Momento de Retroalimentación: Reto Limpieza del Conjunto de Datos

Equipo 3:

Fernando Bustos Monsiváis - A00829931

Ramón Yuri Danzos García - A00227838

Axel Amós Hernández Cárdenas - A00829837

Josué Emmanuel Flores Mendoza - A00833132

Jesus Daniel Martínez García - A00833591

Escuela de Ingeniería y Ciencias, Instituto Tecnológico y de Estudios Superiores de Monterrey

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

Profesores:

Dr. Frumencio Olivas Alvarez

Dr. Hugo Terashima Marín

Dr. Julio Antonio Juárez Jiménez

Dr. Alfredo Esquivel Jaramillo

Lunes 19 de agosto de 2024

Índice

1. Introducción

El hundimiento del Titanic es una de las catástrofes marítimas más infames de la historia. El 15 de abril de 1912, durante su viaje inaugural, el Titanic, considerado por muchos como inhundible, colapsó al colisionar con un iceberg. Desafortunadamente, de los 2 224 pasajeros que se encontraban en la embarcación, solamente sobrevivieron 772 a causa de la falta de botes salvavidas en la embarcación (Cukierski, 2012).

Aunque la suerte siempre es un aspecto a considerar en una situación de vida o muerte, en el caso del Titanic pareciera ser que algunos grupos de personas tenían mejores probabilidades de sobrevivir que otras. Debido a lo anterior, es de interés analizar y determinar, a partir de datos de pasajeros, qué tipo de personas tenían más oportunidades de sobrevivir.

2. Objetivo

El objetivo de este reporte es abordar el proceso de preprocesamiento de datos de una forma manual mientras se sigue el esquema de *Extract, Transform & Load* (ETL) para encontrar las features clave que podrían influir en la supervivencia de los pasajeros. Con el fin de lograr lo anterior, la metodología se describe a continuación:

1. Primero, se hará un análisis del dataset otorgado, realizando un desglose de las features que contiene, el label, los tipos de datos del dataset y, finalmente, si existe la presencia de datos faltantes. Lo anterior tiene la finalidad de poder realizar una comparación con el dataset inicial una vez terminada la limpieza de datos.
2. Segundo, se procederá a utilizar métodos estadísticos y gráficos para determinar y argumentar la inclusión o exclusión de datos del dataset. Así mismo, se justificará como se manejarán las variables categóricas del dataset, como el sexo, puerto de embarque, etc.
3. Finalmente, se comparará el dataset resultante con el inicial, con el fin de observar el cambio en la calidad de los datos que se utilizarán para entrenar en el modelo.

Realizado lo anterior, se espera que el dataset resultante sea lo suficientemente bueno y que las features utilizadas contengan datos consistentes y de buena calidad para que posteriormente se pueda generar un modelo de predicción más preciso, robusto y funcional.

3. Dataset Inicial

El dataset inicial, por nombre de archivo “train.csv”, contiene un total de 891 instancias las cuales están definidas por un total de 12 features. En la siguiente tabla se muestran los features, su tipo de dato y una pequeña descripción de lo que representa en el dataset:

Feature	Descripción	Tipo de Dato
passengerId	ID del pasajero	Int64
survived	Muestra si el pasajero sobrevivió (1) o no sobrevivió (0)	Int64
pclass	Muestra la clase a la que pertenece el ticket: 1 = 1st, 2 = 2nd, 3 = 3rd	Int64
name	Muestra el nombre del pasajero	str
sex	Muestra el sexo del pasajero	str
age	Muestra la edad del pasajero	float64
sibsp	Muestra el número de hermanos / esposas en la embarcación	Int64
parch	Muestra el número de padres / hijos en la embarcación	Int64
ticket	Muestra el número del ticket del pasajero	str
fare	Muestra la tarifa que pagó el pasajero para abordar la embarcación	float64
cabin	Muestra el número de cabina del pasajero	float
embarked	Muestra el puerto de embarque del pasajero	str

Tabla 1. Features del dataset, su descripción y tipo de dato.

Después de haber identificado las características del dataset, su descripción y tipo de dato, es pertinente proceder a evaluar la calidad y completitud de los datos. Esto implica realizar un análisis para detectar la presencia de datos faltantes, o inconsistencias que puedan afectar el rendimiento de los modelos que posteriormente se implementarían.

Por lo anterior, se generó el siguiente gráfico de barras para visualizar de manera más clara la presencia de datos faltantes en cada feature del dataset:

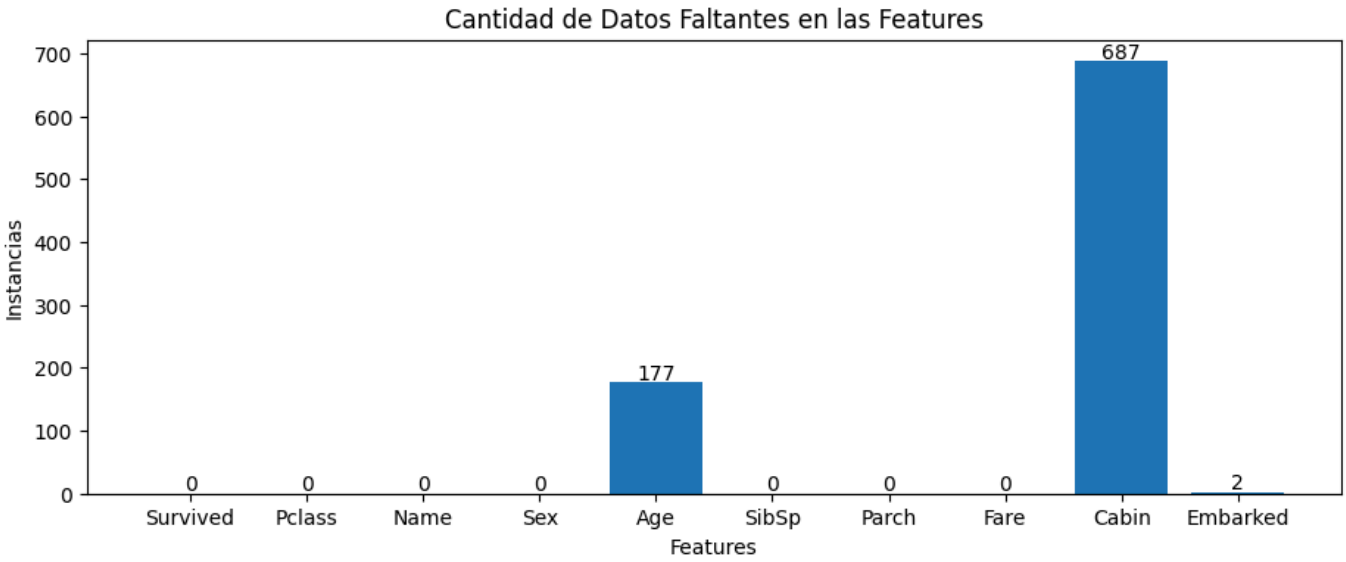

Figura 1. Distribución de datos faltantes por cada feature del dataset

Como se puede observar, los únicos features que presentan datos faltantes son los de 'Age', 'Cabin' y 'Embarked'. A continuación, se presenta una tabla que detalla el número y porcentaje de datos faltantes para cada uno de los features mencionados:

Feature	Número de Datos Faltantes	Porcentaje de Datos Faltantes
age	177	19.87%
cabin	687	77.10%
embarked	2	0.22%

Tabla 2. Número y porcentaje de datos faltantes de los features de 'Age', 'Cabin' y 'Embarked'.

En síntesis, el análisis preliminar del dataset ha proporcionado una visión detallada de las características del mismo, sentando las bases para las decisiones que se tomarán en la siguiente sección: la limpieza de datos.

4. Decisiones que tomamos

El siguiente paso será calcular y analizar los datos faltantes en cada feature del dataset. Este paso es crítico en el proceso, ya que si una feature presenta un alto porcentaje de valores faltantes, null o vacíos, es probable que no sea útil para entrenar el modelo.

4.1 Features con datos faltantes

4.1.1 Cabin

Como se vió anteriormente, el feature de *cabin* presenta un total de 697 valores faltantes, valor que que representa el 77% del total de instancias. Esta alta proporción de datos faltantes dificulta la identificación de patrones confiables para imputar datos, lo que podría introducir ruido en el modelo. Aunque, teóricamente, esta variable podría estar relacionada con la supervivencia debido a la ubicación de la cabina en el barco, el alto nivel de datos faltantes hace que esta información sea poco confiable. Además, la probabilidad de supervivencia también puede ser reflejada en otras variables con datos más completos. Por estas razones, eliminar la columna *Cabin* fue la mejor decisión para evitar comprometer la calidad del modelo.

4.1.2 Age

La *edad* de los pasajeros es un dato relevante para el análisis del problema, ya que ésta está altamente relacionada a la supervivencia. El siguiente gráfico muestra el número de supervivientes y no supervivientes agrupados por un rango de edad.

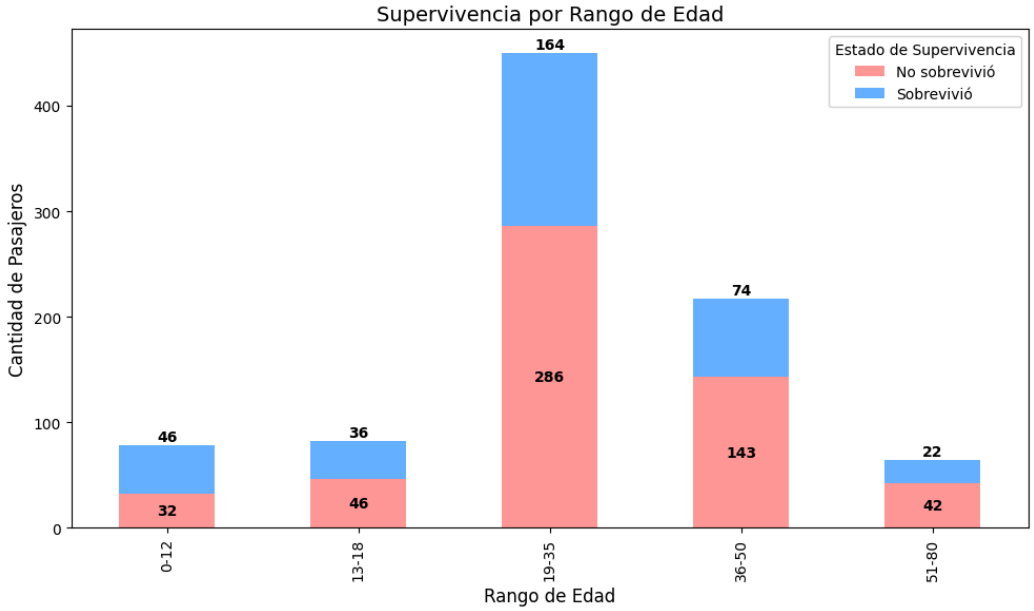

Figura 2. Distribución de supervivencia por rango de edades.

Sin embargo, como se mencionó anteriormente, la variable de edad presenta un total de 177 datos faltantes, representando casi un 20% del total de las instancias y eliminarlas puede convertirse en un problema. Como esta feature se considera relevante, se decidió utilizar una estrategia basada en calcular la edad con los títulos de las personas. La metodología de extracción se describe a continuación:

Primero, se agrupan a los pasajeros según su título al mismo tiempo que se intenta observar un patrón relacionado a sus edades. Por ejemplo, títulos como Master corresponden a pasajeros jóvenes, mientras que Mr. o Mrs. con adultos. De esta forma, se comienza a formar un patrón en el rango de edades a simple vista. Para títulos que también representan adultos (Dr., Rev., Mlle., Major, Col., Countess, Capt., Ms., Sir., Lady, Mme., Don, Jonkheer) se decidió incluirlos en títulos como Mr. y Ms., ya que comparten las características de edad similares y se pretende evitar crear grupos con pocos datos.

