Project Report

Session Fall 2022 – BSAI

Submitted by:

- Safwan Hafeez 22I-0588

Submitted to:

Professor Hassan Raza

Department of Artificial Intelligence

National University of Computer

and Emerging sciences,

FAST

1. Introduction

In this project, we aim to predict flight delays using machine learning techniques. The dataset consists of flight and weather information, with the goal of predicting whether a flight will be delayed and, if so, to what extent. We employ a series of preprocessing, feature selection, and machine learning models to achieve accurate predictions.

We use several tools, including:

- **Data Preprocessing**: Cleaning, handling missing values, and transforming data for model readiness.
- **Feature Engineering**: Mapping weather data to flight data and selecting top features based on correlation.
- **Modeling**: Using Random Forest Classifiers and Regressors for both classification (binary and multiclass) and regression tasks.
- **Hyperparameter Tuning**: Optimizing the performance of the models using RandomizedSearchCV.
- **Evaluation**: Assessing model performance through various metrics and checking Kaggle scores.

2. Data Acquisition and Preprocessing

2.1 Reading Flight and Weather Data

The first step involved reading the flight data, including both training and test datasets, and the weather data. The flight data contains key attributes such as departure and arrival times, flight delays, and other relevant details. The weather data includes environmental conditions (temperature, precipitation, etc.) at the time of the flight.

2.2 Data Cleaning and Preprocessing

- Missing values were handled by either imputation or removal, depending on the severity of the missing data.
- Categorical columns were encoded using One-Hot Encoding, and numerical columns were standardized using StandardScaler to ensure that all features contribute equally to the model.
- The weather data was mapped to both the training and test flight data by matching the appropriate weather attributes to the corresponding flight times.

