**INTRODUCTION**

The sinking of the Titanic is one of the most infamous shipwrecks in history.
On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic
sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone
onboard, resulting in the death of 1502 out of 2224 passengers and crew.
While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

**DATA DESCRIPTION**

*Features Overview*

| Variable | Definition | Key |
|---|---|---|
| **Survival** | Survival | 0 = No, 1 = Yes |
| **pclass** | Ticket class | 1=1st, 2=2nd, 3=3rd |
| **Sex** | Sex(Gender) | |
| **Age** | Age in years | |
| **sibsp** | # of siblings/ spouses aboard the Titanic | |
| **parch** | # pf parents / children aboard the Titanic | |
| **ticket** | Ticket number | |
| **Cabin** | Cabin number | |
| **Embarked** | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

**1. Passenger Details**

- **Passenger ID (PassengerId)**: Unique identifier for each passenger.
    - Count: 712

    - Mean: 444.41
    - Min: 1
    - Max: 891
    - Std. Dev.: 257.47
- **Name**: Passenger names (712 unique values).
- **Sex**: Gender of the passengers.
    - Categories: Male (459 occurrences), Female (253 occurrences).
- **Age**:
    - Count: 575 (missing for 137 passengers).
    - Mean: 29.81 years
    - Min: 0.42 years
    - Max: 80.00 years
    - Std. Dev.: 14.49 years
    - Percentiles:
        - 25%: 21.00 years
        - 50% (Median): 28.50 years
        - 75%: 39.00 years

**2. Family and Companions**

- **SibSp**: Number of siblings/spouses aboard.
    - Mean: 0.49
    - Min: 0
    - Max: 8
    - Std. Dev.: 1.06
- **Parch**: Number of parents/children aboard.

    - Mean: 0.39
    - Min: 0
    - Max: 6
    - Std. Dev.: 0.84

**3. Ticket and Fare Information**

- **Ticket**: Unique ticket numbers (571 unique values; 6 most frequent tickets shared by multiple passengers).
- **Fare**:
    - Mean: $31.82
    - Min: $0.00
    - Max: $512.33
    - Std. Dev.: $48.06
    - Percentiles:
        - 25%: $7.90
        - 50% (Median): $14.45
        - 75%: $31.00

**4. Cabin and Embarkation**

- **Cabin**: Only 160 records available (127 unique cabins; "G6" is the most frequent, appearing 4 times).
- **Embarked**: Port of embarkation.
    - Categories:
        - Southampton (S): 516 passengers
        - Cherbourg (C): Count not provided
        - Queenstown (Q): Count not provided

**5. Survival Information**

- **Survived**: Survival status (0 = Did not survive, 1 = Survived).
    - Mean survival rate: 38.34%
    - Count: 712
    - Std. Dev.: 48.66%