Segundo, una vez que se tiene a los pasajeros agrupados por título, se calcula el promedio y la desviación estándar de la edad para cada grupo. Estos valores proporcionarán una idea del rango típico de edades para cada título. De esta manera, se logra una imputación con el rango basado en la desviación estándar en lugar de imputar un valor exacto del promedio y bajo la idea de generar valores random dentro de este rango con el fin de mantener una distribución natural de las edades según los títulos.

Se puede concluir que esta técnica de imputación tiene algunas ventajas, pues los títulos extraídos tienen un contexto útil para obtener la edad mantiene la distribución al utilizar la desviación estándar y un valor aleatorio dentro de este rango, logrando reducir el ruido en el modelo, pues se respetan las características de los datos originales al insertar los faltantes.

4.1.3. Embarked

El feature de *embarked* sólo cuenta con dos datos faltantes, que representan el 0.22% de las instancias, por lo que se optó analizar si el feature tenía potencial de formar parte del dataset resultante.

Se decidió realizar un gráfico pastel que muestra la distribución entre las tres ciudades de embarque y la tasa de supervivencia por cada una, buscando algún patrón que pudiera mostrar alguna relevancia en los datos. El gráfico se puede ver a continuación.

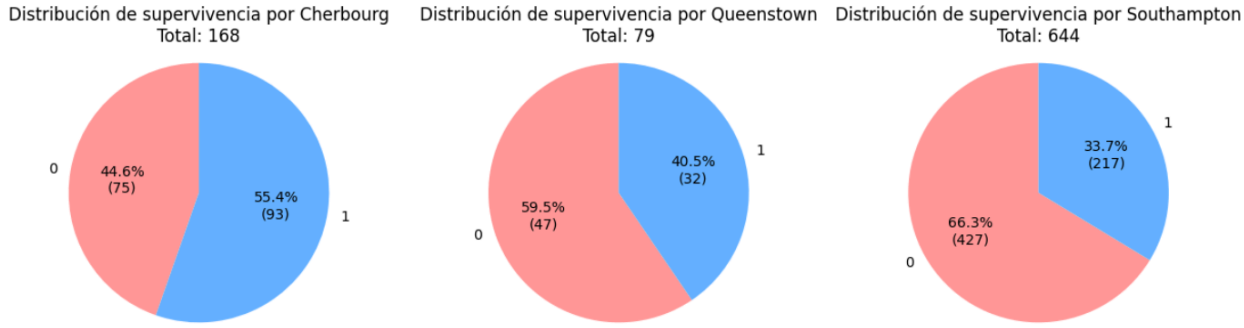

Figura 3. Distribución de supervivencia por cada puerto de embarque.

Como se puede observar, los resultados indicaron una diferencia insignificante para la embarcación en 'Cherbourg' siendo una proporción de 45 contra 55 en porcentaje, en cambio para las ciudades de 'Queenstown' y 'Southampton' la proporción de supervivencia era 1 a 3, indicando que una de cada tres personas sobrevivió, convirtiendo este feature en uno relevante porque en estas ciudades embarcó la mayor parte de los tripulantes, siendo Southampton la principal con 644 de los 889 pasajeros del dataset.

Ahora, existe una duda latente en la interpretación de estos gráficos y es que existe la posibilidad de que estemos se esté encontrando una aparente relación entre la ciudad de embarque y la tasa de supervivencia que puede estar dado por otro feature. Para descartar lo anterior, se decidió realizar la misma comparación pero esta vez separado por sexo, esperando ver si el patrón se repetía, lo cual indicaría que la relevancia real está en el sexo y no en la ciudad de embarque.

