# Predicting Survival on the Titanic Using Machine Learning

## 1. Introduction

The sinking of the RMS Titanic in 1912 remains one of the most tragic maritime disasters in history, claiming the lives of more than 1,500 passengers. From a machine learning perspective, this dataset provides a classical binary classification problem: predicting whether a passenger survived (1) or not (0) based on socio-demographic and travel-related attributes.  
  
This project applies logistic regression as a baseline model to predict survival, followed by the creation of a submission file in Kaggle’s required format.

## 2. Data Description

Three datasets were provided by Kaggle:  
- train.csv – Passenger features with survival outcome (Survived).  
- test.csv – Passenger features without survival outcome.  
- gender\_submission.csv – A sample submission assuming all females survived.  
  
Variables:  
- Target variable: Survived – Survival status (0 = did not survive, 1 = survived).  
- Features:  
 - Pclass – Ticket class (1 = upper, 2 = middle, 3 = lower).  
 - Sex – Gender of passenger.  
 - Age – Age in years.  
 - SibSp – Number of siblings/spouses aboard.  
 - Parch – Number of parents/children aboard.  
 - Fare – Ticket fare.  
 - Embarked – Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

## 3. Data Preprocessing

Several preprocessing steps were applied to clean and prepare the data:  
- Handling missing values:  
 - Age – Missing values imputed with the median age.  
 - Fare – Missing values imputed with the median fare.  
 - Embarked – Missing values replaced with the most frequent port (“S”).  
- Categorical encoding: Sex and Embarked were converted to factors for logistic regression.  
- Feature scaling: Not required, since logistic regression in R handles raw numeric inputs.

## 4. Exploratory Data Analysis (EDA)

Exploratory analysis highlighted strong survival patterns:  
- Gender: Women had a much higher survival rate than men.  
- Class: First-class passengers were more likely to survive than those in third class.  
- Age: Younger passengers, particularly children, had higher survival rates.  
- Fare: Passengers who paid higher fares showed better survival odds.  
  
These findings align with historical accounts and suggest that socio-economic status and gender strongly influenced survival.

## 5. Model Development

A logistic regression model was fitted using the training dataset.  
  
Model formula:  
Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
  
R Implementation:  
model\_logit <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,   
 data = train, family = binomial)  
summary(model\_logit)

## 6. Results

The logistic regression model identified several significant predictors of survival:  
- Sex (male) – Strong negative effect (males were less likely to survive).  
- Pclass – Higher passenger class increased survival odds.  
- Age – Older passengers were less likely to survive.  
- Fare – Higher ticket fares slightly increased survival chances.  
- Embarked (S) – Boarding at Southampton reduced survival odds compared to Cherbourg.  
  
These results validate the well-known narrative: “Women and children first,” with preference for wealthier passengers.

## 7. Prediction and Submission

- Predictions were generated for the test dataset using the trained logistic regression model.  
- Probabilities > 0.5 were classified as “Survived = 1”; otherwise “0.”  
- Final output was formatted into a Kaggle-compatible file (my\_submission.csv) containing:  
 - PassengerId  
 - Survived

## 8. Model Evaluation

- On Kaggle, logistic regression achieves an accuracy score of approximately 77–79%, slightly better than the baseline gender\_submission.csv (≈76%).  
- While logistic regression provides interpretability, it is limited in capturing complex feature interactions.  
- More advanced models such as Random Forests or Gradient Boosting (XGBoost, LightGBM) typically outperform logistic regression (>80% accuracy).

## 9. Conclusion

This project successfully applied logistic regression to predict Titanic passenger survival. The analysis revealed that survival was primarily determined by gender, class, age, and fare. These findings are consistent with historical accounts, demonstrating the social inequalities of the tragedy.  
  
Future Work:  
- Implement ensemble models (Random Forest, Gradient Boosting) for improved accuracy.  
- Perform feature engineering (e.g., family size from SibSp + Parch, extracting titles from names, cabin deck from Cabin).  
- Apply cross-validation for more robust evaluation.