Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Inteligencia Artificial Avanzada para la Ciencia de Datos

Ingeniería en Ciencias de Datos y Matemáticas

Predicción de Pasajeros Supervivientes del Titanic

Reporte

Elaborado por

Fabián Trejo Díaz Barreiro	A01423983
Samantha Daniela Guanipa Ugas	A01703936
Eduardo Martínez Martínez	A01023975
Alexia Elizabeth Naredo Betancourt	A00830440
Miguel Á. Bermea Rodríguez	A01411671
Francia García Romero	A01769680

Monterrey, Nuevo León. 14 de septiembre 2023

1. Introducción

El aprendizaje automático o aprendizaje máquina es un subconjunto de la inteligencia artificial que se enfoca en enseñarle a las computadoras a tomar decisiones a partir de los datos y manejar los mismos sin ser totalmente programadas para esto, sino por medio de métodos de entrenamiento y modelos que permiten que estas tengan una habilidad de identificar patrones entre los datos para hacer predicciones.

El hundimiento del Titanic es uno de los naufragios más tristemente célebres de la historia. El 15 de abril de 1912, durante su viaje inaugural, el considerado insumergible RMS Titanic se hundió tras chocar con un iceberg. Desgraciadamente, no había suficientes botes salvavidas para todos los ocupantes, lo que provocó la muerte de 1,502 de los 2,224 pasajeros y tripulantes. (Geographic, 2023)

Aunque hubo algo de suerte en la supervivencia, parece que algunos grupos de personas tuvieron más probabilidades de sobrevivir que otros. En este reto, se busca construir un modelo predictivo que responda a la pregunta: "¿qué tipo de personas tenían más probabilidades de sobrevivir?ütilizando los datos de los pasajeros (es decir, nombre, edad, sexo, clase socioeconómica, etc.). En el siguiente reporte se evaluarán diferentes técnicas y modelos que resolverán la problemática descrita.

2. Descripción de Variables

2.1. Variables Numéricas

Descripción general

Se utilizó el método *describe()* para obtener un resumen estadístico de las variables numéricas del conjunto de datos. Los resultados muestran que la edad promedio de los pasajeros era de aproximadamente 29.7 años, con una desviación estándar de 14.5 años. El número promedio de hermanos/cónyuges y padres/hijos a bordo era de 0.52 y 0.38, respectivamente. La tarifa promedio pagada por los pasajeros fue de 32.20, con una desviación estándar de 49.69.

Todas las variables numéricas tienen una asimetría positiva (sesgo hacia la derecha), especialmente la variable Fare (la tarifa).

Posteriormente se generó un diagrama de caja y bigotes (boxplot) para visualizar la distribución de las variables numéricas del conjunto de datos, como se muestra a continuación:

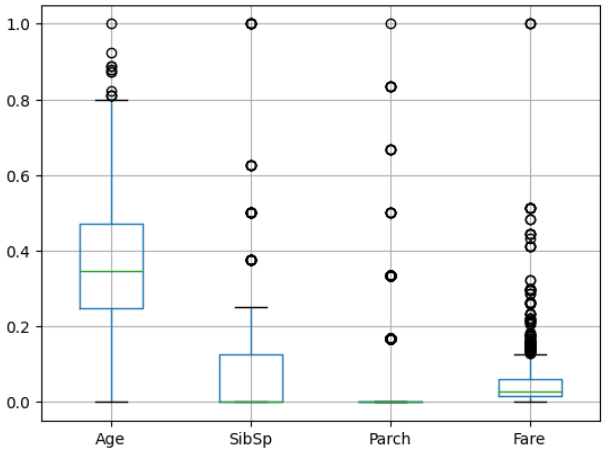

Figura 1: Boxplot de variables numéricas

Después se generó un mapa de calor con el propósito de visualizar la correlación existente entre las variables

numéricas, a continuación se presenta dicho gráfico:

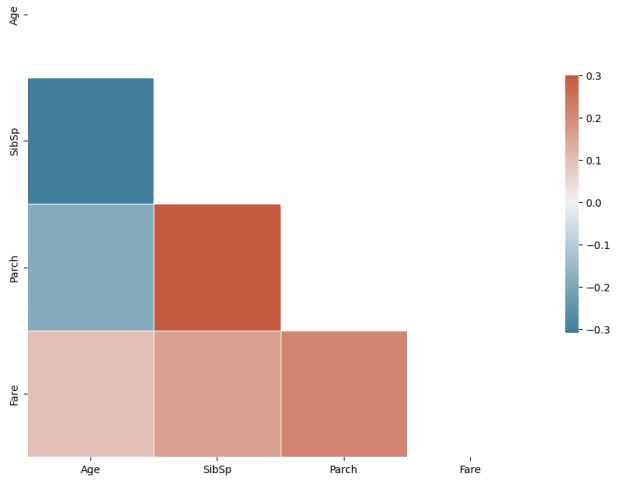

Figura 2: Heatmap de variables numéricas

Posteriormente se hizo un análisis individual a cada una de las variables numéricas del conjunto de datos. Haciendo uso de la representación visual de la variable mediante histograma y boxplot.

Empezando por la variable Age , en la cual se generó su respectivo histograma (Figura 7) y su boxplot (Figura 8). En este último se pudieron observar los outliers de la variable de una manera más clara, mientras que en el histograma se pudo observar que la distribución de datos esta moderadamente sesgada a la derecha.

Después para la variable SibSp nuevamente se generó su respectivo histograma (Figura 9) y su boxplot (Figura 10). En este caso el histograma reflejó que los datos están muy sesgados a la derecha. Por otro lado, los datos atípicos reflejados en el boxplot son más que evidentes.

En cuanto al histograma (Figura 11) y el boxplot (Figura 12) de la variable Parch podemos observar que está muy sesgado a la derecha, en este caso en particular se observa que la inmensa mayoría recae en 0 en el eje X (es decir, la mayoría de tripulantes tiene 0 Number of Parents/Childrens), por lo que el boxplot identifica todos los demás como datos atípicos (outliers).

Finalmente para la variable Fare se observa en el histograma (Figura 13) que también esta muy sesgada a la derecha, además de que observamos los outliers; especialmente el de valor cercano a 500 que se separa de los demás datos atípicos como se visualiza en el boxplot correspondiente (Figura 14).