**Key Observations:**

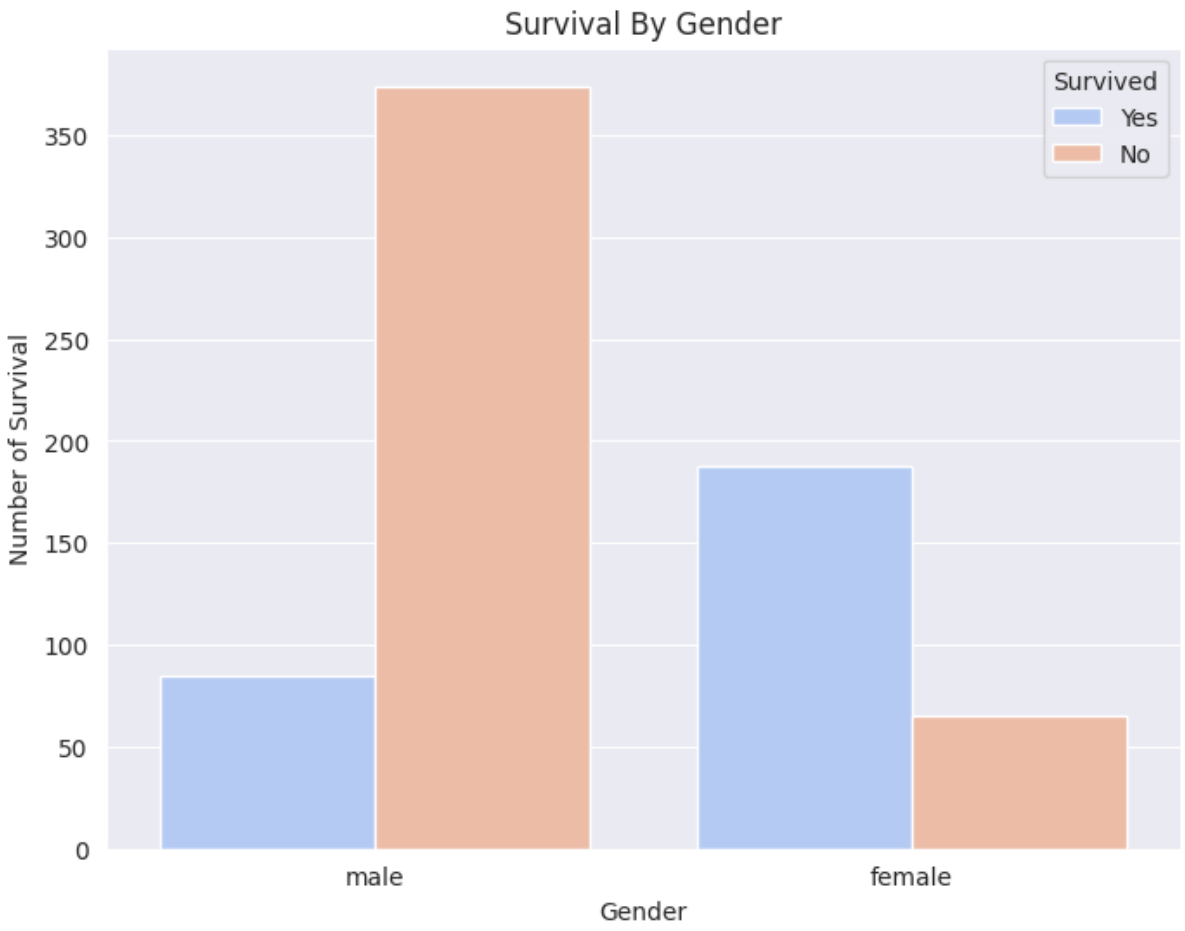
1. **Survival Rate**: Approximately 38% of the passengers survived.
2. **Gender Disparity**: The majority of passengers were male (64.49%).
3. **Age Range**: Ages varied significantly, with the youngest being an infant (0.42 years) and the oldest aged 80 years.
4. **Fare Distribution**: A wide range of ticket prices was observed, with a high fare outlier of $512.33.

**Data Exploration and Visualization**

**Techniques:**

Bar plots and histograms are used to visualize survival trends across gender, age, and class.

**Gender-Based Survival Analysis**




Males:

- The number of males who did not survive (represented by the peach bar) is significantly higher than those who survived (blue bar).
- This indicates that a substantial proportion of male passengers could not survive the disaster.

Females:

- The number of females who survived (blue bar) is notably higher than those who did not survive (peach bar).
- This suggests a considerably higher survival rate among female passengers compared to their male counterparts.