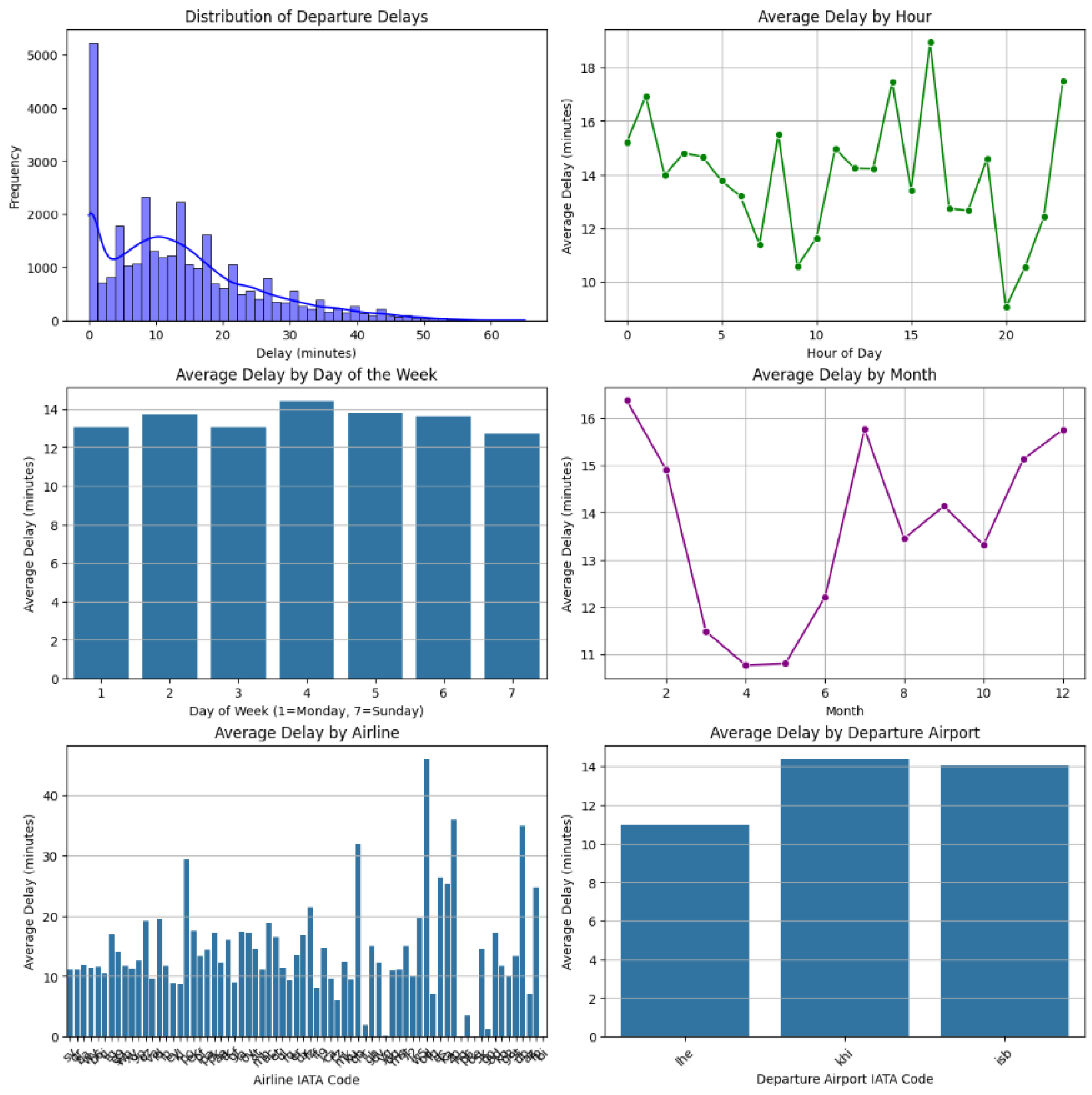
---

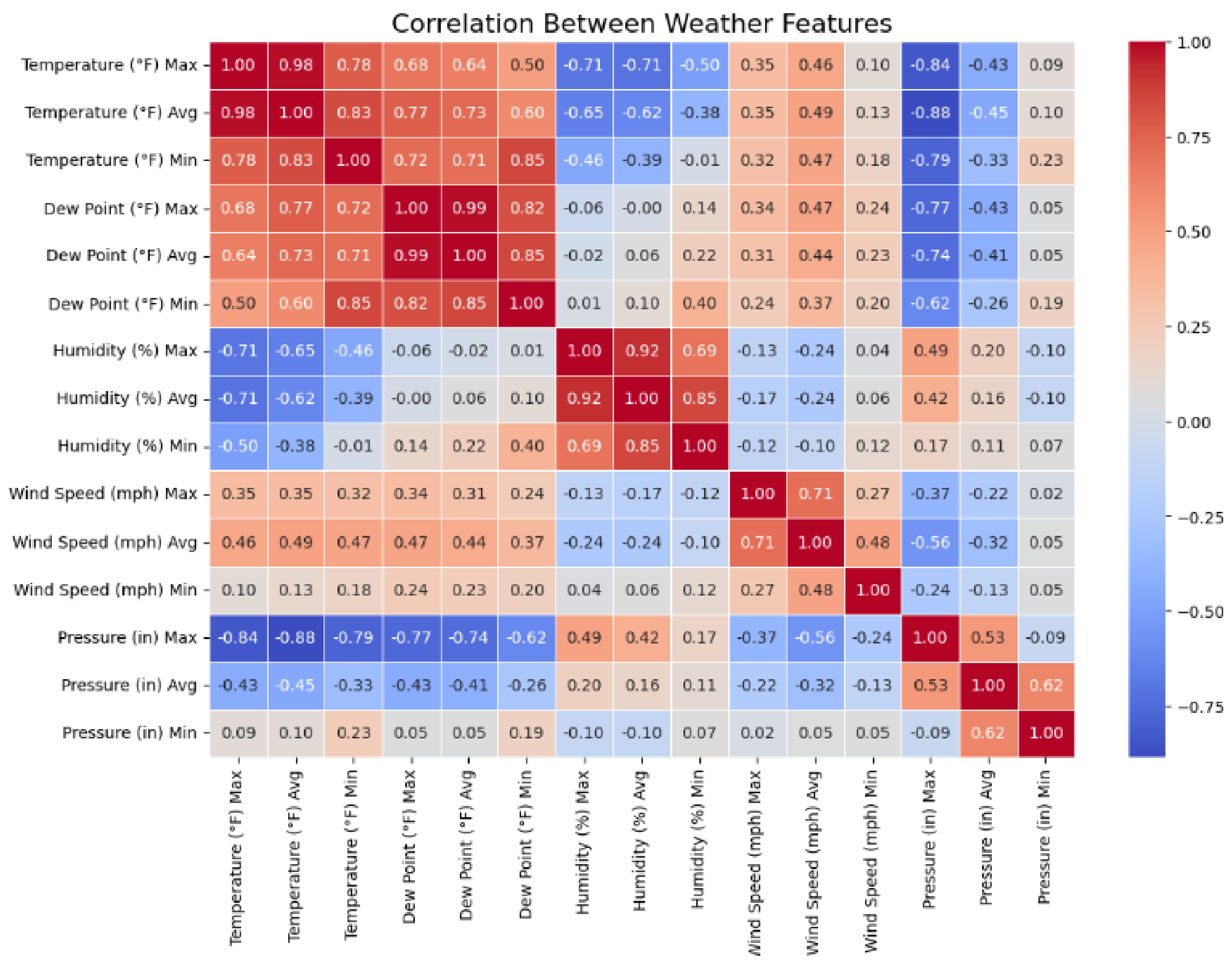
3. Feature Engineering

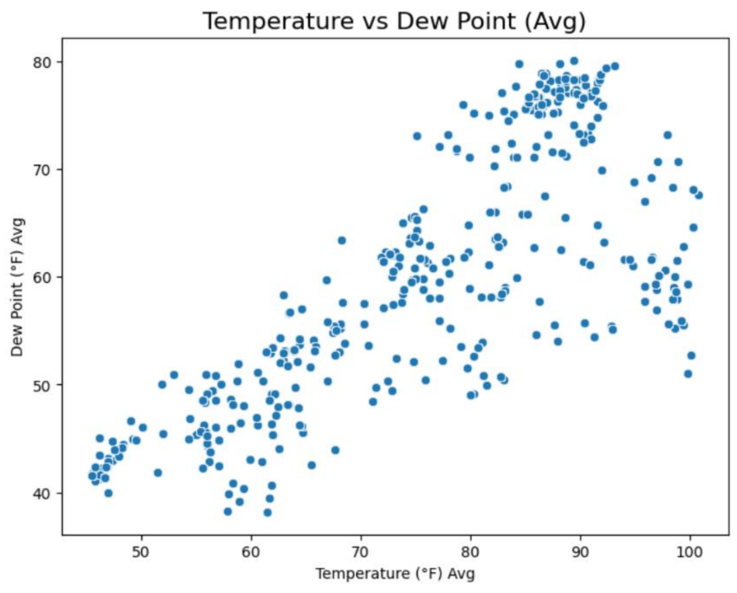
3.1 Selecting Relevant Features

To reduce dimensionality and enhance model performance, the top 5 most correlated features were selected based on correlation with the target variable (flight delay). Additionally, categorical variables such as day of the week, season, and weather-related features were retained for further analysis.