2.2. Variables Categóricas

Descripción general

Con el propósito de describir las variables categóricas resultó indispensable cuantificar la frecuencia de cada una de las categorías de cada una de las variables. Esto se logró con un bucle sencillo para iterar por cada una de las variables: *dataframe[columna de variable categórica].value_counts()*.

Para facilitar la visualización en este reporte de las frecuencia obtenidas se diseñaron las siguientes representaciones tabulares:

Pclass	Count
3	491
1	216
2	184

(a) Variable: Pclass

Sex	Count
male	577
female	314

(b) Variable: Sex

Embarked	Count
S	644
C	168
Q	77

(c) Variable: Embarked

Tabla 1: Frecuencias de valores en Pclass, Sex y Embarked

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

Tabla 2: Frecuencia de valores en Title

Posteriormente se hizo un análisis individual a cada una de las variables categóricas del conjunto de datos. Haciendo uso de gráficos de barras para representar la distribución de la variable y los sobrevivientes por cada categoría de la variable.

Primero como ya se menciono se utilizaron gráficos de barras para representar la distribución por cada categoría de la variable. Dichos resultados que se ven reflejados en las distribuciones incluidas en la sección de Anexos o el el notebook desarrollado, dado que es una por variable.

Estudiando estas distribuciones, se encontró que Pclass de tercera clase es la de mayor frecuencia entre los tripulantes. Además, se observó un gran desbalance en las categorías de la variable Sex siendo que hay casi el doble de hombres que de mujeres. Por otro lado, el puerto/muelle de embarque con mayor número de tripulantes fue Southampton con una inmensa diferencia en comparación a Cherbourg y Queenstown.

Asimismo, también se realizo la distribución de sobrevivientes por categoría nuevamente correspondiente a cada variable (se incluyen en Anexos). De estas se puede visualizar que hubo mayor numero de sobrevivientes en los tripulantes de 1era clase (Pclass : 1) seguido de 3era clase, que ya habíamos mencionado que es la categoría de ticket mas frecuente entre los tripulantes; en el caso de la variable Sex también vemos que la categoría con mayor frecuencia en la distribución nuevamente no es la categoría que tuvo más sobrevivientes, en este caso hubieron más sobrevientes mujeres que hombres, por una diferencia de aproximadamente el doble. Por otro lado, la mayoría de sobrevivientes fueron del puerto de embarque Southampton, aunque hay que considerar que también fueron el de mayor número de tripulantes.

Finalmente se realizó la distribución de sobrevivientes de manera general dado por la columna/variable Survived donde 0: No y 1: Yes. Lo que nos confirma que en el conjunto de datos proporcionado para el entrenamiento (*train.csv*), hay 549 tripulantes que no sobrevivieron y 342 tripulantes que si.

3. Exploración de Datos

En esta sección se expondrán los descubrimientos obtenidos de la descripción de las variables. El conjunto de datos con el cual se estará trabajando (el cual se obtuvo de `Kaggle Titanic Competition`), cuenta con 891 registros de pasajeros, y 12 variables. Como variable dependiente (la variable a predecir) tenemos *Survived*, la cual toma un valor de 1 cuando se trata de un pasajero que sobrevivió y 0 cuando se trata de un pasajero que no sobrevivió. Como variables numéricas tenemos *Age, SibSp, Parch* y *Fare*, las cuales hacen referencia a edad, número de hermanos o cónyuges, número de padres o hijos, y costo del boleto, respectivamente. Como variables categóricas se tiene *Pclass, Sex, Cabin* y *Embarked*, las cuales hacen referencia a la clase de hospedaje, sexo, cuarto y puerto de embarcación de un pasajero. Mientras que *Pclass* es una variable categórica ordinal, el resto son nominales. Por último tenemos columnas con valores de identificación únicos: *PassengerId, Name* y *Ticket*. Aunque no pueden usarse directamente estos valores para predicción, pueden conteenr información que puede ser extraida para su análisis.

Las columnas *Age, Cabin* y *Embarked* contienen valores faltantes, por lo cual se evaluará si es óptimo conservar estas variables u omitirlas. Respecto a *Cabin*, se falta el 70 % de los valores, por lo que lo mejor será eliminarla. La edad, por otra parte, es una variable que puede ser de importancia para la predicción de supervivencia, por lo que se aplicarán técnicas de imputación de datos en la sección de *Preparación de Datos* 4. Para la columna *Embarked*, falta el 0.2 % de los datos, por lo que la mejor opción es eliminar los registros en donde falten valores.

- **Variables Numéricas**

  Graficando la distribución de las variables (ver Anexos 8) y calculando el coeficiente de sesgo, se pudo observar que todas cuentan con asimetria positiva, principalmente la variable *Fare*. Calculando las medidas de dispersión, se pudo también observar que *Fare* es igual la columna con mayor variabilidad. Analizando la matriz de correlación realizada en la sección de *Descripcion de Variables*, podemos observar que no hay correlación significativa entre estas variables, es decir, son independientes entre sí.

  Respecto a los valores atípicos, podemos observar en los gráficos boxplot incluidos en la sección de Anexos 8, que las variables *Fare, SibSp* y *Parch* tienen la mayor cantidad de *outliers*. El manejo de este tipo de datos se abarcará en la sección *Preparación de Datos* 4.

- **Variables Categóricas**

  La variable *Pclass* cuenta con 3 categorias: 1, 2 y 3, y la moda es 3. La variable *Sex* tiene dos categorias: *female* y *male*, y su moda es *male*. En el caso de *Embarked* hay 3 categorias: C, Q y S, siendo esta última la moda. Para la variable que se intentará predecir (*Survived*) es importante notar que hay un desbalance en la distribución de las clases (ver figura **??**): se tiene información de 549 pasajeros que no sobrevivieron, y 342 que sí sobrevivieron. La diferencia no es abismal, sin embargo, si podría representar un sesgo en las predicciones del modelo a implementar.

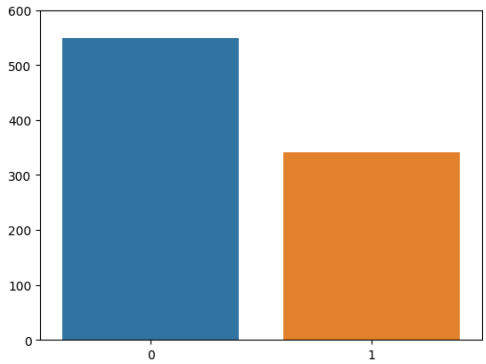

Figura 3: Distribución de supervivientes en el dataset.

