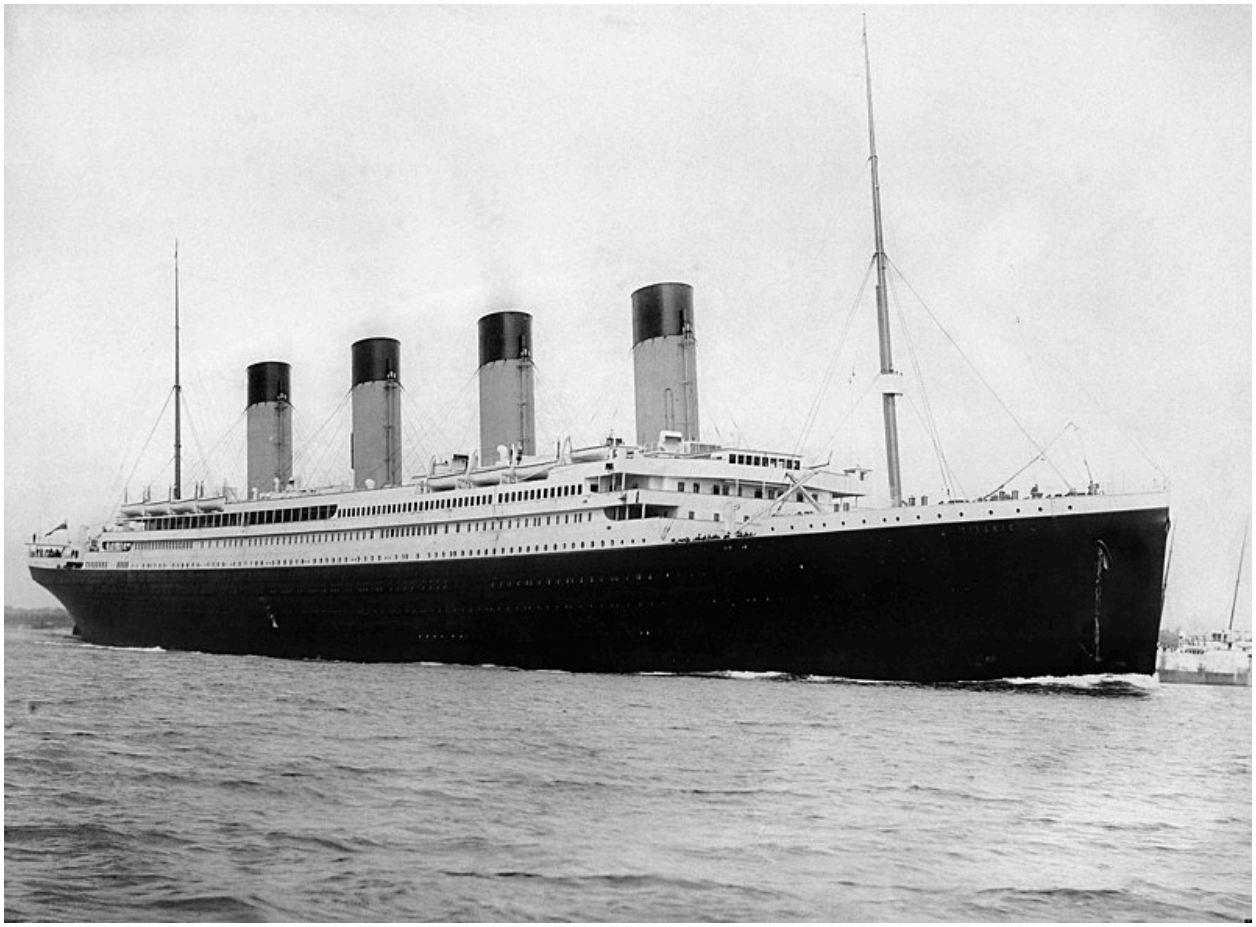
# TITLE PAGE



**Project Title: Statistical Analysis and Machine Learning for Titanic Survival Prediction**

**Bradley Williams**
**Date: 8/18/2024**

# TABLE OF CONTENT

# Introduction

**This project will do an exploratory study of the Titanic dataset, to identify the important elements that determine the passengers' survival. The major goal is to use data analysis and machine learning approaches to create predictive models that can identify the features most strongly associated with survival.**

**The project adopts a comprehensive approach, beginning with thorough data preprocessing to address missing values and transform categorical variables into numerical representations. Following this, a neural network model is constructed, trained, and evaluated to assess its accuracy in predicting passenger survival, offering a data-driven insight into the factors that influence survival outcomes.**

**The project's objectives are to:**

- **Uncover patterns and relationships within the dataset that provide insights into the demographics and circumstances of the passengers.**
- **Develop a predictive model using machine learning techniques to identify the key attributes linked to survival.**

# Dataset Exploration

**The Titanic dataset provides a significant glimpse into the demographics and circumstances of the passengers aboard the ill-fated ship. It encompasses a range of features that offer insights into the passenger's group identities and their chances of survival.**

