar

- **SVM**: Una Maquina de Vectores de Soporte, SVM por sus siglas en inglés, es un modelo de aprendizaje supervizado utilizado para la clasificación binaria. Este modelo clasifica datos generando un hiperplano de $n-1$ dimensiones en un espacio de n dimensiones, donde n es la cantidad de features empleados. El hiperplano generado segmenta el espacio optimizando el margen entre los datos mas cercanos de ambas clases, los datos empleados se denominan vectores de soporte. Dicha separación se utiliza para diferenciar entre ambas clases (IBM, s.f.).

  El modelo de clasificación busca generar una separación lineal entre ambas clases, en dado caso que los datos no sean linealmente separables se utilizar una función kernel para proyectar los datos. Esta proyecta los datos, sin transformarlos, en una dimensión superior en la cual sean linealmente separables (Wilimitis, 2019). Adicionalmente, en aquellos casos en los cuales los datos no sean perfectamente linealmente separables, el margen generado se clasifica como suave, lo que permite ciertos errores de clasificación lo que a su vez, en datos reales, hace de el modelo de clasificación más robusto ante datos atípicos.

- **Random Forest**: El Random Forest es un modelo de aprendizaje supervisado que combina la salida de múltiples árboles de decisión para alcanzar un solo resultado. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de

clasificación y regresión (IBM, s.f.).

Para su funcionamiento, este modelo elige una muestra aleatoria del conjunto de datos. Cada uno de estos se entrena de distinta manera para que el algoritmo combine todos los resultados posibles y así genere resultados basados en la votación. Para su procedimiento se inicia seleccionando aleatoriamente un número 'n"de registros del conjunto de datos original. A partir de cada muestra seleccionada, se construye un árbol de decisión que genera un resultado que se basa en los datos que le fueron asignados. Finalmente, para su clasificación, el resultado final se determina por la mayoría de votos de estos árboles generados (*Random forest, la gran técnica de Machine Learning*, 2023).

- **MLP**: Clasificador Perceptrón Multicapa es un modelo de aprendizaje supervisado que optimiza la función de descenso de gradiente estocástico, a través de redes neuronales. Se compone de múltiples capas de neuronas interconectadas, en las que las salidas de las neuronas de una capa se convierten en entradas para la siguiente capa. La primera capa se llama capa de entrada, la última capa se llama capa de salida y las capas intermedias se llaman capas ocultas, en las cuales cada conexión tiene un peso asociado que se ajusta durante el proceso de entrenamiento (*Qué es Perceptrón Multicapa - MLP | Concepto y definición. Glosario*, s.f.).

  Este modelo es uno de los más utilizados de redes neuronales, debido a su capacidad para modelar tareas de clasificación y regresiones no lineales; además la arquitectura que se implementa, permite que el modelo aprenda representaciones complejas de los datos.

  La elección principal de este modelo se realizó, puesto que es apto para el manejo de relaciones no lineales entre variables complejas como las que posee el Titanic, como ejemplo, se tiene la edad, el género, la clase de boleto, entre otras. Dichas caracterísiticas pueden influir en la probabilidad de supervivencia y por consiguiente el modelo se encargará de capturar estas interacciones y patrones en los datos, permitiendo una predicción más precisa entre los pasajeros que sobrevivieron y los que no.

5.3. Configuración y entrenamiento

Configuren y entrenen el o los modelos y observen su desempeño.

- **SVM**:

Kernel	Accuracy	F1 Score
`Linear`	0.80	0.80
`Poly`	0.64	0.53
`RBF`	0.61	0.48
`Sigmoid`	0.56	0.54

Cuadro 5: SVM Model Performance con diferentes Kernels.

En el cuadro 5, se muestran los hiperparámetros utilizados para la función SVM, para el cual únicamente varió el kernel, junto con los resultados de precisión del modelo. Como se mencionó anteriormente, se muestra el tipo de kernel utilizado, el cual se selecciona entre las diferentes opciones posibles, y se utiliza un margen blando. Es importante mencionar que F1-Score es respecto al promedio macro del modelo, teniendo los demás hiperparámetros sin variación, y se respetan los valores predeterminados establecidos por la función (*SVC-Scikitlearn*, 2024).

