Titanic Survival Prediction - Logistic Regression

1. Project Title

Titanic Survival Prediction using Logistic Regression

2. Objective

The goal of this project is to predict whether a passenger survived the Titanic disaster using a logistic regression model based on passenger features such as age, gender, and passenger class.

3. Dataset Description

The dataset contains information about passengers aboard the Titanic, including whether they survived (Survived column).

Columns:

- PassengerId: Unique passenger ID
- Survived: Target (0 = Did not survive, 1 = Survived)
- Pclass: Passenger Class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name: Name of the passenger
- Sex: Gender
- Age: Age in years
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Fare paid for the ticket
- Cabin: Cabin number (many missing values)
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

4. Exploratory Data Analysis (EDA)

Key Insights:

- Survival Distribution: Majority of passengers did not survive.
- Gender Impact: Females had a higher survival rate.
- Passenger Class: Passengers in 1st class had higher survival rates.
- Age Distribution: Majority between 20-40 years.

- Fare Distribution: Most passengers paid lower fares, with a few paying very high fares.
- Correlation: Gender (Sex) and passenger class (Pclass) were strongly correlated with survival.

(Plots such as survival distribution, gender, class-wise survival, heatmap, age, fare distributions can be added here.)

5. Data Preprocessing

- Missing Values Handled:
  - Age: Filled with median.
  - Embarked: Filled with mode ('S').
- Dropped Columns: PassengerId, Name, Ticket, Cabin.
- Categorical Encoding: Converted Sex and Embarked to numeric using one-hot encoding.

6. Model Building

- Algorithm: Logistic Regression
- Data Split: 80% training, 20% testing
- Model Accuracy: 81.01%

7. Model Performance Evaluation

Metric	Class 0 (Not Survived)	Class 1 (Survived)
Precision	0.83	0.79
Recall	0.86	0.74
F1-Score	0.84	0.76
Accuracy		81.01%

- Confusion Matrix: Shows correct and incorrect predictions.
- ROC Curve: Indicates the model distinguishes well between the two classes.

(Plots for confusion matrix and ROC curve can be inserted.)

8. Conclusion

- Logistic Regression achieved good performance with 81.01% accuracy.
- Gender and Passenger Class were the most influential features.

- The model performed better in predicting non-survivors but maintained decent performance for survivors.
- Future improvements include feature engineering (e.g., FamilySize), hyperparameter tuning, and trying advanced models like Random Forest and XGBoost.

9. References

Dataset: Kaggle Titanic Competition (https://www.kaggle.com/competitions/titanic)