**Key Features and Their Distributions:**

- **Survival: The binary target variable, indicates whether a passenger survived (1) or not (0). The dataset reveals a stark reality, with a majority of passengers not surviving the disaster.**

- **Passenger Class (Pclass): A categorical feature representing the socio-economic status of passengers, ranging from 1st class - to 3rd class. The distribution of passengers across classes highlights the social stratification on board, with a larger proportion belonging to the 3rd class.**

- **Sex (Sex): A categorical feature distinguishing between male and female passengers. The dataset shows a clear gender imbalance, with a significantly higher number of male passengers.**

- **Age (Age): number-based feature representing the age of passengers. The age distribution reveals a diverse range, from infants to elderly individuals, with a concentration of adults in their prime. *The missing values in the 'Age' column were imputed with the median age of all passengers. This approach helps to maintain the overall distribution of ages while avoiding the introduction of extreme or unrealistic values.***

- **Number of Siblings/Spouses Aboard (SibSp): A numerical feature indicating the count of siblings or spouses accompanying a**

**passenger. The distribution underscores the prevalence of individuals traveling alone or with few family members.**

- **Number of Parents/Children Aboard: A number-based feature representing the count of parents or children traveling with a passenger. Similar to 'SibSp', the distribution emphasizes the predominance of solo travelers or those with limited family accompaniment.**

- **Passenger Fare (Fare): A numerical feature reflecting the ticket price paid by passengers. The fare distribution exhibits a wide range, reflecting the varying levels of luxury and accommodation choices available on the Titanic.**

- **Port of Embarkation (Embarked): A categorical feature denoting the port where passengers boarded the ship (C = Cherbourg, Q = Queenstown, S = Southampton). The distribution reveals that the majority of passengers embarked from Southampton. *The missing values in the 'Embarked' column, which denotes the port of embarkation, were filled with the mode (most frequent value) of the column. This ensures that the missing values are replaced with the most likely embarkation point based on the existing data.***

# Data Preprocessing

**The project tackled several data preprocessing challenges to prepare the Titanic dataset for analysis and modeling.**

**Encoding Categorical Variables:**

**The categorical features 'Sex' and 'Embarked' were converted from text labels (e.g., 'male', 'female') into number representations using LabelEncoder. This transformation is essential for machine learning algorithms that primarily operate on the data. Missing Values: The 'Age' column, which contained missing values, was addressed by imputing the median age. Similarly, missing entries in the 'Embarked' column were filled with the most frequent embarkation point.**

**Normalization/Standardization:**

**The numerical features in the dataset were standardized using StandardScaler. This process scales the features to have a mean of 0 and a standard deviation of 1, ensuring that all features contribute equally to the model's learning process, regardless of their original scales**

**Feature Selection:**

**The project specifically mentions dropping several columns that were deemed irrelevant for the prediction task, including 'deck', 'embark_town', 'alive', 'class', 'who', 'alone', and 'adult_male'. This careful selection of features helps to streamline the model and potentially improve its performance by focusing on the most informative attributes.**

# Model evaluation

**Evaluation metric**

**The primary evaluation metric used is accuracy, which measures the overall proportion of correct predictions. The code also generates a classification report that includes precision, recall, and F1-score for both classes (survived and not survived)**

**.Comparison with Baseline Models**

**The baselines through the analysis of survival probabilities are based on features like sex and passenger class. The high survival probability for women (77%) and the low survival probability for men (19%) where survival is predicted solely based on gender. The accuracy of such a baseline is calculated by predicting all females as survivors and all males as non-survivors and then comparing these predictions to the actual outcomes.**
**Similarly, the varying survival probabilities across different passenger c lasses (1st class: 63%, 2nd class: 47%, 3rd class: 24%) where survival is predicted based on passenger class.**
**The accuracy of this baseline is evaluated by assigning survival outcomes based on the majority class within each passenger class and then comparing these predictions to the actual outcomes.**

**Interpretation of Results**

**The provided code snippets focus on evaluating the models' performance on the test set using accuracy and the classification report. The accuracy values and the precision, recall, and F1 scores for each class can be interpreted as follows:**

- **Accuracy: The overall percentage of correct predictions made by the model on the test set.**

- **Precision (for survived class): Out of all passengers predicted to survive, how many survived?**
- **Recall (for survived class): Out of all passengers who survived, how many were correctly predicted by the model?**
- **F1-score (for survived class): A harmonic mean of precision and recall, providing a balanced measure of the model's ability to predict survivors.**
- **Precision, Recall, and F1-score (for not survived class): Similar interpretations apply, but for the class of passengers who did not survive.**

***Neural Network:***

- ***0 (Not Survived): 0.75***
- ***1 (Survived): 0.64***

***Decision Tree:***

- ***0 (Not Survived): 0.68***
- ***1 (Survived): 0.68***

# Feature importance

**Most Important Features:**

- **Ticket Price (Fare): This was the biggest clue for predicting survival. Generally, people who paid more for their tickets had a better chance of surviving.**
- **Gender (Sex): Being male or female made a big difference. Women were much more likely to survive than men.**
- **Age: How old someone was played a big role. Children and younger people had better odds of surviving.**
- **Passenger Class (Pclass): Whether someone was in first, second, or third class mattered a lot. First-class passengers had the best chance of survival.**
- **Family Size: The number of family members traveling together affected survival chances. Small family groups often did better than large families or solo travelers.**

