## **1. Introduction**

The Titanic disaster remains one of the most infamous maritime tragedies in history. The dataset associated with it has become a well-known benchmark in machine learning and data science. The primary objective of this study was to predict whether a passenger survived the sinking of the Titanic using a **classification model**.

For this analysis, the original dataset file train.csv was modified to test.csv, although the structure and attributes remained the same. The machine learning workflow included:

**Data Preprocessing** – Handling missing values, encoding categorical variables, and feature scaling.

**Model Training** – Applying a **Logistic Regression** model for binary classification.

**Model Evaluation** – Assessing model performance using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

## **2. Data Analysis and Preprocessing**

### **Exploratory Data Analysis (EDA)**

A thorough EDA was performed to gain insights into the dataset:

**Dataset Overview**

* The dataset consists of **passenger attributes** such as **Name, Age, Sex, Ticket Class, Fare, Number of Siblings/Spouses aboard, Number of Parents/Children aboard, and Embarkation Port**.
* The **target variable (Survived)** is binary:
  + 0: Did not survive
  + 1: Survived

**Key Insights from EDA:**

* **Gender Impact on Survival:**
  + Female passengers had a significantly higher survival rate (about **75%**), while only **18-20% of males** survived.
* **Class-Based Survival:**
  + First-class passengers had the highest survival rate (**~63%**), whereas third-class passengers had the lowest (**~25%**).
* **Age Distribution & Impact:**
  + Children (age **≤ 10**) had a higher survival rate compared to adults.
* **Fare Influence:**
  + Passengers who paid a higher fare were more likely to survive, suggesting economic status played a crucial role.

### **Data Preprocessing Steps**

To ensure data consistency and improve model accuracy, the following preprocessing steps were applied:

**Handling Missing Values**

* **Age**: Missing values were replaced using the **median age** of the dataset.
* **Embarked**: Missing values were filled using the **most frequent embarkation port (mode)**.
* **Cabin**: Dropped due to excessive missing values (**~77% missing data**).

**Encoding Categorical Variables**

* **Sex** was encoded into numerical values (Male = 0, Female = 1) using **Label Encoding**.
* **Embarked** was transformed using **One-Hot Encoding** to convert categorical values into machine-readable numerical values.

**Feature Scaling**

* **Age** and **Fare** were standardized using **StandardScaler** to bring them to the same scale and improve model performance.

**Train-Test Split**

* The dataset was split into **80% training** and **20% testing** to train and evaluate the model effectively.

## **3. Model Selection and Evaluation**

### **Model Choice: Logistic Regression**

Logistic Regression was chosen as the baseline model for this binary classification task due to its interpretability, simplicity, and strong performance for problems with **linear decision boundaries**.

### **Model Performance Metrics**

After training the model on the dataset, the evaluation metrics were as follows:

| **Metric** | **Score** |
| --- | --- |
| **Accuracy** | **0.8045** |
| **Precision** | **0.7671** |
| **Recall** | **0.7568** |
| **F1 Score** | **0.7619** |

### **Classification Report**

| Class | Precision | Recall | F1 Score | Support |
| --- | --- | --- | --- | --- |
| 0 (Not Survived) | 0.83 | 0.84 | 0.83 | 105 |
| 1 (Survived) | 0.77 | 0.76 | 0.76 | 74 |
| Overall Accuracy | 0.80 |  |  | 179 |

### 

### **Key Observations from Model Performance**

**Confusion Matrix Analysis:**The confusion matrix revealed:

* **True Positives (TP)**: Passengers correctly classified as survivors.
* **True Negatives (TN)**: Passengers correctly classified as non-survivors.
* **False Positives (FP)**: Passengers incorrectly classified as survivors.
* **False Negatives (FN)**: Passengers incorrectly classified as non-survivors.

**ROC-AUC Score:**

* The **ROC Curve** showed an **AUC score of ~0.84**, indicating good discrimination between survivors and non-survivors.
* AUC values closer to **1.0** indicate better classification performance.

**Bias-Variance Tradeoff:**

* Logistic Regression performed **reasonably well**, but further improvements could be made with **ensemble methods like Random Forest or Gradient Boosting**.

## **4. Conclusion**

### **Key Takeaways**

**Logistic Regression achieved an accuracy of 80.45%, meaning the model correctly classified passengers 8 out of 10 times.**

**Precision of 76.71%** shows that when the model predicts a passenger survived, it is correct ~77% of the time.

**Recall of 75.68%** indicates that the model correctly identifies 76% of all actual survivors.

The model performed well, but **improvements** could be made with:  
**Feature Engineering** – Creating new features like "Family Size" or extracting titles from passenger names.  
**Advanced Models** – Trying Random Forest, XGBoost, or Neural Networks for better predictive power.  
**Hyperparameter Tuning** – Optimizing parameters like regularization strength to further improve performance.

This analysis demonstrated how machine learning can be used to predict survival probabilities based on real-world datasets. The Titanic dataset provided valuable insights into how different factors (gender, class, fare) influenced survival chances.

While Logistic Regression served as a solid baseline model, there is potential to improve classification performance with more **sophisticated algorithms and refined features**.