Machine Learning

Final Project Submission

## Name: Maheen Rasool

## Roll Number: 22i – 0626

## Section: AI-B

**Report: Predicting Flight Departure Delays**

**Project Overview:** Flight departure delays are a critical challenge in the aviation industry. These delays impact passenger satisfaction, airline operations, and the overall efficiency of the aviation system. This project analyzes historical flight and weather data to identify patterns in flight delays and builds predictive models to predict future delays. The models developed are used for submission to a Kaggle competition for evaluation.

**Phase 1: Data Preprocessing and Feature Engineering**

**1. Data Integration:**

The weather dataset was integrated with the flight data to provide a comprehensive view of factors that could influence delays. The integration helped incorporate weather variables such as temperature and wind speed, which are critical in assessing potential flight delays.

**2. Data Cleaning and Transformation:**

* **Handling Missing Values:**  
  Missing values were handled by computing the central value (median or mean) of the "Delay" column and filling them accordingly. This central value was added to the "Scheduled Time" to derive the "Actual Runway Time" (ART).
* **Formatting Time Fields:**  
  All time fields (Scheduled, Actual, Estimated Times) were converted into a standard datetime format for consistency across the dataset. This ensures proper handling of time-based features like the hour of the day and day of the week.

**3. Feature Engineering:**

* **Departure Delay Calculation:**  
  Departure delay was computed by subtracting the departure scheduled time from the departure Actual Runway.
* **Merge Weather Data:**  
  Relevant weather features such as temperature and wind speed were merged with flight data to enhance the model's ability to detect weather-related delays.
* **Temporal Features Extraction:**  
  Additional features were derived from the datetime column, such as:
  + Day of the Week (e.g., Monday, Tuesday)
  + Hour of the Day (e.g., 8 AM, 9 AM)
  + Month of the Year (e.g., January, February)

These derived features help the model account for temporal patterns in flight delays.

**Phase 2: Exploratory Data Analysis (EDA)**

**1. Visualizations:**

* **Delay Distributions:**  
  A histogram of delay durations was plotted to understand the spread of delays. The distribution of delays helped us identify common delay ranges and the extent of delays for flights.
* **Temporal Analysis:**  
  Line plots and bar charts were generated to show delays across hours, days, and months. This helped us understand when delays are most frequent (e.g., delays might be more frequent in the afternoon or on certain days of the week).
* **Category-Wise Analysis:**  
  Grouping delays by airline, departure airport, and flight status allowed us to identify whether specific carriers or airports experience higher delays.

**2. Correlation Analysis:**

The correlation between weather features and flight delay data was analyzed using a correlation matrix and visualizations. Three visualizations were used to assess the relationship between:

* Temperature and delay duration
* Wind speed and delay duration
* Flight status and delay

**3. Comparison:**

The delay data from the training and testing datasets was compared to ensure consistency. Any significant differences were handled through normalization or additional feature engineering.

**Phase 3: Analytical and Predictive Tasks**

**1. Classification Tasks:**

**Binary Classification:**

* Flights were classified as either "On-Time" (delay = 0) or "Delayed" (delay > 0).
* A logistic regression model was trained to predict binary outcomes.
* Performance metrics used for evaluation:
  + **Accuracy:** Measures the proportion of correct predictions.
  + **Precision-Recall:** Provides a balance between precision and recall for imbalanced datasets.
  + **F1-Score:** Harmonic mean of precision and recall.
  + **Confusion Matrix:** Helps visualize the classification performance.

**Multi-Class Classification:**

* Flights were categorized into four classes based on delay duration:
  + No Delay (0 min)
  + Short Delay (<45 min)
  + Moderate Delay (45–175 min)
  + Long Delay (>175 min)
* A multinomial logistic regression model was trained for multi-class classification.
* Evaluation metrics:
  + **Accuracy**
  + **Precision-Recall**
  + **F1-Score**
  + **Class-wise Precision-Recall**
  + **Confusion Matrix**

**2. Regression Analysis:**

**Delay Duration Prediction:**

* A regression model was trained to predict the exact delay duration for each flight.
* Cross-validation techniques were applied to assess the model’s generalization performance.
* Performance metrics:
  + **Mean Absolute Error (MAE):** Measures the average absolute error.
  + **Root Mean Squared Error (RMSE):** Measures the square root of the average squared error, emphasizing larger errors.

**Phase 4: Model Optimization and Evaluation**

**1. Hyperparameter Tuning:**

* Techniques such as **Grid Search** were used to optimize hyperparameters of the regression and classification models. These approaches ensured the best model performance by exploring different parameter configurations.

**2. Validation:**

* **K-Fold Cross-Validation:** The dataset was split into multiple folds to validate the model's performance and reduce overfitting. The models were trained and evaluated on each fold to ensure that the model generalizes well to unseen data.