🚢 Titanic Survival Prediction – Report

1. Introduction

Titanic dataset is a classic machine learning problem where the goal is to predict whether a passenger survived or not based on demographic and travel information. This project followed a complete data science workflow: data cleaning, exploratory data analysis (EDA), feature engineering, model building, evaluation, and prediction on unseen test data.

---

2. Data Overview

Two datasets:

- **train.csv** (891 passengers, includes the target variable `Survived`)
- **test.csv** (418 passengers, no `Survived`, predictions required)

Key Features

- **PassengerId**: Unique passenger ID
- **Pclass**: Ticket class (1st, 2nd, 3rd)
- **Name**: Passenger name (titles useful)
- **Sex**: Gender
- **Age**: Age in years
- **SibSp**: Number of siblings/spouses aboard
- **Parch**: Number of parents/children aboard
- **Ticket**: Ticket number
- **Fare**: Ticket fare
- **Cabin**: Cabin number (many missing)
- **Embarked**: Port of embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

3. Data Cleaning & Feature Engineering

Steps taken:

- **Missing Values**:
    - `Age`: filled with median per passenger title (e.g., Mr, Miss, Mrs).
    - `Embarked`: filled with the most common value (S).
    - `Cabin`: dropped but replaced with binary feature `HasCabin`.
- **Feature Engineering**:
    - Extracted **Title** from names, grouped rare titles into `"Rare"`.
    - Created **FamilySize** = SibSp + Parch + 1.
    - Created **IsAlone** = 1 if FamilySize = 1, else 0.
- **Encoding**: Converted categorical variables (`Sex, Embarked, Title`) into numeric via mapping and one-hot encoding.

---

4. Exploratory Data Analysis (EDA)

Key findings:

- **Sex**: Females had much higher survival rates than males.
- **Pclass**: First-class passengers survived more frequently than third-class.
- **Age**: Children had a higher survival rate.
- **FamilySize**: Very large families and solo travelers survived less than medium-sized families.

A correlation heatmap confirmed strong relations between survival and features like `Sex`, `Pclass`, and `Fare`.

5. Modeling

We tested three models: **Logistic Regression ,Decision Tree ,And Random Forest**

Validation accuracy (20% split of training data):

- Logistic Regression: **82.7%**
- Decision Tree: 77.6%
- Random Forest: 77.6%

Best model: Logistic Regression

---

6. Predictions on Test Data

The `test.csv` dataset was cleaned with the same preprocessing steps as the training data, aligned to have identical feature columns, and passed through the Logistic Regression model.

---

8. Conclusion

The Logistic Regression model performed the best, achieving **82.7% validation accuracy**

- The most important predictors of survival were: **Sex, Pclass, Age, Fare, and Title**.
- Future improvements could include feature scaling, hyperparameter tuning, and advanced models like XGBoost or LightGBM.

This project demonstrated the full data science pipeline — from raw data to final predictive model and Kaggle submission.