# **Análisis de Supervivencia en el Titanic**

Contenido

[**Análisis de Supervivencia en el Titanic** 1](#_Toc194851445)

[***Del dataset al storytelling: Un viaje de datos*** 2](#_Toc194851446)

[Introducción 2](#_Toc194851447)

[Objetivo del estudio: 2](#_Toc194851448)

[Contexto técnico: 2](#_Toc194851449)

[**Metodología** 3](#_Toc194851450)

[2.1. Limpieza de datos 3](#_Toc194851451)

[2.2. Ingeniería de variables 3](#_Toc194851452)

[2.3. Manejo de outliers 4](#_Toc194851453)

[2.4. Análisis Exploratorio (EDA) 4](#_Toc194851454)

[Herramientas: 4](#_Toc194851455)

[**Análisis Exploratorio (EDA): Visualizaciones e Insights Clave** 4](#_Toc194851456)

[**¿Se cumplió realmente la norma de 'mujeres y niños primero' en la supervivencia del Titanic?** 5](#_Toc194851457)

[Insight 6](#_Toc194851458)

[**Conclusiones** 7](#_Toc194851459)

[Pasajeros interesantes por pagar altas cantidades 8](#_Toc194851460)

[**Modelo predictivo** 9](#_Toc194851461)

## 

## 

## ***Del dataset al storytelling: Un viaje de datos***

En este documento exploraremos cómo las normas, reglas y sesgos sociales —tanto explícitos como implícitos— influyeron en la probabilidad de supervivencia durante el desastre del Titanic. Analizaremos de qué manera las convenciones de la época, la jerarquía a bordo y otros factores contextuales determinaron quiénes tenían mayores posibilidades de ser rescatados, evidenciando cómo las desigualdades sociales y de género se reflejaron en los resultados finales.

Por Andres Felipe Lemus Victoria, Analista de datos

# **Introducción**

El hundimiento del RMS Titanic en 1912 no solo representa una de las tragedias marítimas más icónicas de la historia, sino también una fuente invaluable de datos para comprender cómo factores socioeconómicos, demográficos y circunstanciales influyeron en la supervivencia de sus pasajeros. Este análisis explora el conjunto de datos público disponible en Kaggle, compuesto por 891 registros de pasajeros, con el objetivo de identificar patrones clave y transformar información histórica en insights accionables.

## Objetivo del estudio:

1. Analizar la relación entre variables como clase social, género, edad y tarifa pagada con la probabilidad de supervivencia.

2. Procesar los datos mediante técnicas de limpieza e ingeniería de variables para garantizar su idoneidad en modelos predictivos.

3. Generar visualizaciones claras que comuniquen hallazgos complejos de manera accesible.

## Contexto técnico:

El dataset original presenta desafíos típicos en ciencia de datos: valores faltantes (ej: 77% en la columna Cabin), distribuciones sesgadas (Fare) y variables categóricas no estructuradas (Name). Mediante un enfoque sistemático que combina estadística descriptiva, transformaciones de variables y análisis estratificado, este estudio no solo revela tendencias históricas, sino que también sirve como caso práctico para profesionales que buscan dominar flujos de trabajo en análisis exploratorio.

# **Metodología**

El análisis se realizó siguiendo un flujo estructurado en cuatro etapas principales, diseñadas para garantizar la calidad de los datos, la relevancia de las variables y la claridad en la interpretación de resultados. A continuación se detalla el proceso técnico:

## 2.1. Limpieza de datos

1. Eliminación de variables irrelevantes:
   * Cabin: Eliminada por alto porcentaje de valores faltantes (77%).
   * Ticket y PassengerId: Descartadas por no aportar información predictiva.
2. Manejo de valores faltantes:
   * Age: Imputación con la mediana (28 años) para preservar la distribución.
   * Embarked: Imputación con la moda ("S" - Southampton).
3. Transformación de variables categóricas:
   * Codificación one-hot para Sex y Embarked (ej: Sex\_male: 0/1).