4. Preparación de Datos

4.1. Feature Engineering

Partiendo de la descripción anterior de las variables, se tiene que la primera modificación de los datos realizada como parte de la preparación para su uso en modelos fue la creación de la columna *Title* extrayendo el título (*Mr., Ms., Master*, entre otros) de la columna *Name*. Esto se realizó mediante la identificación de una expresión regular (*RegEx*) para localizar el titulo dentro del nombre dado. Se identificó que este patrón ubicaba el título de una persona siempre después de una coma y antes de un punto. Una vez habiendo realizado esto, se eliminó la columna de *Name*. La columna *Ticket* también se eliminó ya que al ser igualmente una columna de identificadores, carece de valor predictivo.

4.2. Valores Faltantes

La columna *Cabin* igualmente fue eliminada, como descrito en la sección *Exploración*, debido a la gran cantidad de valores faltantes (77 %) que contenia. La única otra columna con valores faltantes era la correspondiente a la edad (*Age*) de los pasajeros En este caso, se tuvo unicamente un 20 % de datos faltantes, por lo que se recurrió a **Técnicas de Imputacion de Datos**. Se consideraron distintas técnicas para la predicción/llenado de los datos:

- **Random Forest**

  Se realizó un modelo para predecir la edad con base en las demás variables del *dataset* (excluyendo la variable *Survived*).

- **Media**

  Se realizó el llenado de los datos faltantes con la media de todos los datos de la columna.

- **Mediana**

  Se obtuvo la mediana del valor de edad para cada categoría de la variable *Title*, puesto que cada una (por la naturaleza de la variable) tiende a hacer referencia a un grupo de edad muy especifico (*Master*:niños, *Ms*: mujeres jóvenes, entre otros). Se consideró que la mediana seria una mejor estimación de la edad puesto que la distribución por grupo presenta asimetría (como se puede observar en la figura 4.2).

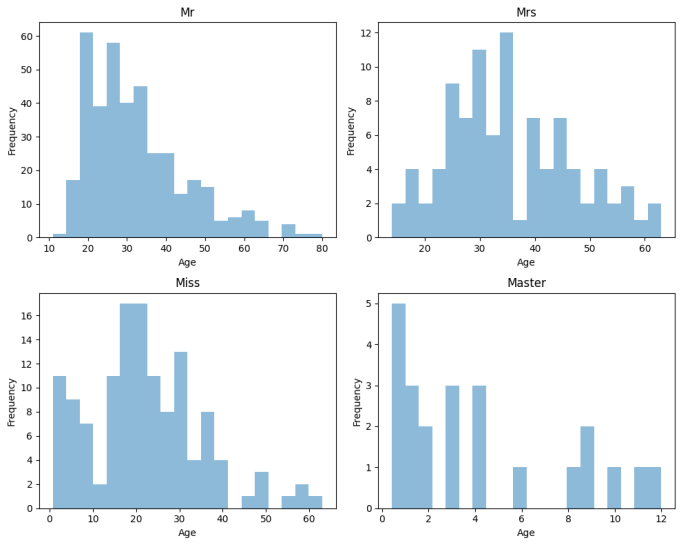

Figura 4: Distribución de la edad por titulo. Asimetría positiva en todas las variables.

Se evaluó cada uno de los métodos prediciendo valores de edad conocidos con cada uno. Los resultados se muestran en la figura 4.2. Podemos observar que la imputación por mediana fue la que obtuvo un mejor desempeño, aunque por muy poco.

```
Random Forest Imputation  | 0.8275
Mean Imputation           | 0.8250
Median Imputation         | 0.8287
```

Figura 5: Resultado de distintas técnicas de imputación de datos.

4.3. Valores Atípicos

- **Variables Numéricas**

  Para la eliminación de valores atípicos numéricos, se tomó como criterio el número de desviaciones estándar que un cierto dato tiene de alejamiento respecto a la media. En este caso, se estableció un número máximo de 3 desviaciones estándar para considerar un valor como atípico. Los registros eliminados contenían datos atípicos en al menos una variable. En total se eliminó un 7% de los registros totales disponibles (1 - 827 atípicos /891 datos totales). Principalmente se eliminaron aquellos registros con título únicos (*Jonkheer, Capt, Col, Major*)

- **Variables Categóricas**

  Los datos atípicos categóricos se consideraron como aquellas categorías con pocas apariciones. Se estableció en este caso la cantidad mínima de apariciones de una categoría como 10. Esto se consideró debido a que las categorías con pocos miembros no son representativas. En este caso, la proporción de datos eliminada fue de 4%.

4.4. Valores Duplicados

Un vez que se completaron los datos faltantes en la columna *Age*, ya que se utilizó un valor constante por categorías, es posible que haya registros que sean exactamente iguales. Esto puede significar redundancia al entrenar el modelo, por lo cual se eliminaron los registros idénticos, dejando solamente una instancia de cada uno.

4.5. Codificación de Variables Categóricas

Para las categorías de cada una de las variables cualitativas, se creó una variable binaria nombrada con la notación *variableOriginal_categoria* para indicar si a un cierto registro le corresponde esa categoría (en cuyo caso se le asigna el valor de 1) o si no le corresponde (se le asigna un valor de 0). Las variables a las que se les aplicó este proceso fueron: *Pclass, Sex, Embarked* y *Title*.

4.6. Selección de Features

El coeficiente de correlación de Spearman fue calculado para observar la interacción entre las nuevas variables y la variable objetivo (como puede observarse en la figura 4.6. Para las variables categóricas también se realizó una prueba chi-cuadrada, en donde se observó que las variables de *Embarked* debian ser omitidas. Esto debido a la presencia de suficiente evidencia estadistica para poder aceptar, con un valor de significancia de 0.03, la hipótesis nula de la prueba (H_0: La relación entre las variables categóricas no es significativa).

