# TITANIC BINARY CLASSIFICATION PROJECT

**Project Title:**
Titanic Survival Prediction Using Binary Classification

**Dataset**:
Kaggle Titanic Dataset – train.csv

**Tools Used:**
Python, Google Colab, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, XGBoost

**NAME:** Aishwarya Gunaseelan

**DOMAIN:** Data Science

# 1. INTRODUCTION

Data science plays a major role in extracting meaningful knowledge from historical data and using it to make intelligent predictions. Binary classification is one of the most widely used machine learning techniques where the target variable has only two possible outcomes such as Yes/No, True/False, or 0/1.

The Titanic disaster dataset is a well-known real-world dataset that contains information about passengers who traveled on the Titanic ship. The dataset includes attributes such as age, gender, passenger class, ticket fare, and embarkation point. The goal of this project is to build a machine learning model that can predict whether a passenger survived or not based on these features.

This project follows a complete end-to-end machine learning workflow starting from raw data loading to final model optimization. The work includes data cleaning, handling missing values, exploratory data analysis, feature engineering, model building using multiple algorithms, evaluation using standard performance metrics, and hyperparameter tuning. The project demonstrates practical implementation of data science concepts using Python in Google Colab environment.

## 2. AIM

The main aim of this project is:

- To develop an efficient binary classification model to predict passenger survival.
- To analyze the influence of demographic and travel features on survival.
- To compare different machine learning algorithms.
- To select the best performing model using evaluation metrics and hyperparameter tuning.

## 3. BUSINESS PROBLEM / USE CASE

The Titanic tragedy resulted in the loss of more than 1500 lives. However, survival was not random; it depended on several social and economic factors such as gender, class, and age.

This project addresses the following real-world problems:

- Identifying factors that increase survival chances
- Understanding risk patterns during disasters
- Helping authorities design better evacuation strategies
- Demonstrating predictive analytics in historical events

Such analysis can be extended to modern transportation safety, disaster management, and risk assessment domains.

## 4. PROJECT WORKFLOW

The project follows these steps:

1. Dataset Selection
2. Data Loading
3. Data Understanding
4. Data Cleaning
5. Exploratory Data Analysis
6. Data Preprocessing
7. Model Training
8. Model Evaluation
9. Hyperparameter Optimization
10. Result Interpretation
11. Documentation

## 5. DATA UNDERSTANDING

### Dataset Details

- Source: Kaggle Titanic Dataset
- File used: train.csv
- Total records: 891
- Total features: 12
- Target: Survived (0 = No, 1 = Yes)

## Feature Description

- PassengerId – Unique ID
- Pclass – Ticket class (1,2,3)
- Name – Passenger name
- Sex – Male/Female
- Age – Age in years
- SibSp – Siblings/Spouse aboard
- Parch – Parents/Children aboard
- Ticket – Ticket number
- Fare – Ticket price
- Cabin – Cabin number
- Embarked – Port of boarding
- Survived – Target variable

## Initial Observations

- Missing values in Age, Cabin, Embarked
- Mix of numerical & categorical data
- Survival distribution is imbalanced
- Some columns not useful for prediction

## 6. DATA CLEANING

The following cleaning steps were performed:

1. Handling Missing Values

- Age → replaced with median
- Embarked → replaced with mode
- Cabin → dropped due to high missing %

2. Removing Irrelevant Features

- PassengerId
- Name
- Ticket
- Cabin

3. Data Consistency

- Checked duplicates
- Verified data types
- Removed inconsistencies

These steps improved data quality for modeling.

## 7. EXPLORATORY DATA ANALYSIS (EDA)

EDA was performed to understand relationships.

### Univariate Analysis

- Survival count showed more non-survivors
- Age distribution mostly 20–40
- Majority traveled in class 3

### Bivariate Analysis

- Females survived more than males
- Class 1 > Class 2 > Class 3 survival
- Higher fare → higher survival

### Correlation Analysis

- Strong link between Pclass and Survived
- Moderate relation with Fare and Age

Visualizations used:

- Count plots
- Histograms
- Bar charts
- Heatmap

## 8. DATA PREPROCESSING

Steps:

1. Encoding

- Sex → Label Encoding
- Embarked → Label Encoding

2. Feature Scaling

- StandardScaler applied

3. Feature & Target

- X → independent features
- y → Survived

4. Train Test Split

- 80% training
- 20% testing

## 9. MODEL TRAINING

Five algorithms were implemented:

1. Logistic Regression
2. Support Vector Machine
3. Decision Tree
4. Random Forest
5. XGBoost

Each model was trained using the same data for fair comparison.

## 10. MODEL EVALUATION

### Metrics Used

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC
- Confusion Matrix

### Tuned Random Forest Result

- Accuracy: 81.56%
- Precision: 83.60%
- Recall: 68.91%
- F1 Score: 75.55%
- ROC AUC: 0.79

Confusion Matrix:

- TN = 95
- TP = 51
- FP = 10
- FN = 23

## 11. MODEL COMPARISON & SELECTION

Observations:

- Logistic Regression – simple baseline
- SVM – good but sensitive
- Decision Tree – overfitting
- Random Forest – best balance
- XGBoost – high but complex

**Selected Model**

👉 Tuned Random Forest

Reasons:

- Higher accuracy
- Less overfitting
- Stable predictions
- Interpretable

# 12. HYPERPARAMETER TUNING

GridSearchCV applied on Random Forest

Parameters:

- n_estimators = [50,100]
- max_depth = [5,10]

Best Parameters:

- max_depth = 5
- n_estimators = 50

Tuned model improved generalization.

# 13. CONCLUSION

The project successfully implemented an end-to-end binary classification system using the Titanic dataset. After performing data cleaning, exploratory analysis, and model comparison, the **Tuned Random Forest** model was selected as the best performer with **81.56% accuracy**.

The analysis proved that:

- Gender and passenger class were the strongest factors
- Females and class 1 had highest survival
- Machine learning can effectively model historical events

**Future Scope**

- Apply SMOTE for imbalance
- Cross validation
- Feature engineering (family size, title)
- Try LightGBM, CatBoost
- Deploy as web app