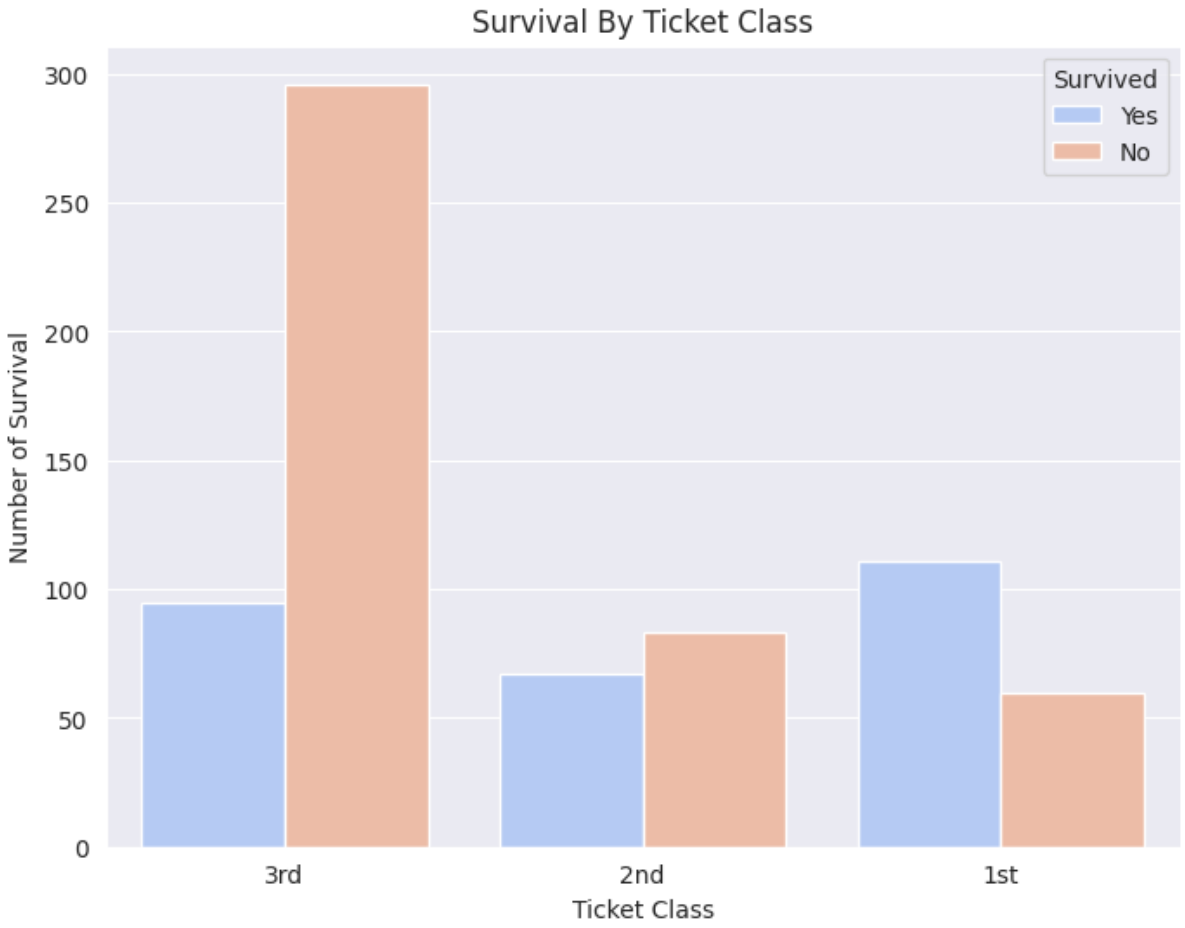
Insights:

- Gender Disparity in Survival Rates: The data reveals a significant disparity in survival outcomes based on gender.

- Female Survival Advantage: A larger proportion of females survived, while males experienced a much lower survival rate.
- Possible Contributing Factors: This disparity may be influenced by evacuation policies or societal norms during emergencies, such as prioritizing women and children for rescue ("women and children first").

The analysis underscores a clear difference in survival outcomes between males and females. Further investigation into the factors contributing to this disparity such as ship evacuation protocols or other socio-cultural influences could provide deeper insights.