3.2 EDA and Visualization




Correlation Between Weather Features
Temperature (°F) Max 1.00 0.98 0.78 0.68 0.64 0.50 -0.71 -0.71 -0.50 0.35 0.46 0.10 -0.84 -0.43 0.09
Temperature (°F) Avg 0.98 1.00 0.83 0.77 0.73 0.60 -0.65 -0.62 -0.38 0.35 0.49 0.13 -0.88 -0.45 0.10
Temperature (°F) Min 0.78 0.83 1.00 0.72 0.71 0.85 -0.46 -0.39 -0.01 0.32 0.47 0.18 -0.79 -0.33 0.23
Dew Point (°F) Max 0.68 0.77 0.72 1.00 0.99 0.82 -0.06 -0.00 0.14 0.34 0.47 0.24 -0.77 -0.43 0.05
Dew Point (°F) Avg 0.64 0.73 0.71 0.99 1.00 0.85 -0.02 0.06 0.22 0.31 0.44 0.23 -0.74 -0.41 0.05
Dew Point (°F) Min 0.50 0.60 0.85 0.82 0.85 1.00 0.01 0.10 0.40 0.24 0.37 0.20 -0.62 -0.26 0.19
Humidity (%) Max -0.71 -0.65 -0.46 -0.06 -0.02 0.01 1.00 0.92 0.69 -0.13 -0.24 0.04 0.49 0.20 -0.10
Humidity (%) Avg -0.71 -0.62 -0.39 -0.00 0.06 0.10 0.92 1.00 0.85 -0.17 -0.24 0.06 0.42 0.16 -0.10
Humidity (%) Min -0.50 -0.38 -0.01 0.14 0.22 0.40 0.69 0.85 1.00 -0.12 -0.10 0.12 0.17 0.11 0.07
Wind Speed (mph) Max 0.35 0.35 0.32 0.34 0.31 0.24 -0.13 -0.17 -0.12 1.00 0.71 0.27 -0.37 -0.22 0.02
Wind Speed (mph) Avg 0.46 0.49 0.47 0.47 0.44 0.37 -0.24 -0.24 -0.10 0.71 1.00 0.48 -0.56 -0.32 0.05
Wind Speed (mph) Min 0.10 0.13 0.18 0.24 0.23 0.20 0.04 0.06 0.12 0.27 0.48 1.00 -0.24 -0.13 0.05
Pressure (in) Max -0.84 -0.88 -0.79 -0.77 -0.74 -0.62 0.49 0.42 0.17 -0.37 -0.56 -0.24 1.00 0.53 -0.09
Pressure (in) Avg -0.43 -0.45 -0.33 -0.43 -0.41 -0.26 0.20 0.16 0.11 -0.22 -0.32 -0.13 0.53 1.00 0.62
Pressure (in) Min 0.09 0.10 0.23 0.05 0.05 0.19 -0.10 -0.10 0.07 0.02 0.05 0.05 -0.09 0.62 1.00
Temperature (°F) Max
Temperature (°F) Avg
Temperature (°F) Min
Dew Point (°F) Max
Dew Point (°F) Avg
Dew Point (°F) Min
Humidity (%) Max
Humidity (%) Avg
Humidity (%) Min
Wind Speed (mph) Max
Wind Speed (mph) Avg
Wind Speed (mph) Min
Pressure (in) Max
Pressure (in) Avg
Pressure (in) Min
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Temperature vs Dew Point (Avg)
Dew Point (°F) Avg
Temperature (°F) Avg
80
70
60
50
40
50
60
70
80
90
100