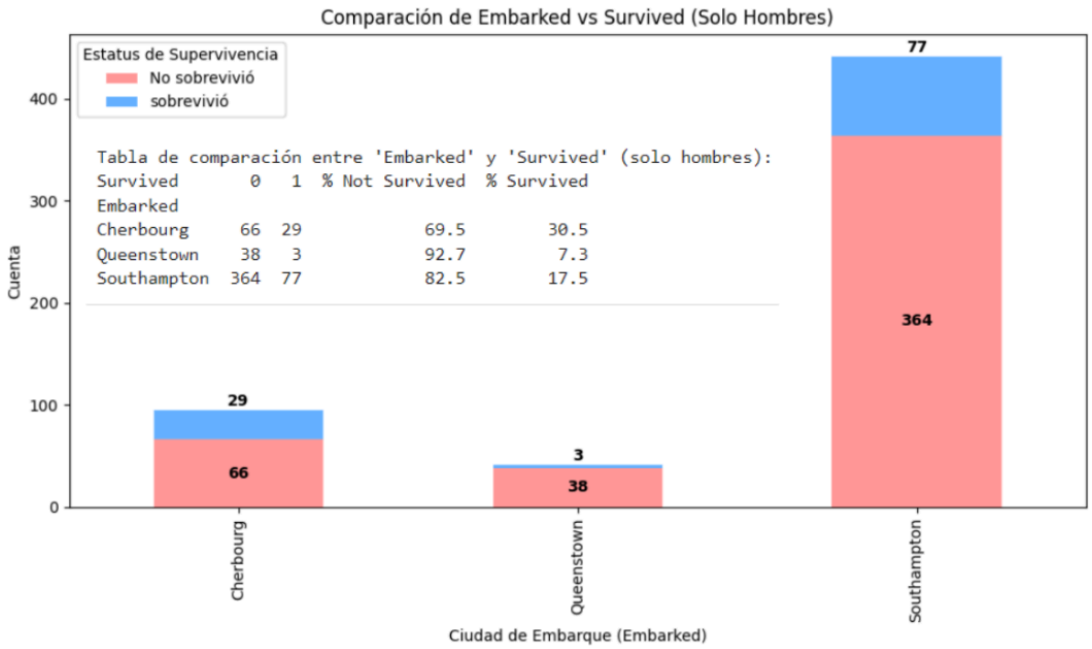

Figura 4. Comparación de Embarked vs Survived para hombres

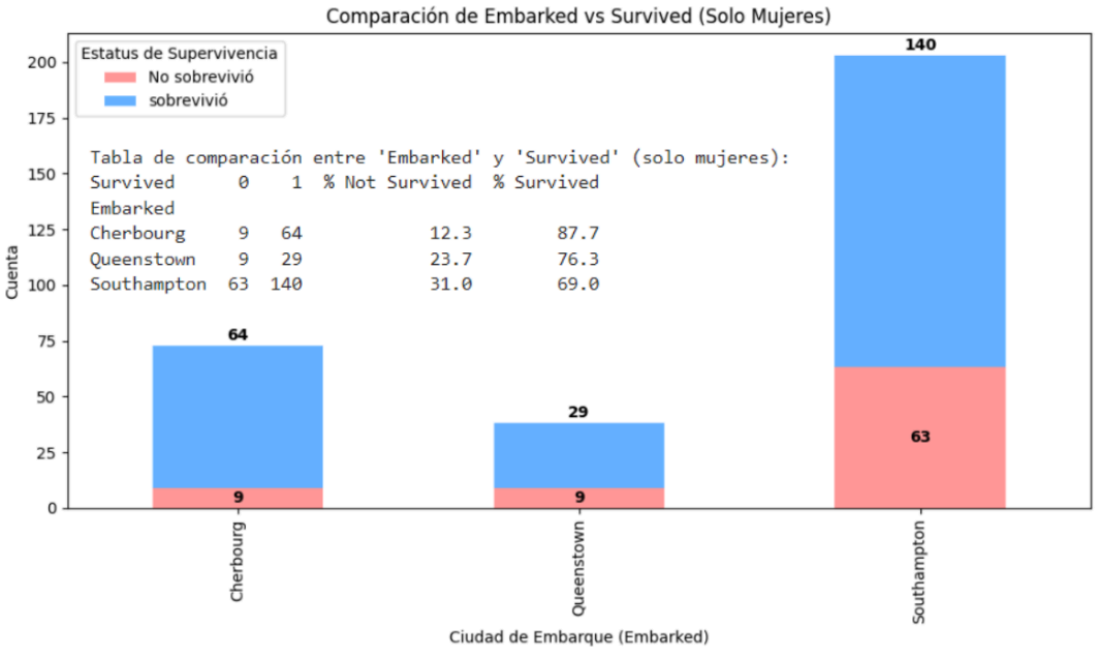

Figura 5. Comparación de Embarked vs Survived para mujeres

Analizando los gráficos realizados, se denota que el patrón se perdía tanto en mujeres como en hombres, aunque en estos se el patrón se mantiene con un incremento en gran medida, indicando cierta relación pero no lo suficiente como para eliminar el feature de Embarked.

4.1.4. Otros métodos de llenado de datos

Aunque ya se explicó en sus respectivas secciones las imputaciones de datos de cada feature que se van a conservar, se decidió que al entrenar el modelo se probarán otros métodos de imputación para comparar la precisión del modelo para verificar si se tomó la mejor decisión.

4.2 Features no relevantes

4.2.1. PassengerId

El dataset contiene un feature no relevante para el análisis que se realizará, el cual es el passengerId, cuya única función es identificar a las personas del mismo dataset.

4.2.2. Ticket

Existen un total de 681 tickets únicos en el dataset, pero esta feature contiene un total de 891 instancias. Se analizó más a fondo y se encontró que los tickets faltantes son sólo tickets que

identifican a grupos de personas, generalmente familias o parejas. Por lo anterior, se puede concluir que esta feature tiene, prácticamente, la misma función que el passengerId: identificar a las personas que abordaron el Titanic.

4.2.3. Fare

Fare, que representa el costo del boleto, está estrechamente relacionada con la feature de Pclass, la cuál indica la clase del boleto. Dado que usar ambos datos era redundante, decidimos conservar únicamente la columna Pclass y descartar el feature de fare..

4.3 Features relevantes

4.3.1. Sex

Para sex, primero se graficó la relación entre el sexo y la supervivencia. En la siguiente gráfica, se observa una fuerte correlación entre el sexo y la probabilidad de sobrevivir. Se denota que las mujeres tenían un porcentaje de supervivencia significativamente mayor que el de los hombres.

