Final Project (BS-AI: Fall-2024)

Advanced Flight Departure Delay Analysis Project

AI-3002: Machine Learning **Due Date: 9th Dec,2024 (11:55 PM)**

Instructions

1. You can not convert .docx files to tables by using any tool, software, or extension.
2. You can use the Sklearn library for training models and code in Python.
3. The goal of the project is to calculate **departure delays** and analyze patterns affecting delays using the provided data. You will preprocess the data, calculate delays, and create predictive models to classify and predict delay durations.
4. Specify any assumptions made during data preprocessing or modeling in the report.
5. Use meaningful visualizations to support your analysis.
6. The **test dataset** provided for this project contains flight information without delay values, as your task is to predict these departure delays using models trained on the **training dataset**. Begin by ensuring that the **test data** is preprocessed consistently with the train data, including formatting time fields, merging relevant weather data, and extracting additional features such as time-based attributes (e.g., hour, day of the week). Once your model is trained and validated using the train data, use it to predict the delays for each flight in the test dataset. The predicted delays should be saved in the format specified for the Kaggle competition: a CSV file with columns for Flight Number (unique identifier) and predicted delay.To submit your predictions, first join the Kaggle competition using the link provided by your instructor. Upload the name_submission.csv file directly to the competition page. Kaggle will automatically evaluate your predictions based on its scoring metric and display your rank on the leaderboard.
7. Ensure the submission file has the correct structure and no missing or erroneous values before uploading. You can iterate on your models and submit improved predictions to climb the leaderboard

Note: There is no extension in the deadline. Please submit it on time. There should be no comments provided on each line of code.

Deliverables:

Cleaned Dataset: Processed data including the specified features.

Report: A detailed report about insights from data analysis, regression, and classification models with performance metrics and practical strategies based on analytical findings.

ZIP File Submission

Put all your files inside one folder and name it according to the following convention:

RollNo_Sec_MLProject.ZIP

You must upload only one ZIP file.

Problem Statement

Flight departure delays are a critical challenge in the aviation industry. Such delays affect passenger satisfaction, airline operations, and overall efficiency. You are provided with raw Excel files (test, train, and weather data) and are tasked with calculating departure delays. Using these datasets, you will analyze delay patterns and build predictive models to identify key factors contributing to delays.

Objective

The main goal is to predict **departure delays** for flights in the test dataset. You will:

1. Analyze the train data, test data, and weather data.
2. Build predictive models based on the train data.
3. Generate predictions for the test data.
4. Submit your predictions to a Kaggle competition for evaluation.

Phase 1: Data Preprocessing and Feature Engineering

1. Data Integration

- Integrate the weather dataset with the training dataset for any further processing.

2. Data Cleaning and Transformation

1. **Handle Missing Values.**
2. **Format Time Fields:**
    - Convert time fields (e.g., Scheduled, Actual, Estimated Time) into a standard datetime format for consistency.

3. Feature Engineering

1. **Calculate Departure Delay:** Compute the departure delay.
2. **Merge Weather Data:** Extract relevant weather features (e.g., temperature, wind speed) and join them with flight data.
3. **Extract Temporal Features:**
    - Derive additional features such as:
        - Day of the week (e.g., Monday, Tuesday).
        - Hour of the day.
        - Month of the year.
    - You may derive more features, the above given are just a few examples.

Phase 2: Exploratory Data Analysis (EDA)

1. Visualizations

- **Delay Distributions:**
    - Histogram of delay durations.
- **Temporal Analysis:**
    - Line plots or bar charts showing delays across hours, days, or months.
- **Category-Wise Analysis:**
    - Group delays by airline, departure airport, or flight status.

2. Correlation Analysis

- The relationship between weather and flight data. (at least 3 different visualizations)

3. Comparison

- Compare delays across training and testing datasets to check for data consistency.

Phase 3: Analytical and Predictive Tasks

1. Classification Tasks

Binary Classification

1. Classify flights as **on-time** or **delayed** based on the following criteria:
    a. delay = 0: on-time
    b. delay > 0: delayed
2. Train one model for binary classification.
3. Evaluate performance using metrics like:
    a. Accuracy
    b. Precision-Recall
    c. F1-Score
    d. Class-wise Precision-Recall
    e. Confusion matrix.

Multi-Class Classification

1. Categorize flights into:
    - No Delay (0 min)
    - Short Delay (<45 min)
    - Moderate Delay (45–175 min)
    - Long Delay (>175 min)
2. Train one model for multi-class classification.
3. Evaluate performance using metrics like:
    - Accuracy
    - Precision-Recall
    - F1-Score
    - Class-wise Precision-Recall
    - Confusion matrix.

2. Regression Analysis

Delay Duration Prediction

Predict the exact delay duration for each flight.

1. Train one regression model.
2. Validate models using cross-validation techniques.
3. Evaluate performance using:
    - Mean Absolute Error (MAE)

        - Root Mean Square Error (RMSE)

Phase 4: Model Optimization and Evaluation

1. **Hyperparameter Tuning:**
    - Use techniques like grid search or random search to optimize predictive models.
2. **Validation:**
    - Apply k-fold cross-validation to assess model performance.
3. **Model Comparison:**
    - Compare the models.

Phase 5: Model Testing:

1. Use the trained models to make predictions on the test dataset.
2. Save predictions in the Kaggle submission format:
    - For regression, predict the exact delay.
    - For classification, predict delay categories or binary outcomes (on-time/delayed). (Your delay column must contain data in a string format, for example, “on-time” or “delayed”; **Do not write in 0 or 1**)
3. **Submission:**
    - Ensure the file meets Kaggle's submission requirements and upload it for evaluation.
    - Final Test.csv format for Kaggle Submission:
        - **File Name:** Indicates if data originated from train or test files.
        - **Flight Number:** Unique flight identifier.
        - **Type:** Type of flight (departure).
        - **Status:** Status of the flight (e.g., active, canceled).
        - **Departure IATA Code:** IATA code of the departure airport.
        - **Departure ICAO Code:** ICAO code of the departure airport.
        - **Scheduled Time:** Scheduled departure time.
        - **Arrival IATA Code:** IATA code of the arrival airport.
        - **Arrival ICAO Code:** ICAO code of the arrival airport.
        - **Arrival Estimated Time:** Estimated arrival time.
        - **Delay:** Calculated delay in minutes.
    - There are three Kaggle competitions for each model (Regression, Binary Classification, and Multi-Classification).