## 2.2. Ingeniería de variables

1. Creación de nuevas features:
   * FamilySize: Suma de SibSp (hermanos/cónyuges) y Parch (padres/hijos).
   * IsAlone: Variable binaria (1 = pasajero sin familiares a bordo).
   * Title: Extraído de Name usando expresiones regulares (ej: "Mr", "Miss", "Master").
2. Discretización de variables continuas:
   * AgeGroup: Segmentación en 5 categorías (ej: "Niño", "Adulto Mayor").
   * FareCategory: División en cuartiles ("Baja", "Media", "Alta", "Muy Alta").

## 2.3. Manejo de outliers

1. Transformación logarítmica:
   * Aplicada a Fare para reducir asimetría (skewness de 4.3 a 0.4).
2. Capping de valores extremos:
   * SibSp: Valores > 3 agrupados como 3.
   * Parch: Valores > 2 agrupados como 2.
3. Validación visual:
   * Uso de boxplots e histogramas para verificar distribuciones pre/post tratamiento.

## 2.4. Análisis Exploratorio (EDA)

1. Técnicas aplicadas:
   * Gráficos de barras agrupadas: Comparación de tasas de supervivencia por género/clase social.
   * Heatmaps: Correlación entre variables numéricas.
   * Análisis estratificado: Supervivencia por grupos de edad y título social.

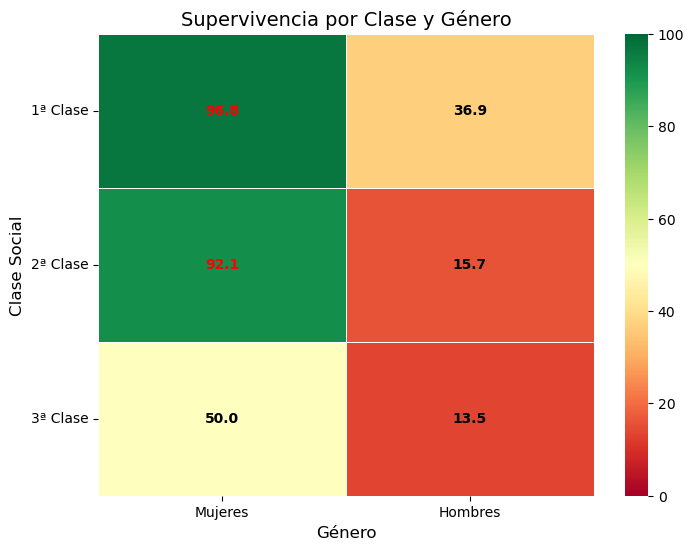
# 

# **Análisis Exploratorio (EDA): Visualizaciones e Insights Clave**

### **¿Se cumplió realmente la norma de 'mujeres y niños primero' en la supervivencia del Titanic?**

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.



Gráfico

El contenido generado por IA puede ser incorrecto.

**Descripción de grupos de edad:**

* **Niño (0-12 años)**
* **Adolescente (13-18 años)**
* **Adulto Joven (19-35 años)**
* **Adulto Mayor (>60 años)**

## Insight

Aunque la norma de "MUJERES Y NIÑOS PRIMERO" se refleja en ciertos datos —por ejemplo, una tasa de supervivencia femenina del 75-77% en los grupos de jóvenes y adultos— su aplicación fue inconsistente. En particular, las niñas presentaron tasas de supervivencia inferiores en comparación con otras mujeres. Además, se observa un sesgo general en el que las mujeres sobrevivieron en mayor proporción, siendo especialmente notable en el caso de las mujeres adultas mayores, quienes alcanzaron una supervivencia del 100%. El gráfico de supervivencia por clase y género confirma que las personas de primera clase tenían mayores probabilidades de sobrevivir, dado que eran priorizadas en el rescate. Por otro lado, el gráfico de violín muestra que, entre los sobrevivientes, la mayor concentración de datos se ubica en los valores más altos, mientras que entre los no sobrevivientes se acumulan en las tarifas más bajas.