**Survival by Ticket Class Analysis**

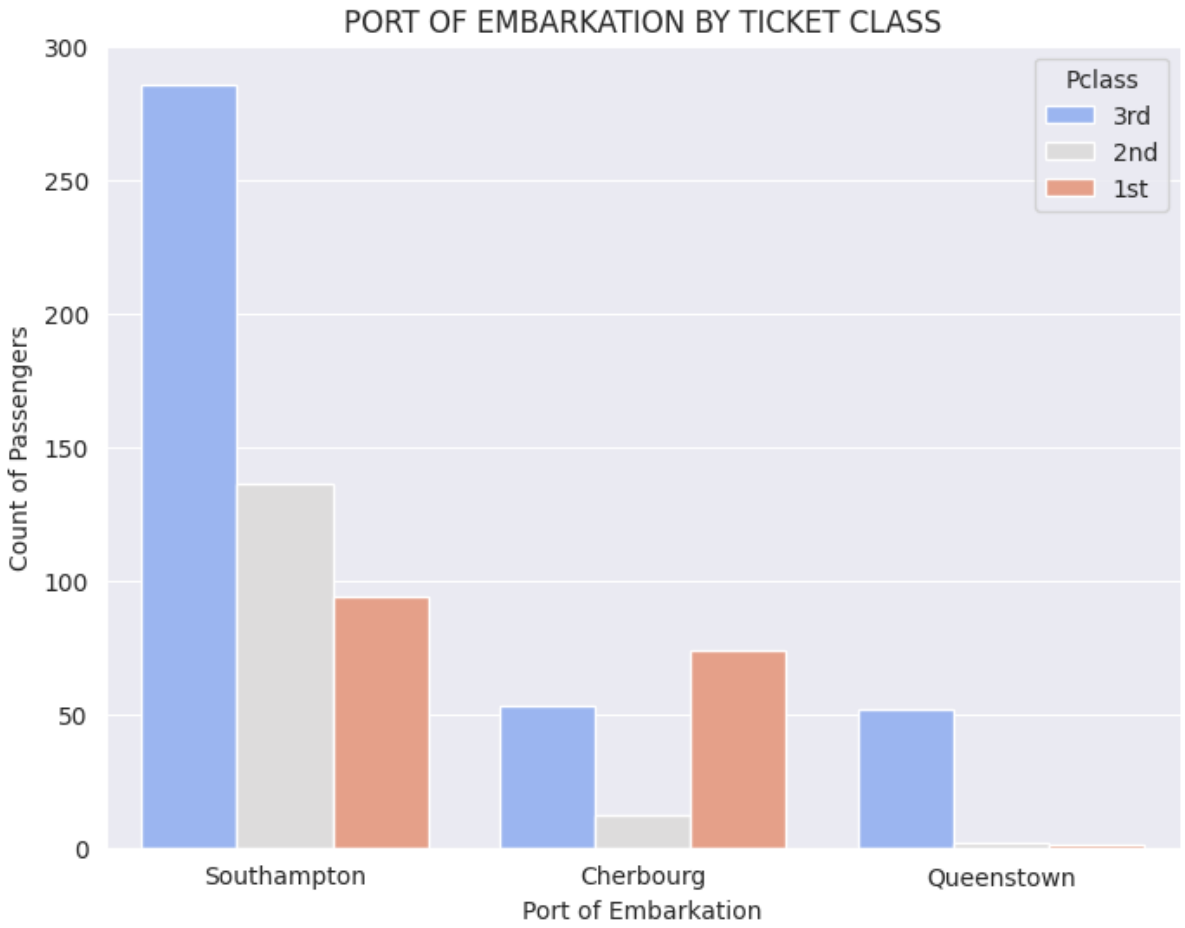


**Insights:**

1. **Ticket Class and Survival Correlation**:
    - Passengers in 1st class experienced the highest survival rates, while those in 3rd class had the lowest.
    - There is a positive correlation between ticket class and survival likelihood.
2. **Potential Contributing Factors**:
    - Higher ticket classes may have provided better access to lifeboats or were prioritized during rescue operations.
    - Physical proximity to safety and socio-economic status could also have influenced survival rates.

The analysis underscores a significant relationship between ticket class and survival outcomes. Passengers in higher ticket classes enjoyed better survival prospects, reflecting potential socio-economic disparities in resource access and rescue priorities during emergencies. Further exploration of these factors could offer valuable insights into the dynamics of survival in crises.

**Port of Embarkation by Ticket Class Analysis**




**Insights:**

1. **Socio-Economic Differences by Port**:
    - **Southampton and Queenstown**: Primarily served 3rd class passengers, indicating these ports catered to more economically disadvantaged travelers.

    - **Cherbourg**: Predominantly served 1st class passengers, suggesting it was a hub for wealthier individuals.

2. **2nd Class Distribution**:
    - 2nd class passengers were relatively evenly distributed between Southampton and Cherbourg, with negligible representation from Queenstown.