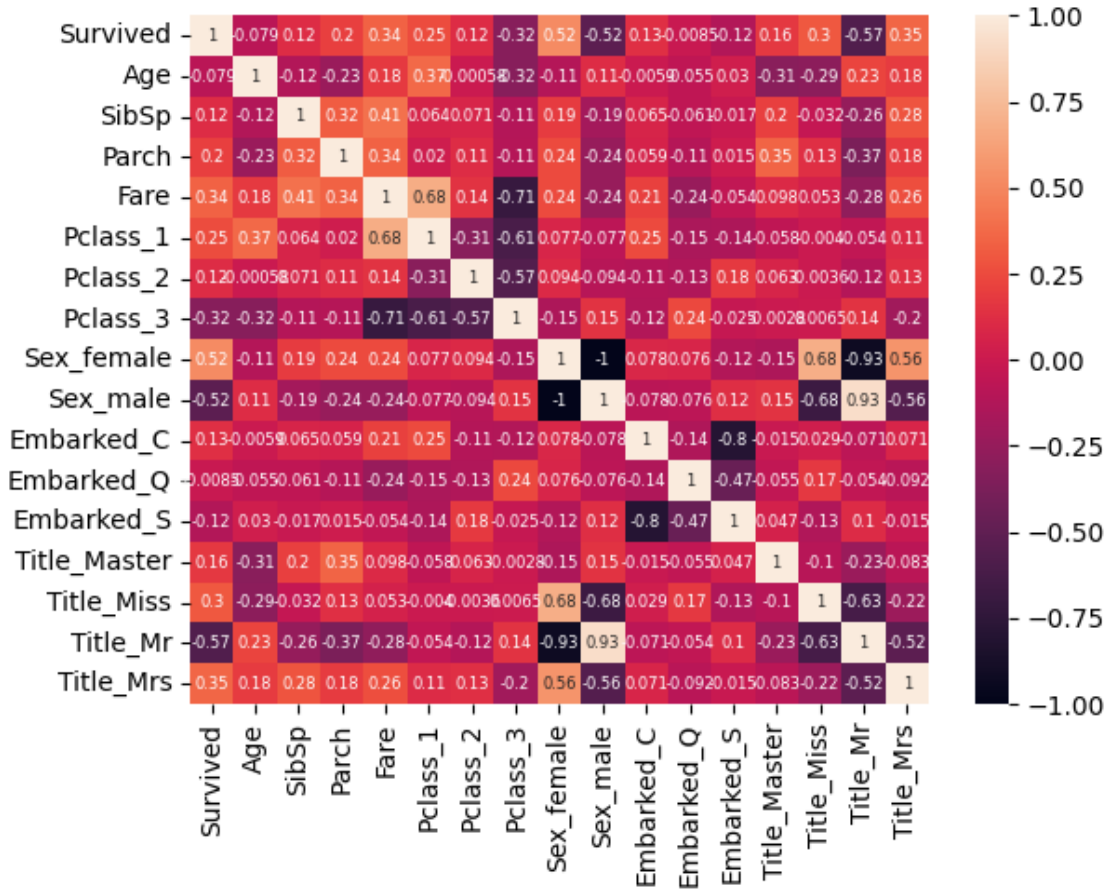

Figura 6: Resultado de distintas técnicas de imputación de datos.

Tomando ambas pruebas (correlación y chi-squared) como base, se eligieron las siguientes variables por el nivel de relevancia de su relación con la variable *Survived*:

- Parch
- Fare
- Pclass_3
- Sex_female
- Title_Master
- Title_Miss
- Title_Mrs

4.7. Transformación de los Datos

Como último paso, se aplicó una estandarización *minmax* al set de datos para poder tener un mejor *performance* en el entrenamiento de los modelos que se presentan en la siguiente sección. Asimismo, se creo un dataset con los datos normalizados para poder ser utilizado en los modelos que supongan normalidad en los datos.

5. Modelos

5.1. Random Forest

Descripción

Esta es una técnica de aprendizaje automático supervisado que se basa en árboles de decisión, que cuentan con una serie de preguntas con respuestas de ”sí o noz estas respuestas obtenidas llevarán a la decisión final. (IBM, 2023c)

Podemos describirlo como un algoritmo de aprendizaje automático basado en la técnica de ensemble. Como se mencionó, unciona construyendo múltiples árboles de decisión durante el proceso de entrenamiento y luego combinando sus predicciones para obtener un resultado final. Cada árbol se entrena con una muestra aleatoria de los datos y una selección aleatoria de características, lo que reduce el riesgo de sobreajuste. Luego, cuando se necesita hacer una predicción, los resultados de los árboles individuales se promedian en problemas de regresión o se vota en problemas de clasificación. Esto proporciona un modelo robusto y generalmente preciso que es útil en una variedad de aplicaciones. En resumen, es un método de conjunto en el que cada árbol se entrena como un subconjunto, y luego se combinan los resultados de todos los árboles para tomar una decisión final. Este método es muy útil para casos como el planteado en el reto, en problemas de clasificación binaria. (IBM, 2023c)

Proceso de entrenamiento

Para entrenar este modelo, primeramente se crearon dos dataframes, uno que es ”X_r”(que almacenará los features o variables para evaluar si la persona sobrevivió o no) y otra z_r”(que contendrá el valor 0 o 1 de si sobrevivió o no la persona).

Posteriormente, se dividió el dataset entre 60 % para entrenamiento y el otro 40 % para las pruebas. Por último, se creó una variable ”forest”que contendrá el bosque aleatorio, y se utilizó la función RandomForestClassifier”de la librería ”Sklearnçon 550 estimadores, utilizando el principio de entropía y con una profundidad de 200. Al tener el bosque, se utilizó la función ”fit()”.

Para entender cómo se selecciona una característica para una división en el árbol, es esencial entender las métricas de impureza. Una métrica común es la entropía, definida como:

$$H(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

Donde p_+ y p_- son las proporciones de ejemplos positivos y negativos en el conjunto S, respectivamente.

Con base en la entropía, se calcula la ganancia de información (IG) para decidir cuál característica es la mejor para dividir:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v)$$

Aquí, S es el conjunto actual, A es el atributo que estamos considerando, $Values(A)$ son los diferentes valores que A puede tomar, y S_v es el subconjunto de S donde A toma el valor v.

Resultados

Se hizo el cálculo del *accuracy* para el entrenamiento y la prueba de este modelo, y se obtuvo un 90 % de *accuracy* para entrenamiento y un 78.9 % *accuracy* para pruebas, por lo que se puede concluir que es un

buen modelo a implementar para este proyecto.