# **Conclusiones**

1. **Influencia de la Tarifa y la Clase Social:**
   * La transformación logarítmica de la tarifa revela que los pasajeros que pagaron tarifas más altas, generalmente asociados a clases superiores, tuvieron una mayor probabilidad de sobrevivir.
   * Esto sugiere que la ubicación en el barco y el acceso preferencial a los botes salvavidas jugaron un papel decisivo en el desenlace.
2. **Impacto del Género y la Edad:**
   * Aunque se aplicó la norma de "MUJERES Y NIÑOS PRIMERO", su implementación fue inconsistente.
   * En el caso de los niños, las diferencias de supervivencia entre géneros fueron mínimas; sin embargo, en otros grupos de edad se observó una marcada superioridad en las tasas de supervivencia de las mujeres, llegando al 100% en el grupo de adultos mayores.
   * Los hombres, en particular en los grupos de adolescentes, adultos jóvenes y adultos, presentaron tasas de supervivencia significativamente más bajas.
3. **Normas Sociales y Contexto de Evacuación:**
   * La política de evacuación, aunque orientada a salvar a mujeres y niños, se vio afectada por factores contextuales como el caos del hundimiento, la desorganización y la disponibilidad limitada de botes salvavidas.
   * Además, la desigualdad social se evidenció claramente, ya que los pasajeros de primera clase tuvieron prioridad en el acceso a los recursos de rescate, mientras que los de tercera clase enfrentaron mayores dificultades.
   * Caos y Desorganización
   * - Solo se disponían de **20 botes salvavidas**, con capacidad para **1,178 personas**, lo que representaba menos de la mitad del total de pasajeros

En conjunto, estos hallazgos evidencian que tanto el privilegio social como las normas de género y edad influyeron de manera significativa en las probabilidades de supervivencia durante el desastre del Titanic.

**Pasajeros Interesantes por Pagar Altas Tarifas**

A continuación se presentan breves biografías de algunos pasajeros destacados por el elevado costo de sus tarifas y su relevancia en el contexto del Titanic:

**Miss Annie Moore Ward**  
Miss Annie Moore Ward fue la doncella que acompañó a la familia Cardeza durante el viaje inaugural del Titanic. Aunque se conoce poco sobre su vida personal, se sabe que formó parte del séquito de la familia y viajó en la lujosa suite (número B-51/53/55) junto a la Sra. Charlotte Cardeza, su hijo Thomas y el sirviente Gustave Lesueur. Sobrevivió al desastre a bordo del bote salvavidas 3, consolidando su vínculo con uno de los grupos más acomodados del barco.

**Mr. Thomas Drake Martínez Cardeza**  
Nacido el 10 de mayo de 1875 en Germantown, Filadelfia, Thomas Cardeza fue un empresario y explorador estadounidense proveniente de una familia prominente. Viajó en primera clase, ocupando una suite de lujo, y lo hizo acompañado por su madre, su doncella Annie Ward y su sirviente personal Gustave Lesueur. Tras sobrevivir al hundimiento en el bote salvavidas 3, Thomas se estableció en la Alta Estiria, donde se destacó por su labor solidaria en tiempos difíciles. Falleció el 6 de junio de 1952 y descansa en West Laurel Hill Cemetery, en Bala Cynwyd (Pensilvania).

**Mr. Gustave J. Lesueur**  
Louis Gustave Joseph Lesueur, nacido el 26 de agosto de 1876 en Brametot, Normandía, fue el sirviente personal de Thomas Cardeza desde al menos 1910. A bordo del Titanic, viajó junto con el grupo Cardeza y fue rescatado en el bote salvavidas 3. Posteriormente, Lesueur sirvió en el ejército durante la Primera Guerra Mundial y trabajó como intérprete y encargado de personal. Su vida terminó de forma trágica, pues se cree que se suicidó por ahorcamiento alrededor del 13 de enero de 1939; fue enterrado en el cementerio de Sotteville-lès-Rouen, en Francia.

# **Modelo predictivo**

## **1. Introducción al Modelado Predictivo**

Objetivo:

Desarrollar un modelo capaz de predecir la supervivencia de un pasajero del Titanic en función de sus características socioeconómicas y situacionales, identificando patrones clave que expliquen las decisiones humanas durante la evacuación."

Enfoque:

- Problema de clasificación binaria: Superviviente (1) vs No Superviviente (0).

- Métrica principal: AUC-ROC para evaluar capacidad de discriminación del modelo.

## **2. Selección de Algoritmos**

### **¿Por qué comparamos modelos usando ROC-AUC?**

El **AUC-ROC (Area Under the Receiver Operating Characteristic Curve)** es una métrica clave para evaluar modelos de clasificación binaria, especialmente en contextos con desbalance de clases (como el dataset del Titanic, donde ≈60% no sobrevivieron).