The analysis highlights a strong relationship between socio-economic status (as reflected by ticket class) and the ports of embarkation. Southampton and Queenstown primarily catered to 3rd class passengers, while Cherbourg was a significant port for 1st class passengers. This distribution reflects the socio-economic diversity of the Titanic's passengers and provides insight into the demographic patterns associated with each port.

**Analysis of Survival Rates Based on Family Relationships Aboard**

The two bar charts provide insights into the survival rates among passengers based on their family relationships aboard the ship. One chart focuses on **parents and children**, while the other examines **siblings and spouses**.

**Observations**:

- **Passengers with 0 parents/children**:
    - The majority did not survive (peach bar), with fewer surviving (blue bar).
- **Passengers with 1 or 2 parents/children**:
    - Survival outcomes are more balanced, with survival counts (blue) slightly exceeding or equaling non-survival counts (peach).
- **Passengers with 3 to 6 parents/children**:
    - This group is small, with most passengers not surviving.

**Insights**:

- **Solo travelers** faced a higher risk of not surviving.
- **Small family groups (1 or 2 parents/children)** improved survival chances.
- **Large family groups (3 or more parents/children)** were rare and exhibited low survival rates.

---

**2. Survival Rate Among Siblings and Spouses Aboard**

**Axes**:

- **X-axis**: Number of siblings and spouses aboard.
- **Y-axis**: Count of passengers.

**Observations**:

- **Passengers with 0 siblings/spouses**:
    - Most did not survive (peach bar), but a notable number survived (blue bar).
- **Passengers with 1 sibling/spouse**:
    - Survival outcomes are more balanced, with survival numbers slightly higher.
- **Passengers with 2 to 8 siblings/spouses**:
    - This group is small, and most passengers did not survive.

**Insights**:

- **Solo travelers** faced a higher risk of not surviving.
- **Small groups (1 sibling/spouse)** showed improved survival odds.
- **Larger groups (2 or more siblings/spouses)** were uncommon and had lower survival rates.

**Data Cleaning and Preprocessing**

**Handling Missing Values**

- **Columns with Missing Values**:
    - **Age**: Contains missing entries.
    - **Embarked**: Some values are missing.
    - **Cabin**: Many missing values observed.

- **Imputation Strategy**:
    - **Age**: Missing values were replaced using **mean imputation**, where the average age is calculated and used to fill the gaps. This approach ensures a reasonable approximation of the missing values without introducing bias.
    - **Embarked**: Missing values were handled using **mode imputation**, where the most frequently occurring port of embarkation is used to replace the missing entries. This method maintains consistency in categorical data.
    - **Cabin**: Due to the significant number of missing values, imputing meaningful data was challenging. Therefore, mode imputation was employed as a simple and practical solution.

**2. Encoding Categorical Variables**

- **Purpose**:
  Machine learning models require numerical input. Hence, categorical variables were transformed into numerical representations.
- **Encoding Technique**:
    - **Label Encoding** was used to convert categories into integer values.
        - **Pclass**: Ticket classes were encoded numerically (e.g., 1st = 0, 2nd = 1, 3rd = 2).
        - **Sex**: Gender was encoded (e.g., Male = 0, Female = 1).
        - **Embarked**: Ports of embarkation were encoded (e.g., Southampton = 0, Cherbourg = 1, Queenstown = 2).
        - **Survived**: The target variable (Survival) was encoded (No = 0, Yes = 1).

**3. Dropping Irrelevant Columns**

- **Criteria**:
  Columns with **low correlation** to the target variable, Survived, were considered irrelevant for prediction and removed.
- **Dropped Columns**:
  - Examples include Name, Ticket, and Cabin. These features either lacked direct relevance to survival or posed challenges in extracting meaningful numerical values.
- **Resulting Dataset**:
  The cleaned dataset was reduced to 8 key features:
  - Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, and Survived.

**4. Scaling the Features**

- **Need for Scaling**:
  Features like Fare have values on a different scale compared to others like Age or SibSp. This variation can affect model performance, especially for distance-based algorithms.

