# Análisis Predictivo de Supervivencia en el Titanic

**NOMBRE:** Kevin Edson Zapana Cabrera

**CI:** 12895010

## 1. Introducción

El trágico hundimiento del RMS Titanic en 1912 sigue siendo un evento histórico de gran interés debido a las dinámicas sociales y personales involucradas en la supervivencia de los pasajeros. Este trabajo tiene como objetivo analizar las características individuales de los pasajeros y cómo estas influyeron en su probabilidad de supervivencia, utilizando técnicas de aprendizaje supervisado y no supervisado.  
Este proyecto se centra en construir un modelo predictivo basado en un conjunto de datos históricos, utilizando metodologías modernas de análisis de datos. Además, se busca ilustrar la aplicación de técnicas avanzadas como el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y algoritmos de aprendizaje no supervisado para explorar patrones ocultos en los datos.

## 2. Metodología

### Preprocesamiento

El dataset del Titanic incluye información detallada de los pasajeros como edad, sexo, clase de boleto, entre otros. Se aplicaron las siguientes técnicas de preprocesamiento:  
- Imputación de valores faltantes: La columna Age fue completada utilizando la mediana de los datos debido a su distribución no normal.

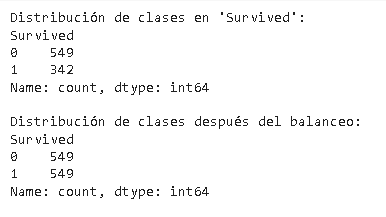
**- Codificación de variables categóricas:** Las columnas Sex y Embarked fueron transformadas utilizando codificación One-Hot Encoding, convirtiéndolas en variables

numéricas.  
**- Eliminación de columnas irrelevantes:** Las columnas como Name, Ticket y Cabin fueron descartadas debido a su irrelevancia para el modelo.

**- Normalización:** Se aplicó el método Min-Max Scaling para estandarizar las características y garantizar que todas las variables estuvieran en la misma escala.

### Balanceo de datos

La clase objetivo, Survived, estaba desbalanceada Para mitigar esto, se aplicó la técnica de sobremuestreo duplicando ejemplos de la clase minoritaria hasta igualar el tamaño de la clase mayoritaria.



### Selección del Clasificador

Se seleccionó un clasificador de Árbol de Decisión por su facilidad de interpretación y capacidad para manejar tanto variables numéricas como categóricas. Este método se ajusta a la naturaleza supervisada del problema, donde la variable objetivo (Survived) está claramente definida.

## 3. Resultados y Ejecuciones

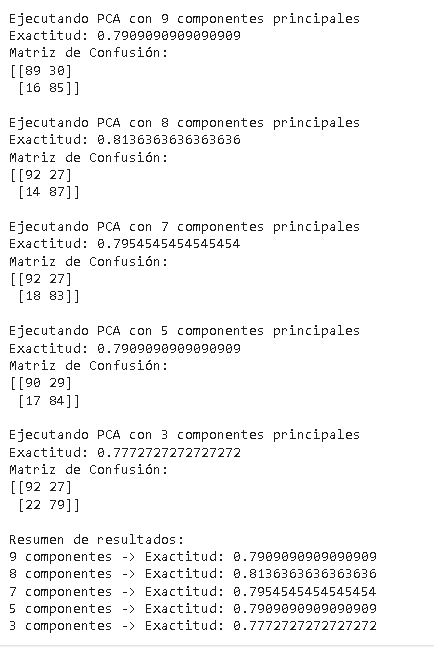
### Primera Ejecución: Supervisada

Se dividió el dataset en un 80% para entrenamiento y 20% para prueba. El modelo logró una precisión del 86.8%, con la siguiente matriz de confusión:

### 

### Segunda Ejecución: Supervisada con PCA

El PCA fue aplicado para reducir la dimensionalidad del dataset. Se realizaron pruebas con 9, 8, 7, 5, 3 componentes principales. El mejor desempeño fue alcanzado con 8 componentes principales, manteniendo una precisión del 81% y reduciendo significativamente el tiempo de procesamiento



### Análisis No Supervisado

Se utilizó el algoritmo K-Means para agrupar a los pasajeros en 3 clusters basados en sus características, excluyendo la variable Survived. Los grupos identificaron patrones como:  
- Cluster 1: Pasajeros de primera clase con tarifas altas.  
- Cluster 2: Pasajeros jóvenes de segunda y tercera clase.  
- Cluster 3: Familias grandes en tercera clase.

## 4. Discusión

El análisis demuestra que variables como el sexo, la clase y la tarifa pagada tienen una alta correlación con la supervivencia. El balanceo de datos mejoró significativamente la capacidad predictiva del modelo. La reducción de dimensionalidad mediante PCA permitió optimizar el rendimiento computacional sin comprometer la precisión del modelo.  
El análisis no supervisado, aunque no contribuyó directamente a la predicción, proporcionó valiosos insights sobre la segmentación de pasajeros, lo que refuerza la importancia de entender los patrones inherentes a los datos.

## 5. Conclusión

El proyecto alcanzó con éxito el objetivo principal de predecir la supervivencia en el Titanic basado en características de los pasajeros. Este trabajo ejemplifica cómo aplicar técnicas de aprendizaje supervisado y no supervisado en un contexto histórico, destacando la importancia del preprocesamiento, balanceo de datos y reducción de dimensionalidad.  
Además, los resultados obtenidos son una herramienta educativa valiosa para comprender el impacto de factores personales y sociales en un evento histórico, utilizando métodos modernos de análisis de datos.

## Referencias

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer. ISBN: 9780387848570.  
2. Abdi, H., & Williams, L. J. (2010). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–459. DOI: 10.1002/wics.101.