5.2. Logistic Regression

Descripción

La Regresión Logística es una extensión del modelo lineal generalizado, diseñado para manejar resultados de variables dependientes categóricas. (IBM, 2023a) En nuestro caso, abordamos un problema de clasificación binaria donde el objetivo es modelar la probabilidad condicional $P(Y = 1|\mathbf{x})$ dadas las características $\mathbf{x}$.

No es un método de regresión, sino que se utiliza para predecir la probabilidad de que una instancia pertenezca a una clase particular. Utiliza la función logística para modelar esta probabilidad, lo que asegura que la salida esté entre 0 y 1. Los coeficientes se ajustan para maximizar la verosimilitud de los datos observados, lo que permite realizar predicciones basadas en las características de entrada. (IBM, 2023a)

La formulación matemática del modelo es:

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}))}$$

Aquí, β_0 es el término de intercepción y $\boldsymbol{\beta}$ es el vector de coeficientes asociados con las características $\mathbf{x}$. La función logística asegura que la probabilidad estimada se encuentre en el intervalo mencionado.

En este trabajo, el modelo incluye características como 'Parch', 'Fare', 'Pclass_3', 'Sex_female', 'Title_Master', 'Title_Miss', y 'Title_Mrs', cada una con su coeficiente asociado que indica la importancia relativa para la clasificación.

Proceso de Entrenamiento

Optimización y Ajuste del Modelo

Se empleó un algoritmo de optimización basado en el método de máxima verosimilitud para estimar los parámetros β_0 y $\boldsymbol{\beta}$. Los coeficientes resultantes fueron $[-0.315, 1.013, -1.169, 1.727, 2.700, 0.676, 1.051]$ y la intercepción fue -0.973.

Validación Cruzada

Se realizó una validación cruzada de 5-folds para evaluar la robustez del modelo. El accuracy media obtenida fue del 79 % con una desviación estándar de 0.021, indicando un rendimiento sólido y consistente a través de diferentes particiones del dataset de entrenamiento.

Métricas de Evaluación

Además del accuracy, se podrían considerar métricas adicionales como la curva ROC, la precisión, el recall y el F1-score para una evaluación más completa del modelo. Sin embargo, el accuracy por sí sola ya ofrece una evaluación razonablemente buena de la eficacia del modelo en este caso.

Consideraciones de Overfitting y Regularización

Dado que el modelo muestra una variabilidad baja en las puntuaciones de accuracy a través de los folds en la validación cruzada, se considera que el modelo es robusto y no sufre de sobreajuste. Sin embargo, para conjuntos de datos más grandes o más complejos, técnicas de regularización como Lasso o Ridge podrían implementarse para mejorar la generalización del modelo.

Resultados

Rendimiento del Modelo

El rendimiento del modelo de Regresión Logística se evaluó utilizando validación cruzada de 5-folds en el conjunto de entrenamiento. El accuracy media resultante fue del 79 %, con una desviación estándar de 0.021. Este nivel de accuracy y variabilidad sugiere que el modelo es tanto preciso como robusto para el problema en cuestión.

Interpretación de Coeficientes

Los coeficientes del modelo, estimados a través del método de máxima verosimilitud, fueron los siguientes:

$$\text{Coeficientes: } \begin{bmatrix} -0.315 & 1.013 & -1.169 & 1.727 & 2.700 & 0.676 & 1.051 \end{bmatrix}$$

$$\text{Intercepción: } -0.973$$

Cada coeficiente representa el cambio en los log-odds de la variable dependiente por una unidad de cambio en la variable predictora correspondiente, manteniendo todas las demás constantes. Por ejemplo, el coeficiente para 'Fare' es 1.013, lo que indica que un aumento en la tarifa normalizada se asocia con un aumento en la probabilidad de supervivencia.

5.3. K Nearest Neighbors

Descripción

Esta es una técnica de aprendizaje automático supervisado que se basa en la distancia entre los puntos de datos, que se representan como vectores en un espacio multidimensional. Funciona considerando los K puntos de datos más cercanos a una instancia de prueba y asignando una etiqueta o valor basado en la mayoría de las etiquetas de sus vecinos (para clasificación) o en el promedio de sus valores (para regresión). La elección de K y la métrica de distancia (por ejemplo, Euclidiana o Manhattan) son parámetros esenciales que afectan el rendimiento del algoritmo. (IBM, 2023b)

En resumen, es un método de instancia, en el que cada punto se clasifica según la mayoría de votos de sus k vecinos más cercanos, donde k es un parámetro definido por el usuario. Este método es muy útil para casos como el planteado en el reto, en problemas de clasificación binaria.

Proceso de Entrenamiento

Para entrenar este modelo, primeramente se crearon dos dataframes, uno que es "X_k"(que almacenará los features o variables para evaluar si la persona sobrevivió o no) y otra ƶ_k"(que contendrá el valor 0 o 1 de si sobrevivió o no la persona).

Posteriormente, se dividió el dataset entre 80 % para entrenamiento y el otro 20 % para la validación. Por último, se creó una variable "knn"que contendrá el clasificador de k vecinos más cercanos, y se utilizó la función "KNeighborsClassifier"de la librería "Sklearnçon 5 vecinos, utilizando la distancia euclidiana como métrica:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Al tener el clasificador, se utilizó la función "fit()".

Finalmente, se iteró a través de diferentes valores de K para maximizar el accuracy. Encontramos que el valor k=12 alcanzaba la mayor precisión.

Para clasificar un punto p basado en sus k vecinos más cercanos, se realiza la siguiente votación:

$$\text{Clase de } p = \begin{cases} 1 & \text{si } \frac{\text{Número de vecinos de clase 1}}{k} > 0.5 \\ 0 & \text{en caso contrario} \end{cases}$$

Resultados

Se hizo el cálculo de precisión para el entrenamiento y la prueba de este modelo, y se obtuvo un 87% de accuracy para entrenamiento y un 79% de accuracy para validación, por lo que se puede concluir que es un modelo aceptable a implementar para este proyecto.