- **Random Forest**:

  - Número de estimadores: 100
  - Criterio de separación: varía
  - Máxima profundidad: 50
  - El resto de los parametros corresponden a aquellos por defecto del RandomForestClassifier de sklearn, de igual manera como en SVM el F1-Score es respecto al promedio macro del modelo (Scikitlearn, 2024).

Criterion	Precisión	F1 Score
`Gini`	0.79	0.77
`Entropy`	0.82	0.80
`log loss`	0.82	0.80

Cuadro 6: Desempeño de Random Forest variando criterio.

- **MLP**:
  - F1-Score:
    - Macro promedio: 0.81
    - Promedio ponderado: 0.81
  - Accuracy: 0.82

5.4. Comparación de Modelos

Para evaluar los resultados de los modelos, se seleccionaron las siguientes métricas:

- **F1-Score:** Ayuda a encontrar un equilibrio óptimo entre identificar sobrevivientes correctamente (recall) y evitar predecir falsamente demasiados (precisión).
- **Exactitud (Accuracy):** Mide qué tan bien el modelo clasifica tanto los positivos como los negativos. Puede ser engañosa si las clases están desequilibradas. Este no sería útil, puesto que se da un peso a las variables, más la fuente "Kaggle" trabaja con el resultado de exactitud, así que es considerado para verificar la exactitud del modelo obtenido.

Los modelos también muestran los resultados de las métricas "Precisión", "Recall" y "Soporte". Sin embargo, estas no nos son relevantes en el contexto actual de los datos de sobrevivientes del Titanic, ya que no tendrían un impacto significativo en la actualidad.

Utilizando los valores obtenidos por el entrenamiento de cada modelo, se estarán comparando en el Cuadro 7:

Modelo	**F1-Score**	**Exactitud (Accuracy)**
MLP	82 %	81 %
SVM	80 %	80 %
Random Forest	84 %	84 %

Cuadro 7: Comparación de Modelos

6. Conclusión

El análisis y modelado del conjunto de datos del Titanic permitió identificar las variables más influyentes en la supervivencia de los pasajeros y desarrollar un modelo predictivo eficaz. El uso de Random Forest demostró ser la mejor opción debido a su capacidad para evitar el sobreajuste y manejar la complejidad del conjunto de datos. La elección del número de árboles se optimizó para alcanzar una estabilidad en el rendimiento del modelo. En conclusión, este trabajo no solo proporciona una visión sobre los factores que influyeron en la supervivencia durante el naufragio, sino que también establece un proceso robusto para la limpieza y preparación de datos en futuros análisis de aprendizaje automático.

Referencias

Arteaga, F., y Ferrer-Riquelme, A. (2009). 3.06 - missing data. En S. D. Brown, R. Tauler, y B. Walczak (Eds.), *Comprehensive chemometrics* (p. 285-314). Oxford: Elsevier. Descargado de https://www.sciencedirect.com/science/article/pii/B9780444527011001253 doi: https://doi.org/10.1016/B978-044452701-1.00125-3

Cukierski, W. (2012). *Titanic - machine learning from disaster.* Kaggle. Descargado de https://kaggle.com/competitions/titanic

Encyclopedia Titanica. (2016). *Arthur jackson brewe.* Descargado de https://www.encyclopedia-titanica.org/titanic-victim/arthur-jackson-brewe.html (ref: #41, last updated: 8th July 2016, accessed 19th August 2024 08:31:49 AM)

IBM. (s.f.). *Support vector machine.* Descargado de https://www.ibm.com/topics/support-vector-machine#:~:text=What%20are%20SVMs%3F,in%20an%20N%2Ddimensional%20space.

Qué es Perceptrón Multicapa - MLP | Concepto y definición. Glosario. (s.f.). Descargado de https://gamco.es/glosario/percepatron-multicapa-mlp/#:~:text=El%20Perceptr%C3%B3n%20Multicapa%20(MLP%2C%20por,el%20campo%20del%20aprendizaje%20autom%C3%A1tico

Random forest, la gran técnica de Machine Learning. (2023, 1). Descargado de https://www.inesdi.com/blog/random-forest-que-es/

Scikitlearn. (2024, Jul). *Randomforestclassifier.* Descargado de https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

SVC-Scikitlearn. (2024). Descargado de https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Wilimitis, D. (2019, Feb). The kernel trick in support vector classification. *Towards Data Science*. Descargado de https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f