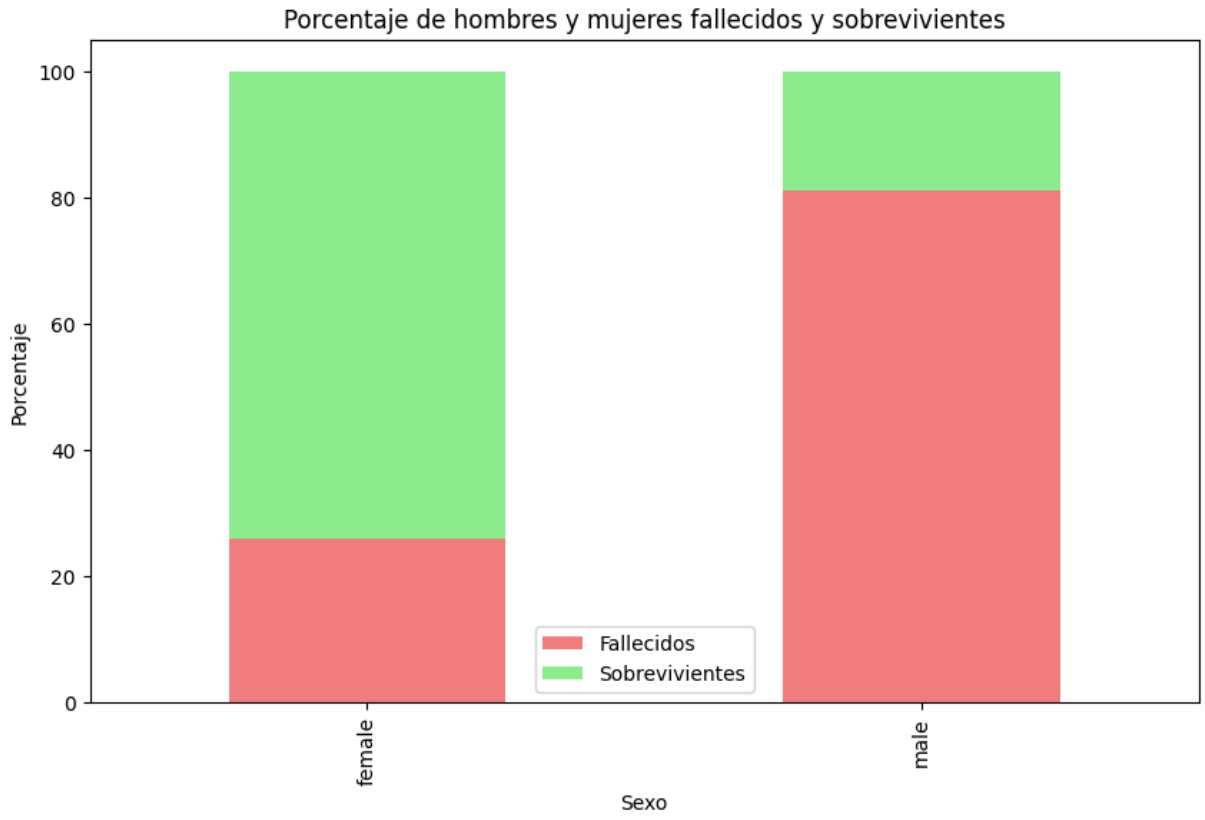

Figura 6. Porcentaje de supervivientes y fallecidos por sexo (sex)

4.3.2. Pclass

Esta feature es relevante al analizar la supervivencia ya que captura la categoría del boleto que compró el pasajero. En la siguiente gráfica se observa una clara correlación entre la clase del boleto y la supervivencia, donde se destaca que la proporción de supervivencia para clase baja es mucho menor que la de las clases altas.

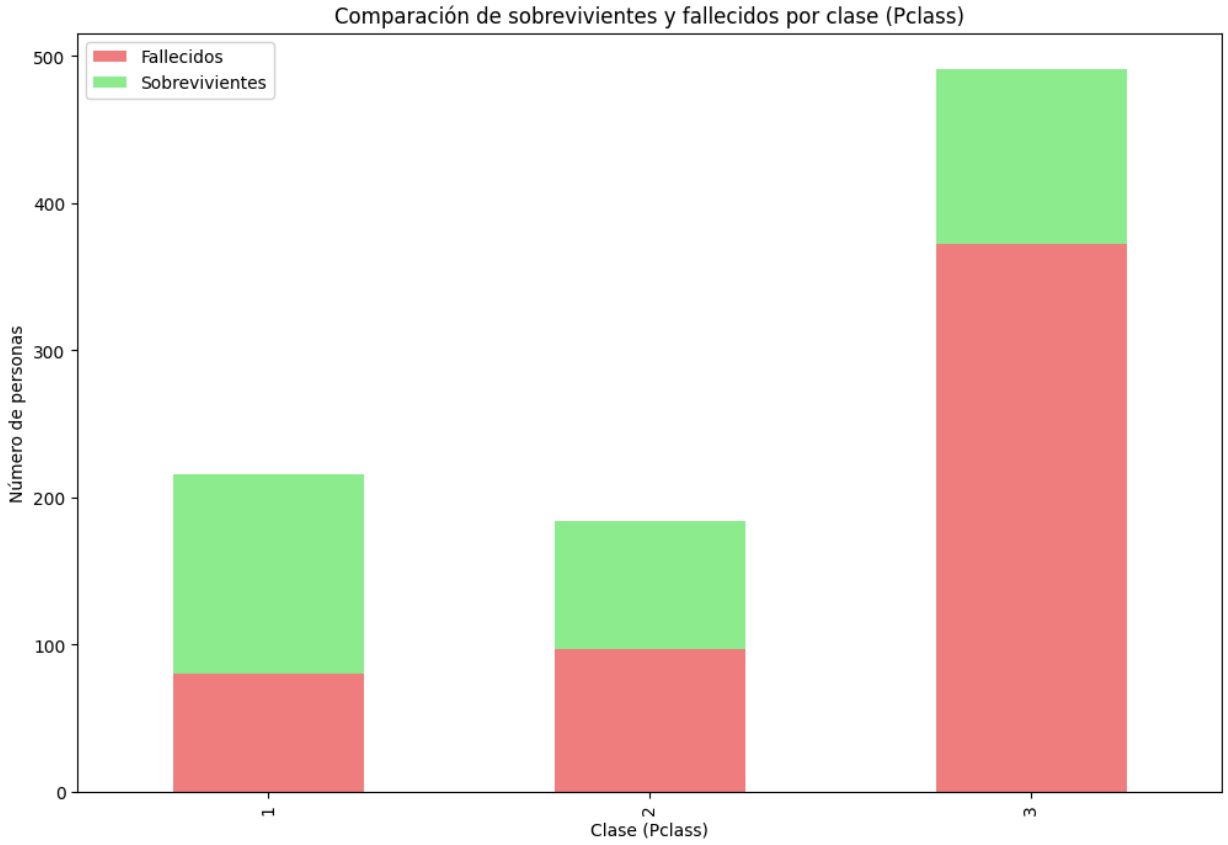

Figura 7. Comparación de sobrevivientes y fallecidos por clase (Pclass)

4.3.3. SibSp

Con el fin de determinar si esta feature sería de utilidad para predecir la supervivencia de una persona a bordo del Titanic, se realizó una gráfica de correlación entre SibSp y Survived, donde se encontró una serie de relaciones interesantes entre diferentes datos. Lo anterior permitió establecer patrones de supervivencia con base en la cantidad de hermanos/cónyuges a bordo.

Para esta feature, se aplicó la misma técnica que en el feature de embarked, la cual era determinar que la relación no se debiera a otro feature, por lo que también se realizó la comparación por sexo obteniendo los mismos resultados que en el antes mencioado.