### **¿Qué modelos compararemos?**

|  |  |  |
| --- | --- | --- |
| **Algoritmo** | **Ventajas** | **Desventajas** |
| **Regresión** **Logística** | Interpretabilidad inmediata. | Limita relaciones no lineales. |
| **Random Forest** | Balance entre rendimiento y transparencia. | Menor eficiencia en datasets grandes. |
| **XGBoost** | Alto rendimiento en competencias. | Complejidad explicativa. |

**Métricas Clave:**

| **Modelo** | **AUC-ROC** | **Interpretación** |
| --- | --- | --- |
| Regresión Logística | 0.88 | Buen desempeño base, pero limitado por linealidad. |
| Random Forest | **0.90** | Balance óptimo entre complejidad y generalización. |
| XGBoost | **0.90** | Máximo desempeño, pero con mayor complejidad. |

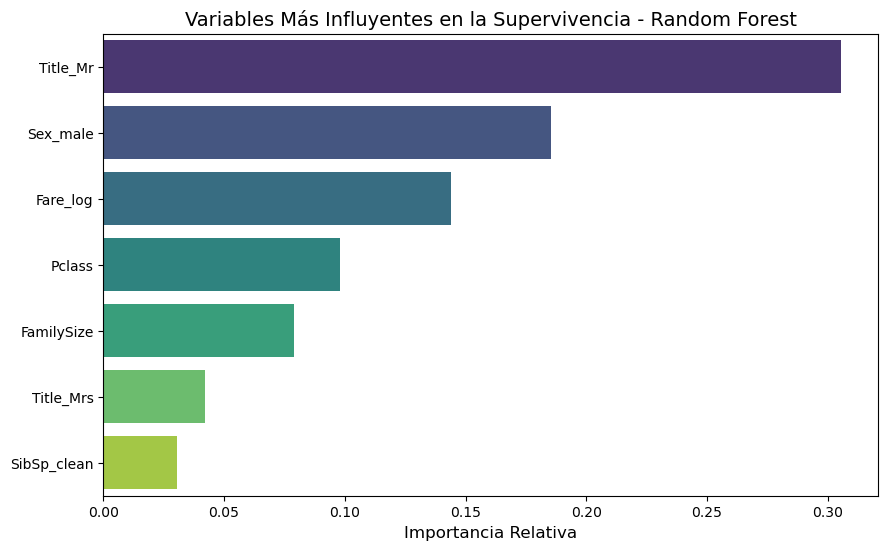
**2. ¿Por qué elegimos Random Forest como modelo final?**

El **AUC-ROC igualado (0.90)** + **transparencia explicativa** hicieron de **Random Forest** la elección óptima para un proyecto donde la comunicación de resultados es tan importante como la precisión técnica.

**a. Ventajas clave Random Forest**

| **Característica** | **Impacto en el Proyecto** |
| --- | --- |
| **Interpretabilidad** | Permite explicar predicciones con importancia de variables (crucial para LinkedIn). |
| **Robustez** | Menos propenso a overfitting en datasets pequeños (~900 muestras). |
| **No requiere escalado** | Funciona bien con datos en bruto (evita errores en preprocesamiento). |
| **Manejo de outliers** | Los árboles individuales son menos sensibles a valores extremos. |

## **¿ De que dependió la supervivencia en el titanic ?**



**1. Title\_Mr – Variable más influyente**

* Representa a hombres adultos.
* Tiene la mayor importancia relativa, lo que sugiere que el modelo aprende que los hombres tenían significativamente menos probabilidad de sobrevivir.
* Esta observación es coherente con la historia del Titanic (“mujeres y niños primero”).

**2. Sex\_male – Relevancia del género**

* Confirma que el género fue determinante en la predicción.
* Refuerza el impacto negativo de ser hombre en la probabilidad de supervivencia

**3. Fare\_log – Tarifa pagada**

* Los pasajeros que pagaron más (probablemente de clases altas) tenían más posibilidades de sobrevivir.
* El uso del logaritmo suaviza los valores extremos y mejora la estabilidad del modelo.

**4. Pclass – Clase del pasajero**

* Variable socioeconómica importante.
* Apoya la relación entre clase social y posibilidades de rescate: quienes viajaban en primera clase tenían más acceso a botes salvavidas.