5.4. Support Vector Machine

Descripción

El Support Vector Machine (SVM) es un algoritmo de aprendizaje automático usado en problemas clasificación y regresión. Busca encontrar un hiperplano óptimo en un espacio de características de alta dimensión para separar dos clases. El objetivo es maximizar el margen entre los puntos de datos de diferentes clases. SVM puede manejar tanto problemas lineales como no lineales mapeando los datos a un espacio de características más elevado. Los vectores de soporte son ejemplos de datos que están más cerca del hiperplano de decisión y son cruciales para determinar la posición del hiperplano. SVM es efectivo en la clasificación de datos complejos y se utiliza en una amplia gama de aplicaciones, desde la detección de spam hasta el diagnóstico médico. (IBM, 2023d)

La ecuación general del hiperplano es:

$$w \cdot x + b = 0$$

Cuando los datos no son linealmente separables, el SVM utiliza un kernel para transformar los datos a un espacio de mayor dimensión. Un kernel común es el kernel radial (RBF):

$$K(x, x') = e^{-\gamma \|x - x'\|^2}$$

(IBM, 2023d)

Proceso de Entrenamiento

Para entrenar este modelo, se dividieron los datos en características y etiquetas, y de estos datos un 80% para entrenamiento y 20% para pruebas.

Posteriormente se ajustaron los datos para tener la misma escala, y a partir de esto se construyó el modelo SVM con un kernel radial. La función objetivo de SVM busca maximizar el margen entre las clases y se puede representar como:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{sujeto a: } y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$

Resultados

El modelo acertó en aproximadamente el 77 % de las veces cuando predijo que alguien sobreviviría (accuracy) y tuvo una accuracy general del 78 % en la clasificación de supervivencia (exactitud).

6. Elección y Afinamiento de Modelo

Aunque K-Nearest Neighbors tuvo el *accuracy* más alto para las pruebas de entrenamiento, este modelo obtuvo el mismo valor de la misma métrica para las predicciones con el set de pruebas que Regresión Logistica. Dado que este último cuenta con menor variabilidad entre las métricas de performance del modelo, y no tiene sesgo significativo, se eligió como el modelo más apropiado para los datos. Para poder afinar los hiper parámetros de este modelo, se aplicó la técnica de **grid search** para la búsqueda de los valores óptimos.

6.1. Grid search

Teniendo en cuenta que "este tipo de búsqueda prueba todas las posibles combinaciones de valores que se le proporcione en el grid de parámetros"(Francisco Sanz, 2023), se especificó una cuadrícula de hiperparámetros que se irá explorando durante el proceso de búsqueda. Los hiperparámetros incluyen: el tipo de penalización "penalty", el parámetro de regularización "C", el algoritmo de optimización "solver", y el número máximo de iteraciones "max_iter".

Para determinar la combinación óptima de hiperparámetros, se utilizó el "grid search" en conjunto con "cross validation". Esta combinación implica entrenar y evaluar el modelo de regresión logística con todas las combinaciones posibles de los valores de hiperparámetros definidos previamente. El método cross validation se empleó para evaluar el rendimiento del modelo en diferentes divisiones del conjunto de entrenamiento, lo que ayuda a evitar el sobreajuste. Después de completar el "grid search", se identifican los mejores hiperparámetros que resultaron en el mejor rendimiento en términos de accuracy, que en este caso fueron : "C": 10, "max_iter": 100, "penalty": l1, "solver": liblinear.

Con los hiperparámetros óptimos en mano, se procede a entrenar un nuevo modelo de regresión logística utilizando estos valores en el conjunto de entrenamiento completo, y con esto se obtuvo un *accuracy* de 0.7929, o 79.29 % para el set de pruebas, que es muy similar al que había obtenido antes del ajuste de los parámetros, pero se observa una pequeña mejora, por lo que finalmente se utilizarán estos nuevos hiperparámetros para nuestro modelo.

7. Conclusiones

El problema que se abarcó fue un problema de clasificación binaria, dado que la variable objetivo unicamente indica si una persona sobrevivió, o no, al hundimiento del Titanic. El reto consisitió en utilizar la información disponible de los pasajeros para evaluar patrones o tendencias que pudieran indicar variables clave que influyeron en la supervivencia de cierto sector de la población. Al final del análisis exploativo de los datos, se pudieron definir aquellas variables con mayor influencia sobre *Survived*. Estas variables fueron: *Parch, Fare, Pclass_3, Sex_female, Title_Master, Title_Miss* y *Title_Mrs*.

La correlación entre las variables *Sex_female, Title_Mrs* y *Title_Miss* resultó ser positiva, mientras que *Title_Mr* tuvo una correlación negativa, por lo que se puede establecer una tendencia del sexo femenino a haber sobrevivido, mientras que ocurre lo contrario para el sexo masculino. La variable —*Fare* igualmente tiene un coeficiente positivo, mientras que *Pclass_3* tiene uno negativo, por lo que se tiene entonces una tendencia a sobrevivir de aquellas personas que pagaron más por su boleto, y por ende podría intuirse que tienen una clase social más elevada sobre aquellas personas que ingresaron como tercera clase.

Con estas variables se distintos modelos de clasificación: KNN, Regresion Logística, Random Forest y Support Vector Machine. Al calcular las métricas de desempeño de cada una, se pudo observar que en los 4 casos se obtuvieron resultados aceptables. Sin embargo, aquella que tuvo los resultados más altos fue Random Forest. Cabe mencionar que este modelo tuvo un mejor desempeño en términos de *Accuracy* y *Recall*. Para *Precision* en especifico, KNN obtuvo un 91 % en promedio, sobre un 76 % en promedio de Random Forest, por lo que si se desea que el modelo tenga una baja cantidad de falsos positivos, es decir, que no clasifique a un pasajero como superviviente cuando no es el caso, entonces la mejor opción sería KNN. No obstante, los valores de Regresión Logística fueron los que tuvieron un balance más apropiado en cuanto a varianza y sesgo entre resultados con el set de entrenamiento y set de pruebas.

Por lo anterior, se decidió elegir la regresión logística para afinamiento. Aplicando Grid Search se encontró que los mejores hiperparámetros son “C”: 10, “max_iter”: 100, “penalty”: l1, “solver”: liblinear. Con esto el resultado fue bastante similar a como era antes del ajuste de parámetros. Sin embargo, tomando en cuenta que a comparación del 79 % de accuracy en promedio que se obtuvo sin el ajuste de los hiperparámetros ahora se obtiene un 79.29 % de *accuracy* en promedio para el modelo de regresión logística, se decidió implementar el modelo con estos nuevos hiperparámetros,logrando que exista un buen fit para el análisis de la supervivencia de los pasajeros del Titanic, y logrando el objetivo de este proyecto.