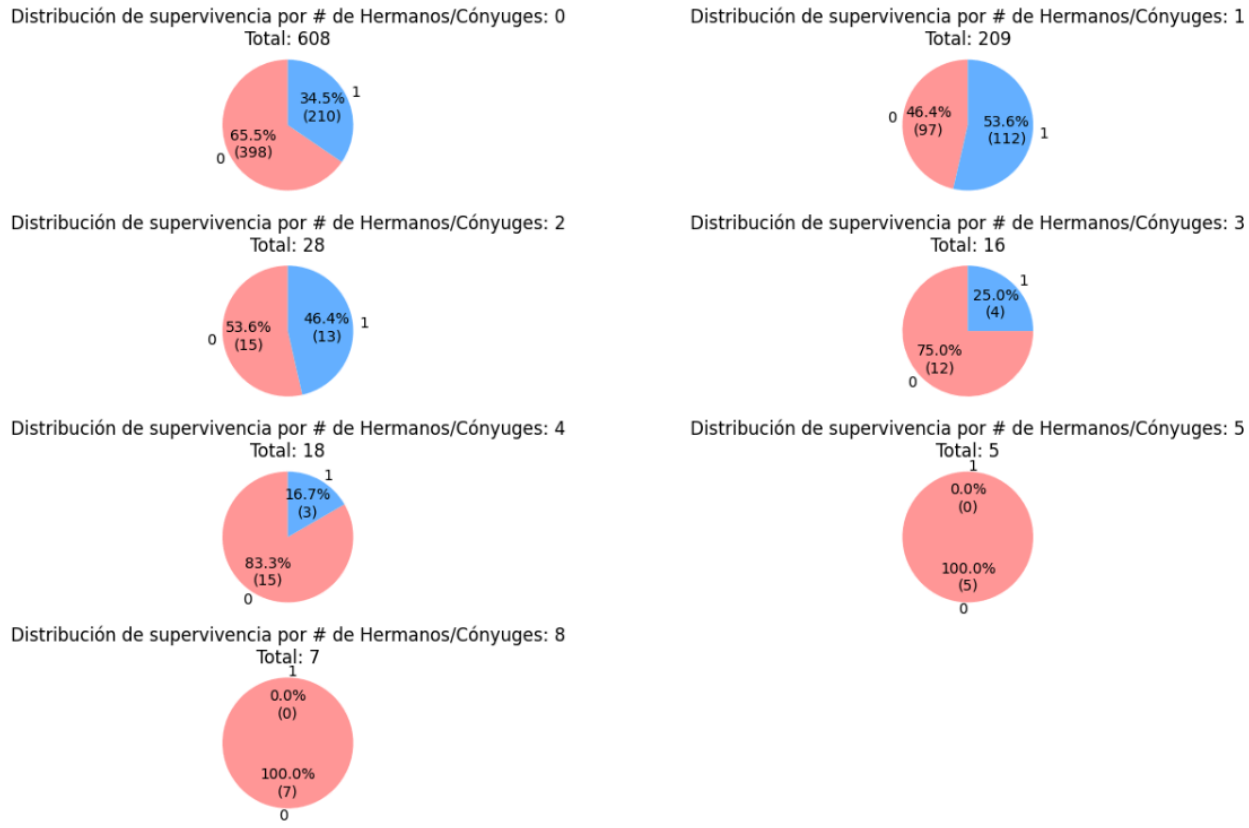

Figura 8. Distribución de supervivencia por número de hermanos / cónyuges

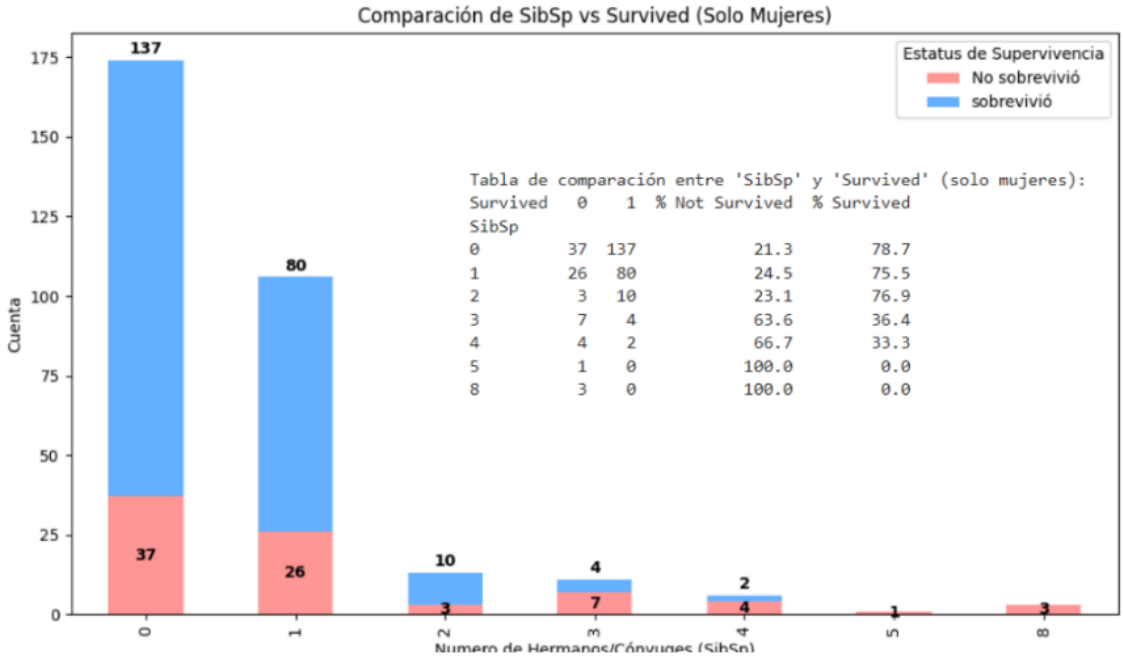

Figura 9. Comparación de SibSp vs Survived para mujeres

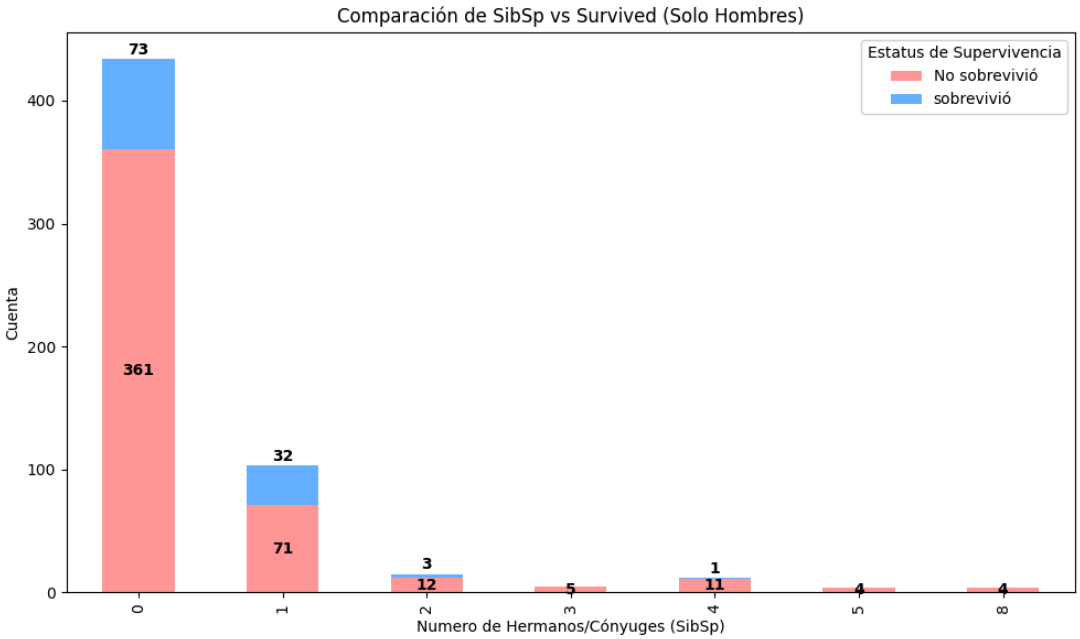

Figura 10. Comparación de SibSp vs Survived para hombres

4.3.4. Parch

Finalmente, para determinar si la feature de Parch sería de utilidad para predecir la supervivencia de una persona a bordo del Titanic, se realizó, una vez más, una gráfica de correlación entre Parch y Survived. Se encontró una serie de relaciones interesantes entre diferentes datos, permitiendo encontrar patrones de supervivencia en base a la cantidad de hijos/padres a bordo, Similar para la feature de embarked y SibSp, se volvió a aplicar la técnica que permite determinar que la relación no se deba a otro feature, por lo que también se realizó la comparación por sexo obteniendo los mismos resultados que en los antes mencionados.