**StandardScaler** was applied to standardize the features by removing the mean and scaling to unit variance. This ensures that each feature contributes equally to the model, preventing dominance by features with larger numerical ranges.

## 3. Feature Engineering

### Feature Retention Criteria

- **Correlation with Target Variable (Survived)**:
  Features with a strong relationship to survival outcomes were prioritized for retention.
- **Model-Based Importance Analysis**:
  Features contributing significantly to predictive accuracy, as indicated by model analysis, were selected.
- **Mutual Information**:
  A statistical technique used to measure the dependency between features and the target variable. Features with the highest mutual information scores were retained as they provide the most predictive value for survival.

## 2. Selected Features

Based on the above criteria, the following six features were identified as the most informative:

1. **Pclass (Passenger Class)**:
   - Represents socio-economic status (1st, 2nd, 3rd class).
   - Strongly correlated with survival rates, with 1st class passengers showing a higher likelihood of survival.
2. **Sex**:
   - Gender plays a significant role, as females had a much higher survival rate compared to males.
3. **Age**:
   - Reflects demographic factors impacting survival, with higher survival rates among children and young adults.
4. **SibSp (Siblings/Spouses Aboard)**:

- Indicates family connections aboard. Moderate family sizes (1 sibling/spouse) positively impact survival rates, while larger families decrease survival likelihood.

5. **Fare**:
    - Proxy for socio-economic status, as passengers paying higher fares were often in higher classes with better survival rates.

6. **Embarked (Port of Embarkation)**:
    - Highlights geographical and socio-economic differences. For example, Cherbourg had a higher concentration of 1st-class passengers, reflecting better survival chances.

### 3. Mutual Information for Feature Selection

- Mutual information measures the dependency between a feature and the target variable. Features with higher scores contribute more unique information to predicting survival.
- The selected features had the highest mutual information scores, ensuring an optimal balance between dataset size and predictive capability.

### 4. Model Training and Evaluation

#### Models Used for Survival Prediction

To predict survival outcomes, the following machine learning models were implemented and evaluated:

#### 1. Logistic Regression

- A statistical model that estimates the probability of survival using a linear combination of input features.
- Advantage: Interpretable and effective for binary classification tasks.

**2. Random Forest Classifier**

- An ensemble learning method combining multiple decision trees to improve accuracy and reduce overfitting.
- Advantage: Handles non-linear relationships and provides feature importance insights.

**3. Support Vector Machine (SVM)**

- A model that finds the hyperplane best separating survivors and non-survivors in feature space.
- Advantage: Effective in high-dimensional spaces.

**4. Decision Tree Classifier**

- A simple, interpretable tree-based algorithm for classification.
- Advantage: Easy to visualize and understand.

**5. Gradient Boosting Classifier**

- An advanced ensemble technique that builds models sequentially to minimize errors from previous iterations.
- Advantage: Highly accurate and effective in capturing complex relationships.

**6. K-Nearest Neighbors (KNN)**

- A non-parametric algorithm that predicts survival based on the majority class of the closest neighbors in feature space.
- Advantage: Simple and intuitive.

**Metrics for Evaluation**

The models were evaluated based on the following metrics:

1. **Accuracy**:
    - Proportion of correct predictions out of total predictions.
    - Provides a general measure of model performance.
2. **Precision, Recall, and F1-Score**:
    - **Precision**: Measures the proportion of correctly predicted positive cases out of all predicted positives.
    - **Recall**: Measures the proportion of correctly predicted positives out of all actual positives.
    - **F1-Score**: Harmonic mean of precision and recall, balancing the trade-off between the two.
3. **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)**:
    - Evaluates the model's ability to distinguish between classes.
    - A higher AUC indicates better classification performance.

**Model Results**

The performance of the models was analyzed and compared based on the above metrics.

- **Gradient Boosting Classifier**:
    - Achieved the highest accuracy among all models **before optimization**.
    - Excelled in capturing complex patterns and provided superior classification quality.

## 5. Model Optimization

To improve the performance of the models, hyperparameter tuning was applied using **GridSearchCV**.