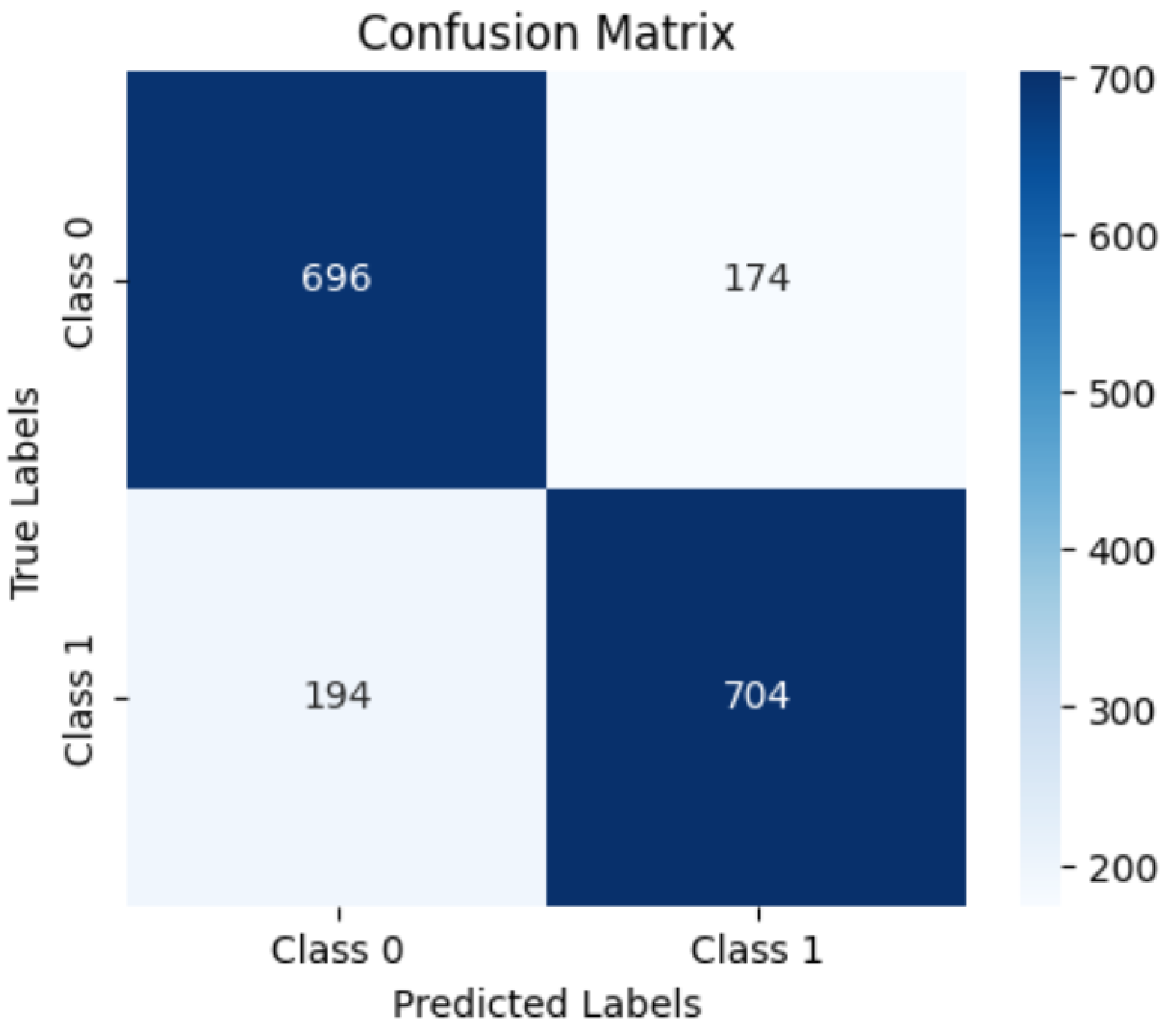
Graphs like correlation matrices, feature distributions, and weather-related visualizations were generated to understand the relationships between different features and flight delays.

---

4. Modeling and Evaluation

4.1 Random Forest Classifier for Binary Classification

The first step in modeling was to build a binary classifier to predict whether a flight would be delayed or not. The Random Forest Classifier was chosen for this task due to its robustness and ability to handle both categorical and numerical features. The model was trained using preprocessed data, and the performance was evaluated using accuracy, precision, recall, and F1 score.