Bibliografía

Geographic, N. (2023). Las vidas truncadas del Titanic: las historias de los sobrevivientes. https://historia.nationalgeographic.com.es/a/vidas-truncadas-titanic_11387

IBM. (2023a). ¿Qué es el aprendizaje supervisado? https://www.lifeder.com/sobrevivientes-del-titanic/

IBM. (2023b). K-Nearest Neighbors (KNN). https://www.ibm.com/topics/knn

IBM. (2023c). Random Forest. https://www.ibm.com/mx-es/topics/random-forest

IBM. (2023d). Support Vector Machine (SVM): ¿Cómo funciona? https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works

8. Anexos

8.1. Visualización de variables numéricas

Age

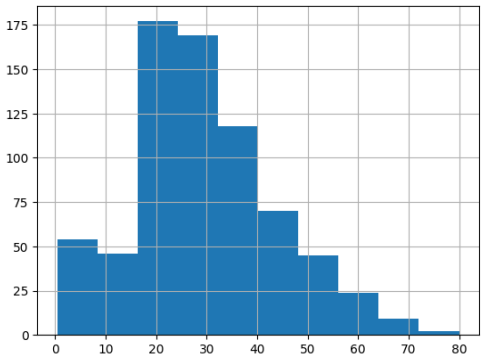

Figura 7: Histograma de Age

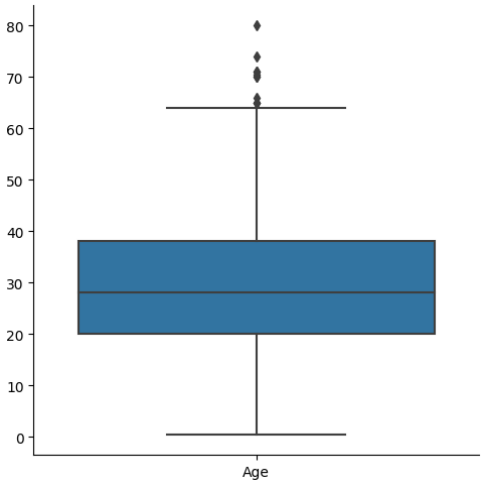

Figura 8: Boxplot de Age

Number of Siblings/Spouses

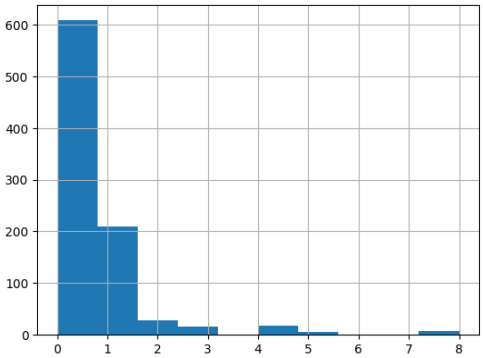

Figura 9: Histograma de SibSp

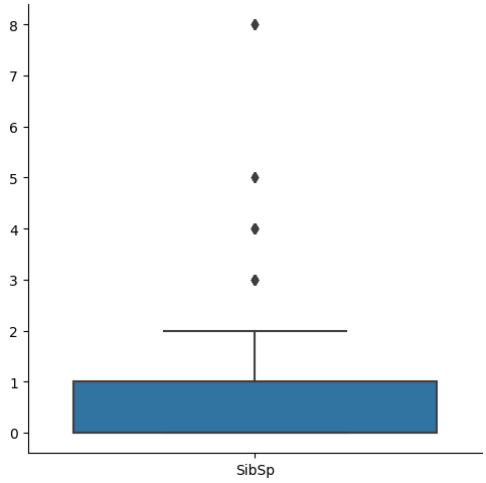

Figura 10: Boxplot de SibSp

Number of Parents/Childrens

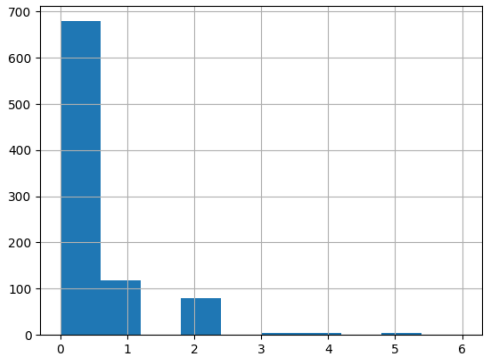

Figura 11: Histograma de Parch