### Hyperparameter Tuning Process

- **GridSearchCV**:
    - A systematic method for searching over a predefined hyperparameter grid.
    - Evaluates all possible combinations of specified hyperparameters to identify the optimal configuration.
    - Cross-validation ensures that the model generalizes well to unseen data by splitting the dataset into training and validation sets multiple times.

### Optimization Results

- The **Decision Tree Classifier** emerged as the best-performing model after optimization.
- Key improvements:
    - **Optimized Parameters**: Fine-tuned hyperparameters (e.g., max depth, minimum samples per leaf) helped balance bias and variance.
    - **Final Accuracy**: The optimized model achieved an accuracy of **81%**, representing a notable improvement over the baseline performance.

## 6. Testing and Submission

### Prediction

After optimizing the models and selecting the best-performing one, the final model was applied to the **test dataset** for evaluation.

- **Test Dataset**:
    - The final model, the **Decision Tree Classifier**, was used to predict survival outcomes on the test dataset provided in Data/test.csv.
    - **Achieved Accuracy**: The model achieved an accuracy of **78%**, indicating that it performs well on unseen data.

**Conclusion**

This project has provided a thorough exploration of machine learning techniques for predicting survival outcomes from the Titanic dataset. After applying several machine learning models, **Decision Tree Classifier** emerged as the best-performing model, demonstrating an accuracy of 81% after optimization and 78% on the test dataset.

**Key Insights from the Data Analysis:**

1. **Economic Stratification**:
    - Southampton and Queenstown primarily served 3rd-class passengers, indicating that these ports were departure points for more economically disadvantaged individuals. This was in contrast to **Cherbourg**, which catered predominantly to wealthier 1st-class passengers, suggesting its prominence among affluent communities.
    - The socio-economic differences in the passengers' class were reflected in their embarkation points, with Cherbourg being a hub for the wealthiest and Southampton and Queenstown serving lower-class travelers.
2. **Class Inequality**:
    - The analysis clearly shows the role of **ticket class** in survival outcomes. The **1st-class passengers** had a significantly higher chance of survival, with most surviving, as compared to **3rd-class passengers** who had the lowest survival rate.
    - The **3rd-class passengers** were more likely to have boarded at Southampton and Queenstown, further emphasizing the class divide between the embarkation points and their associated socio-economic status.

3. **Gender Disparities**:
    - The gender-based survival analysis showed that **females** had a much higher survival rate than **males**, supporting historical accounts of the "women and children first" policy. The higher survival rate among females could also be attributed to evacuation policies that prioritized women during the crisis.
4. **Family Influence on Survival**:
    - Passengers traveling **alone** had a significantly lower chance of survival. Conversely, those traveling with **1 or 2 family members** had better odds of survival. Larger family groups, however, had a very low survival rate, likely due to difficulties in reuniting during the chaotic evacuation process.
5. **Port of Embarkation and Ticket Class**:
    - The relationship between **port of embarkation** and **ticket class** was evident in the data. **Southampton** and **Queenstown** saw a higher proportion of **3rd-class passengers**, while **Cherbourg** was the primary boarding point for **1st-class passengers**, highlighting the socio-economic stratification that existed among passengers.
6. **Historical Implications on Survival**:
    - The socio-economic and class-related disparities are likely to have significantly influenced survival outcomes during the **Titanic disaster**. The wealthier **1st-class passengers** had better access to lifeboats and other safety measures, contributing to their higher survival rates.
    - In contrast, **3rd-class passengers**, often from lower socio-economic backgrounds, faced challenges in accessing lifeboats, which may have contributed to their lower survival rates.

**Final Model Performance:**

The **Decision Tree Classifier** achieved a solid performance, with an accuracy of **78% on the test dataset** after fine-tuning, validating its effectiveness in predicting survival outcomes based on the dataset's features.

The project has provided a deep understanding of the relationship between socio-economic factors, ticket class, and survival outcomes during the Titanic disaster. The Decision Tree Classifier, optimized through hyperparameter tuning, emerged as the best model for this prediction task. The findings highlight the enduring impact of **economic stratification** and **class inequality** on the Titanic disaster, influencing survival chances in ways that reflect broader historical and social inequalities.