**What This Tells Us:**

- **Money Mattered: People who could afford expensive tickets (usually in first class) had better access to lifeboats and were more likely to survive.**
- **Women and Children First: This wasn't just a saying - it was practiced. Women and kids did have a better chance of getting on lifeboats.**
- **Age Was Important: Being young helped your chances of survival.**
- **Class Differences: Where you were on the ship (which depended on your ticket class) affected your chances of getting to a lifeboat.**
- **Family Factor: Having a small family group seemed to help, possibly because it was easier to keep track of everyone in the chaos**

# Tree visualization

**Visual Representation:**

- **Input Data = The process initiates with input data comprising "Gender," "Class," and "Age Group."**
- **Data Preparation: The data undergoes preprocessing, which likely involves encoding categorical variables (like Gender and Class) into numerical formats and potentially scaling numerical features (like Age Group) for consistency.**
- **Generate Predictions: A model (unspecified type) utilizes the prepared data to generate predictions.**
- **Model Prediction: This central node branches into four filters based on "Gender" (Male/Female) and "Class" (Rich/Poor).**
- **Prediction Branches: Each filter leads to a specific prediction output:**
    - **Male Prediction: Predicts survival probability for males.**
    - **Female Prediction: Predicts survival probability for females.**
    - **Rich Prediction: Predicts survival probability for individuals in Class 1 (Rich).**
    - **Poor Prediction: Predicts survival probability for individuals in Class 3 (Poor).**
- **Output = Each prediction branch ultimately yields the "Probability of Survival" for its respective group.**

**Interpretation of Decision Rules:**

**The core decision rule in this visualization is the segregation of predictions based on "Gender" and "Class." The model implicitly assumes that these two factors significantly influence survival probability.**

- **Gender: The model generates separate predictions for males and females, suggesting that their survival rates differ.**
- **Class: Similarly, distinct predictions are made for "Rich" (Class 1) and "Poor" (Class 3) individuals, implying a class-based disparity in survival chances.**

# Results and Conclusion

## Summary of Findings:

- **Survival Rates: The data showed a clear disparity in survival rates between different groups. The overall survival rate was approximately 38%, with a significantly higher proportion of women surviving compared to men. First-class passengers had a much higher survival rate than those in second or third-class.**
- **Factors Affecting Survival: The analysis highlighted the importance of gender and class in determining survival. Women and first-class passengers were more likely to survive, suggesting that they were given priority in the evacuation process. Age also played a role, with children having a relatively high survival rate.**
- **Boarding Location: The port of embarkation also appeared to have a slight impact on survival, with passengers boarding at Cherbourg having a slightly higher survival rate than those boarding at Southampton or Queenstown.**
- **Fare and Age: The analysis of fare and age distributions revealed that most passengers paid lower fares and that the age distribution was skewed towards younger adults. While there was some correlation between higher fares and survival, the relationship was not straightforward, likely due to the confounding influence of class.**

## Insights on Survival Probabilities:

- **Gender Bias: The significant difference in survival rates between men and women underscores the gender bias prevalent at the time, where women and children were prioritized in rescue efforts.**
- **Socioeconomic Disadvantage: The lower survival rates for passengers in second and third class highlight the socioeconomic disparities that existed, with those in lower classes having less access to lifeboats and safety measures.**
- **Limited Resources: The overall low survival rate emphasizes the tragic reality of the Titanic disaster, where the limited number of lifeboats and the chaotic evacuation process resulted in a high loss of life.**

## Challenges and Limitations:

- **Missing Data: The dataset contained missing values, particularly for the 'Age' column, which could have affected the accuracy of some analyses.**
- **Sample Size: While the dataset included a substantial number of passengers, it represents only a portion of the total population on board, limiting the generalizability of the findings.**

# References


**Excel PMT, NPER, PV, NPV, and IRR Functions**
https://youtu.be/2hilrnElkyo?si=O9GPgTLwVVTSKURQ
**Stacey Bolin**

**PV(), NPER(), RATE(), PMT() & FV() - Basic Excel Finance Functions | Corporate Finance 4/8**
**https://www.youtube.com/watch?v=sl9qpUn0z_8**
**Joefessor**

**Titanic Survival Prediction in Python - Machine Learning Project**
**NeuraNine**
**Titanic Survival Prediction in Python - Machine Learning Project**

**TensorFlow 2.0 Complete Course - Python Neural Networks for Beginners Tutorial**
**https://youtu.be/tPYj3fFJGjk?si=KTDnJxcu1P_el5zD**

**Titanic Survival Prediction using Tensorflow Implementation**
**https://youtu.be/GG7rAfqX5IA?si=T-V5TX9ljoHpSTso**