**Titanic Survival Prediction Project Documentation**

**By: Suravi Poudel**

**1. Dataset Description**

The Titanic dataset contains information about passengers aboard the RMS Titanic. The dataset includes various features about each passenger and whether they survived or not.

* **Features include:**
  + PassengerId: Unique ID of passenger
  + Pclass: Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)
  + Name: Passenger name
  + Sex: Gender (male or female)
  + Age: Age in years (some missing values)
  + SibSp: Number of siblings/spouses aboard
  + Parch: Number of parents/children aboard
  + Ticket: Ticket number
  + Fare: Ticket fare paid
  + Cabin: Cabin number (many missing)
  + Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)
  + Survived: Target variable (0 = No, 1 = Yes)

The goal is to predict passenger survival (Survived) based on other features.

**2. Data Cleaning and Transformation Steps**

**Loading and inspecting the data**

* Loaded the dataset using pandas.read\_csv.
* Used df.head(), df.info(), and df.describe() to get an overview.
* Checked for missing data visually using seaborn heatmap and df.isnull().sum().

**Handling Missing Data**

* Noted significant missing values in the Age column.
* Filled missing Age values with the **median age** to avoid bias.
* Dropped other columns with many missing values such as Cabin.

**Outlier Detection and Removal**

* Used boxplots to analyze the distribution of Age and Fare.
* Calculated Interquartile Range (IQR) for Age to identify outliers.
* Removed outliers outside the range:  
  [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR] to improve data quality.

**Feature Engineering**

* Created a new feature FamilySize by summing SibSp and Parch.
* Converted categorical variables:
  + Sex: mapped male → 0, female → 1.
  + Embarked: one-hot encoded into Embarked\_C, Embarked\_Q, Embarked\_S.

**Feature Scaling**

* Standardized Age and Fare using StandardScaler to normalize numerical ranges for model training.

**3. Exploratory Data Analysis (EDA) Insights**

**Missing Data Visualization**

* Heatmap showed missing values primarily in Age.

**Distribution Plots**

* Histograms and boxplots of Age and Fare showed mostly normal distribution with some outliers removed.

**Survival Counts**

* Countplots revealed:
  + Overall survival rate.
  + Females survived at a much higher rate than males.
  + First-class passengers had higher survival rates than lower classes.
  + Differences in survival by port of embarkation (Embarked).

**Correlation Analysis**

* Heatmap of numerical variables showed:
  + Strong negative correlation between Pclass and survival.
  + Positive correlation between Fare and survival.
  + Sex (encoded) highly correlated with survival.

**4. Model Development, Evaluation, and Deployment**

**Feature Importance**

* Used Random Forest classifier to compute feature importances.
* Selected key features for modeling:  
  Pclass, Age, Fare, FamilySize, Sex\_enc, Embarked\_enc.

**Data Preparation**

* Filled any remaining missing Age values.
* Split data into training (80%) and testing (20%) sets with stratification.

**Model Training and Comparison**

* Trained multiple classification algorithms:
  + Logistic Regression
  + Decision Tree
  + Random Forest
  + Support Vector Machine (SVM)
  + K-Nearest Neighbors (KNN)
  + Naive Bayes
* Evaluated on accuracy, precision, recall, and F1-score.
* Logistic Regression performed well and was chosen for deployment.

**Cross-Validation and Hyperparameter Tuning**

* Performed 5-fold cross-validation to estimate model stability.
* Used Grid Search to tune logistic regression parameters (C, solver).
* Achieved best accuracy with tuned parameters.

**Model Evaluation on Test Set**

* Achieved accuracy, precision, recall, and F1-scores reported with classification report and confusion matrix.

**Deployment**

* Saved trained logistic regression model using pickle.
* Built a Streamlit app to accept user inputs, transform features similarly, and predict survival interactively.

**Summary**

The project demonstrates the full lifecycle of a machine learning pipeline on the Titanic dataset: data loading, cleaning, EDA, feature engineering, model training, evaluation, tuning, and deployment. The final model allows users to input passenger details and get a survival prediction via a web app.