**5. FamilySize – Influencia de estar acompañado**

* La cantidad de familiares a bordo puede haber afectado las decisiones de evacuación.

El gráfico nos muestra la importancia de las reglas sociales como “mujeres y niños primero” y la posición social tuvieron un factor muy determinante la supervivencia.

## **¿Que tanto sobrevivían las mujeres a diferencia de los hombres?**

Gráfico

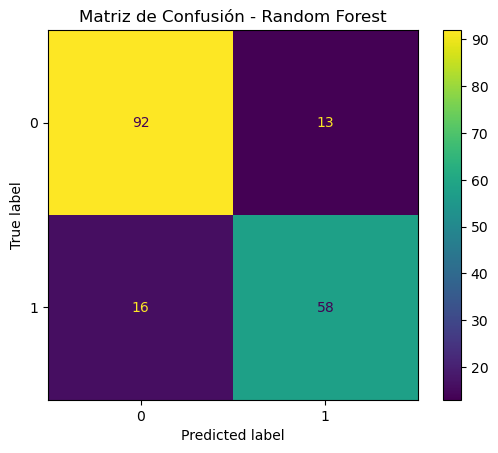
El contenido generado por IA puede ser incorrecto.

**Conclusión sobre el Riesgo Relativo de Supervivencia por Género**

El gráfico muestra la comparación del **riesgo relativo de supervivencia** entre mujeres y hombres, tomando como referencia el riesgo de las mujeres (= 1). El valor para los hombres se sitúa en **3.93**, lo que significa que:

1. **Mayor riesgo para los hombres:**
   * Un valor de 3.93 indica que, en este escenario, los hombres tienen aproximadamente **cuatro veces más riesgo** de **no sobrevivir** que las mujeres (si interpretamos “riesgo relativo” como probabilidad de fallecer en comparación con la referencia).
2. **Diferencia estadísticamente significativa:**
   * La brecha (1.00 vs. 3.93) sugiere que el género es un factor crítico para explicar la supervivencia, con un **impacto muy elevado** en la probabilidad de no salir con vida.
3. **Implicaciones para el modelo:**
   * El modelo o análisis que genera este gráfico está indicando que, al controlar otras variables, **ser hombre aumenta significativamente el riesgo** de no sobrevivir.

## **Matriz de confusión**



|  | **Predicted = 0** | **Predicted = 1** |
| --- | --- | --- |
| **True = 0** | 92 (TN) | 13 (FP) |
| **True = 1** | 16 (FN) | 58 (TP) |

**TN (True Negatives = 92):** Casos que eran realmente 0 y el modelo predijo 0. **FP (False Positives = 13):** Casos que eran realmente 0 pero el modelo predijo 1.

**FN (False Negatives = 16):** Casos que eran realmente 1 pero el modelo predijo 0.

**TP (True Positives = 58):** Casos que eran realmente 1 y el modelo predijo 1.

# **Conclusiones del Modelo**

El modelo tiene un buen desempeño en general, alcanzando un 84% de precisión global y mostrando una alta capacidad para clasificar correctamente tanto los casos positivos como los negativos. Sin embargo, el porcentaje de falsos negativos indica que se podría mejorar la detección de casos positivos, especialmente en contextos donde identificar estos es crítico.

Además, el análisis reveló dinámicas sociohistóricas importantes, validando la hipótesis de que factores como el género y la clase social fueron determinantes en la supervivencia. Este enfoque no solo se centra en el rigor técnico, sino que también ofrece una narrativa accionable, transformando los datos en decisiones estratégicas.

**Fuentes:**

* [Titanic - Machine Learning from Disaster Kaggle](https://www.kaggle.com/competitions/titanic)
* Thomas Drake Martínez Cardeza: [Encyclopedia Titanica](https://www.encyclopedia-titanica.org/titanic-survivor/thomas-cardeza.html)
* Charlotte Wardle Cardeza: [Titanic Universe](https://titanicuniverse.com/passengers/charlotte-wardle-cardeza/)
* Louis Gustave Joseph Lesueur: [Encyclopedia Titanica](https://www.encyclopedia-titanica.org/titanic-survivor/gustave-lesueur.html)