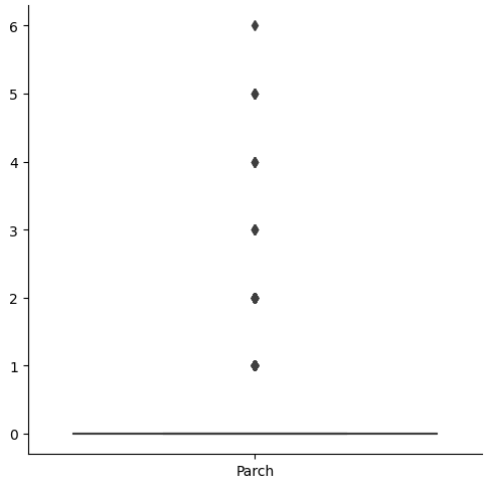

Figura 12: Boxplot de Parch

Fare

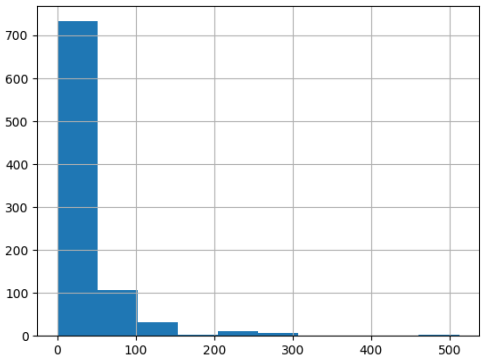

Figura 13: Histograma de Fare

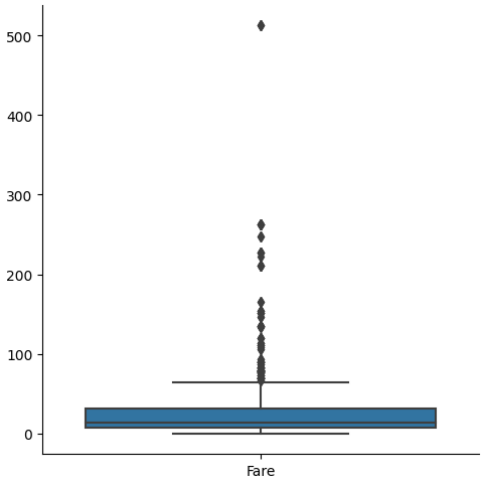

Figura 14: Boxplot de Fare

8.2. Visualización de variables categóricas

Pclass

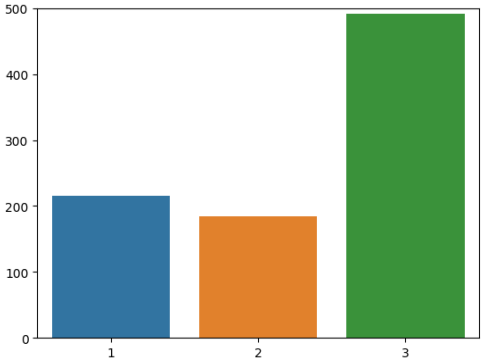

Figura 15: Distribución de 'Pclass'

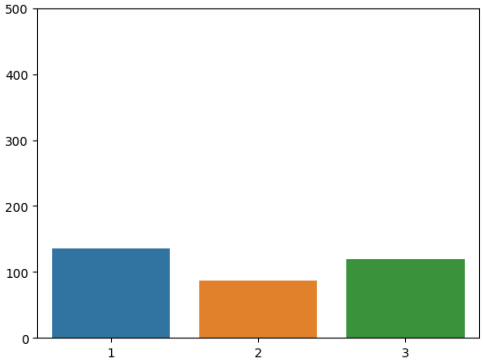

Figura 16: Sobrevivientes por 'Pclass'

Sex

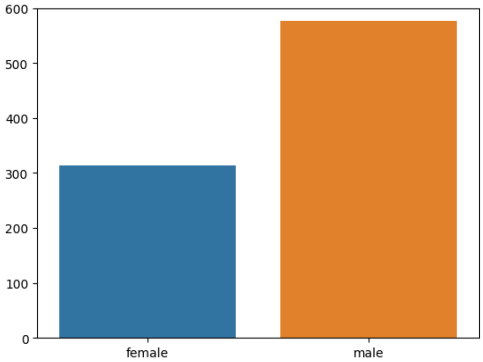

Figura 17: Distribución de 'Sex'

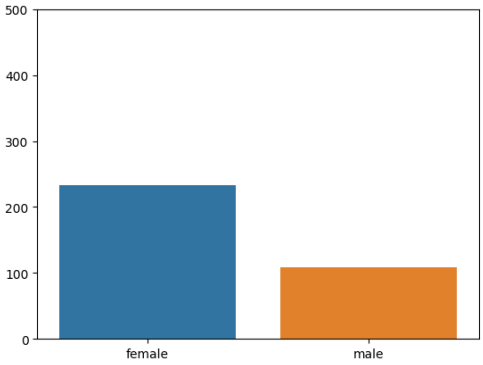

Figura 18: Sobrevivientes por 'Sex'

Embarked

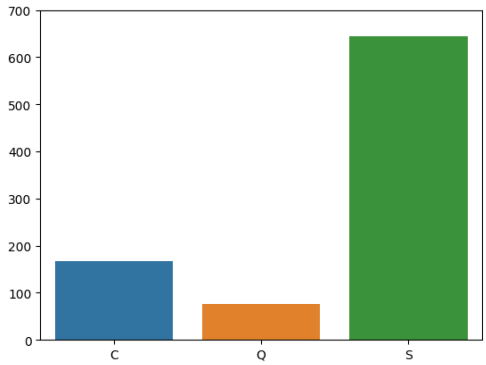

Figura 19: Distribución de 'Embarked'

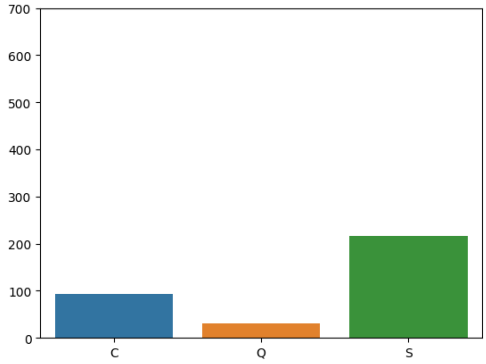

Figura 20: Sobrevivientes por 'Embarked'

Title

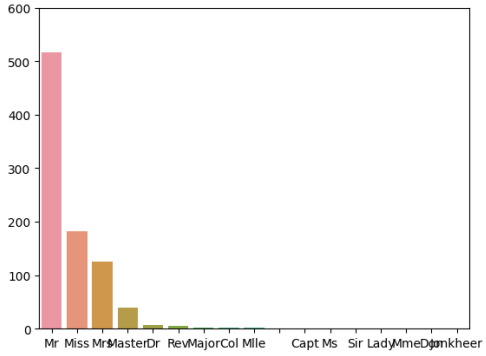

Figura 21: Distribución de 'Title'

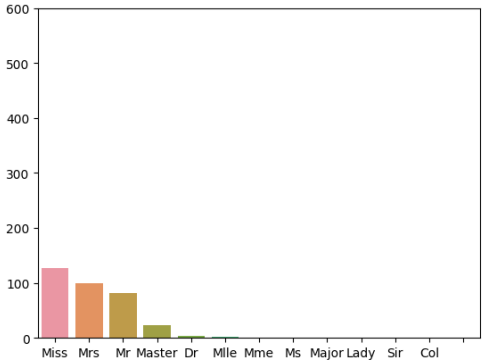

Figura 22: Sobrevivientes por 'Title'

Survived

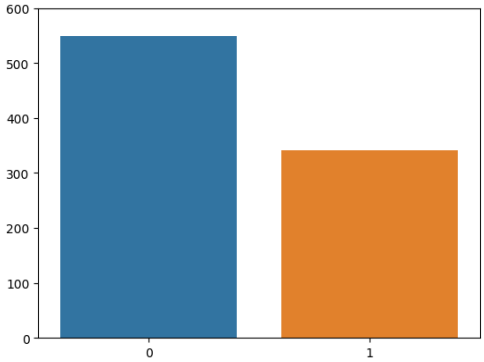

Figura 23: Distribución de 'Survived'