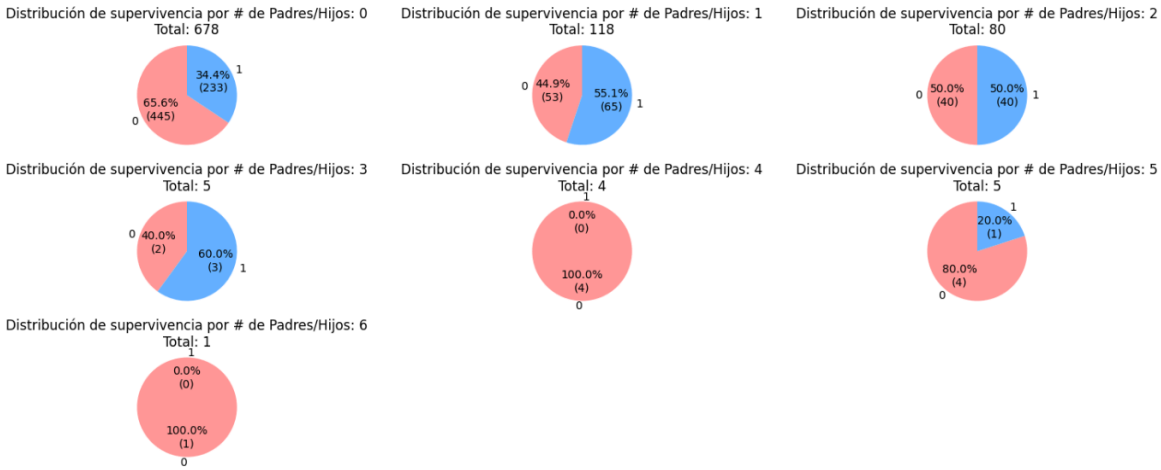

Figura 11. Distribución de supervivencia por número de padres / hijos

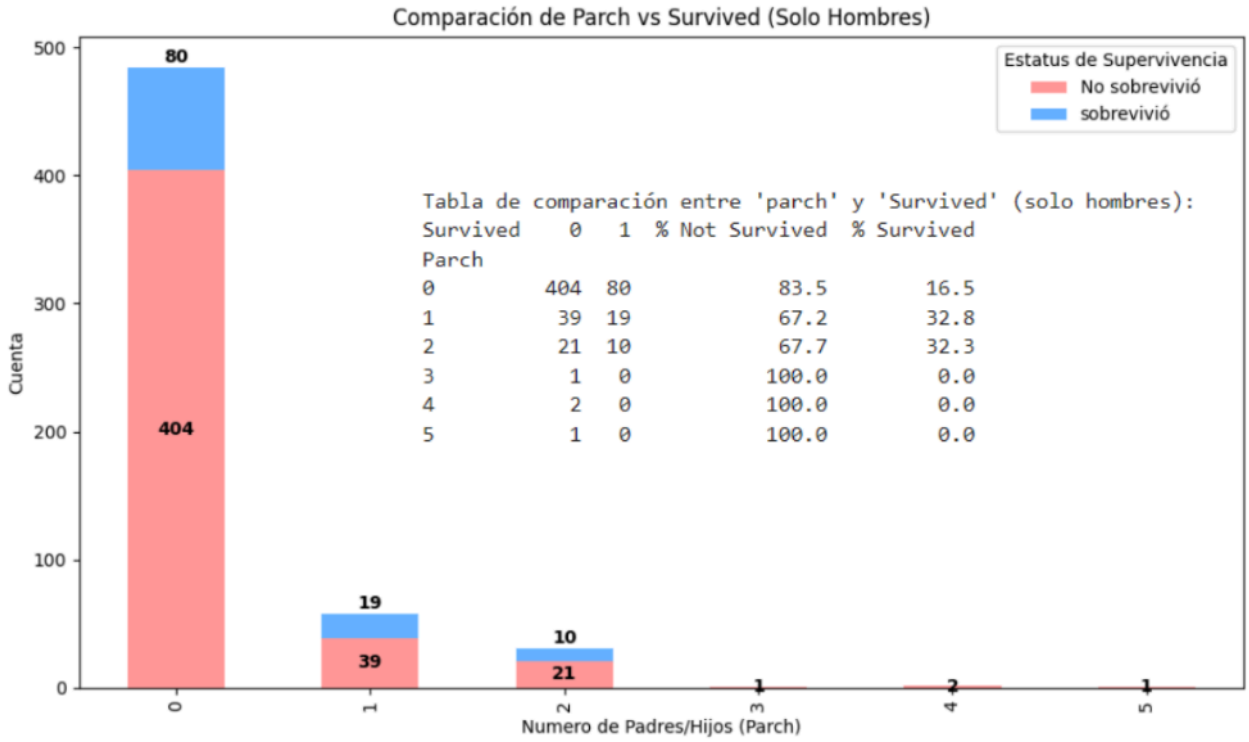

Figura 12. Comparación de Parch vs Survived para hombres

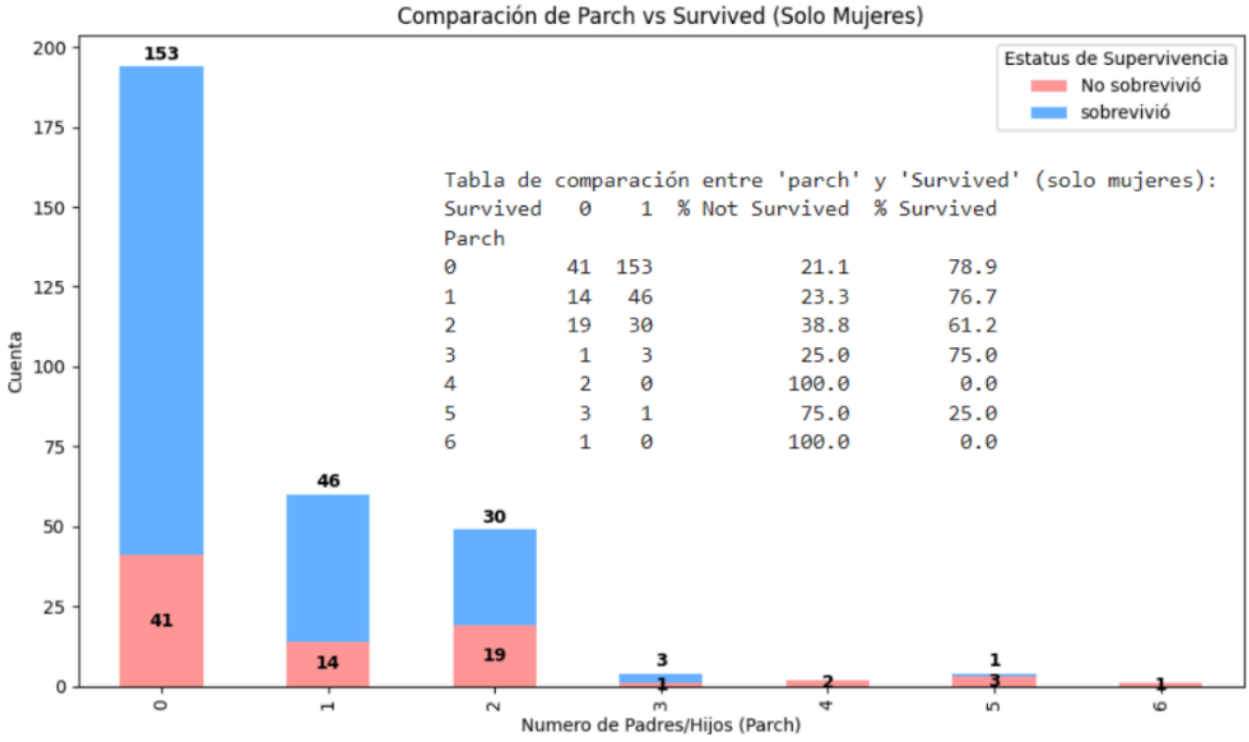

Figura 13. Comparación de Parch vs Survived para mujeres

5. Dataset Final

Después de tomar las decisiones explicadas en la sección anterior, el dataset limpio consta de seis features: ‘Pclass’, ‘Sex’, ‘Age’, ‘SibSp’, ‘Parch’ y ‘Embarked’; y ‘Survived’ como label. Todos los datos son numéricos, no hay valores nulos (se llenaron los datos faltantes como se explicó en la subsección 4.1) y se tiene un total de 891 instancias.

Así mismo, se transformaron los datos de ‘Sex’ (male: 1 y female: 0) y ‘Embarked’ (Southampton: 0, Cherbourg: 1 y Queenstown: 2) a valores numéricos para poder comparar y trabajar con todas las features por igual.

6. Referencias

Barrera, P. (2022, June 9). *Las probabilidades de sobrevivir en el Titanic, según la Ciencia de los Datos*. Actualidad UVG. Retrieved August 19, 2024, from https://noticias.uvg.edu.gt/probabilidades-sobrevivir-titanic-ciencia-de-los-datos-power-bi/

Brewster, H., & Coulter, L. (1999). *Tout ce que vous avez toujours voulu savoir sur le "Titanic"*. Ed. Glénat.

Cukierski, W. (2012). *Titanic - Machine Learning from Disaster*. Kaggle. Retrieved August 16, 2024, from https://www.kaggle.com/competitions/titanic