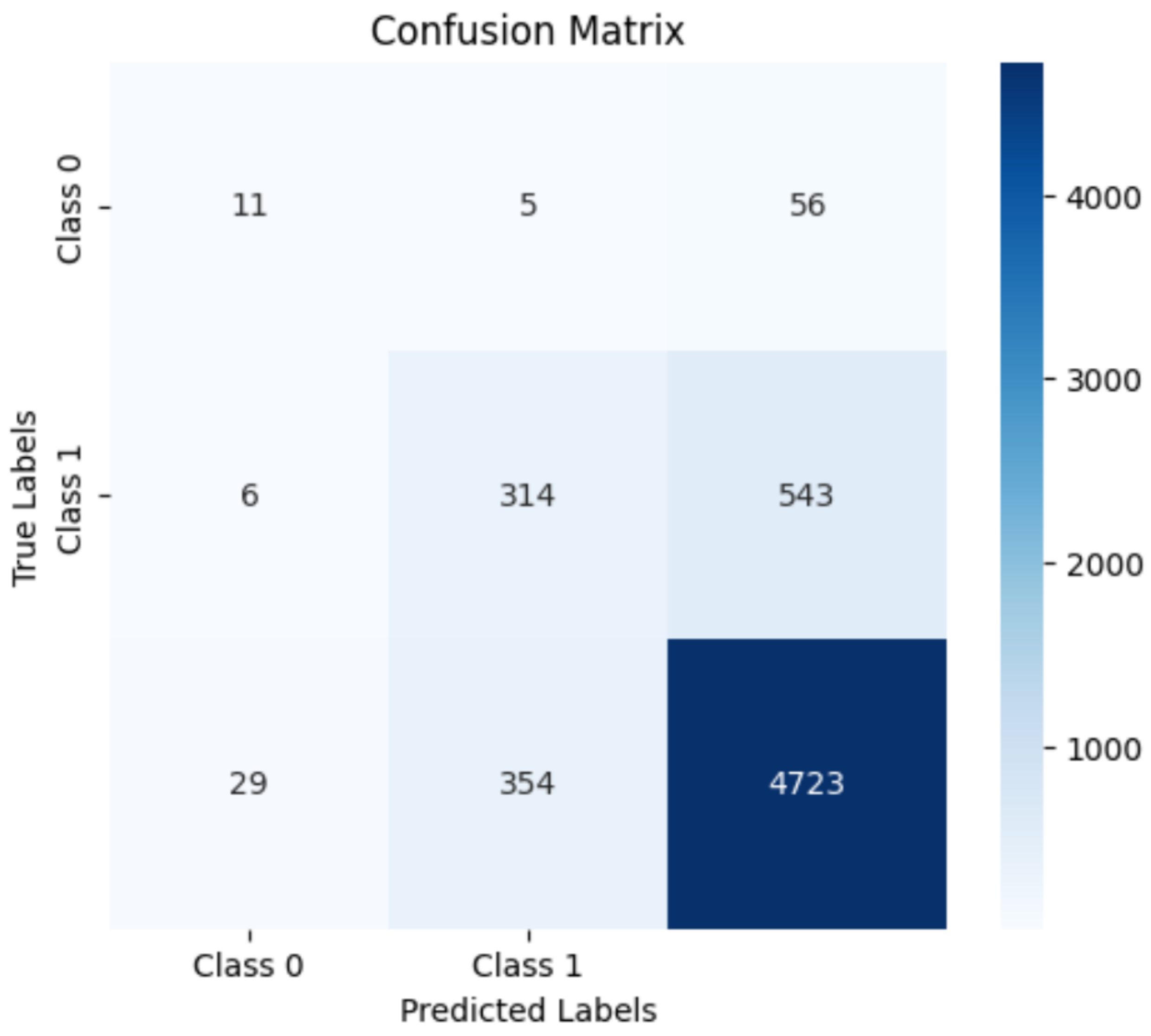


4.2 Random Forest Classifier for Multiclass Classification

Next, a multiclass Random Forest Classifier was trained to predict the extent of the flight delay. The delay categories were classified as:

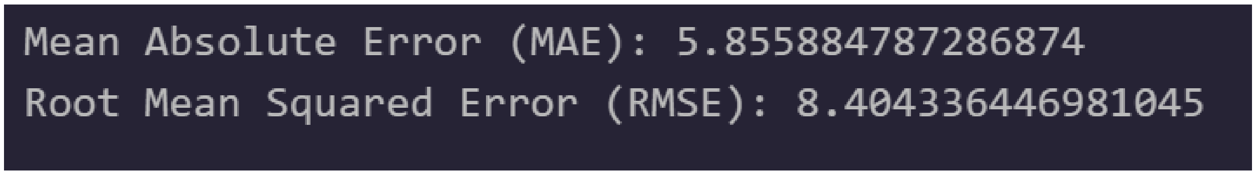
- **No delay**
- **Short delay**
- **Moderate delay**
- **Long delay**

The model's performance was evaluated using multi-class metrics such as accuracy and classification report, which provides a detailed breakdown of precision, recall, and F1 score for each class.



4.3 Random Forest Regressor for Exact Delay Prediction

A Random Forest Regressor was used to predict the exact delay time. This model helps calculate the continuous delay values rather than just the categorical class of delay. The performance was evaluated using mean squared error (MSE) and mean absolute error (MAE).

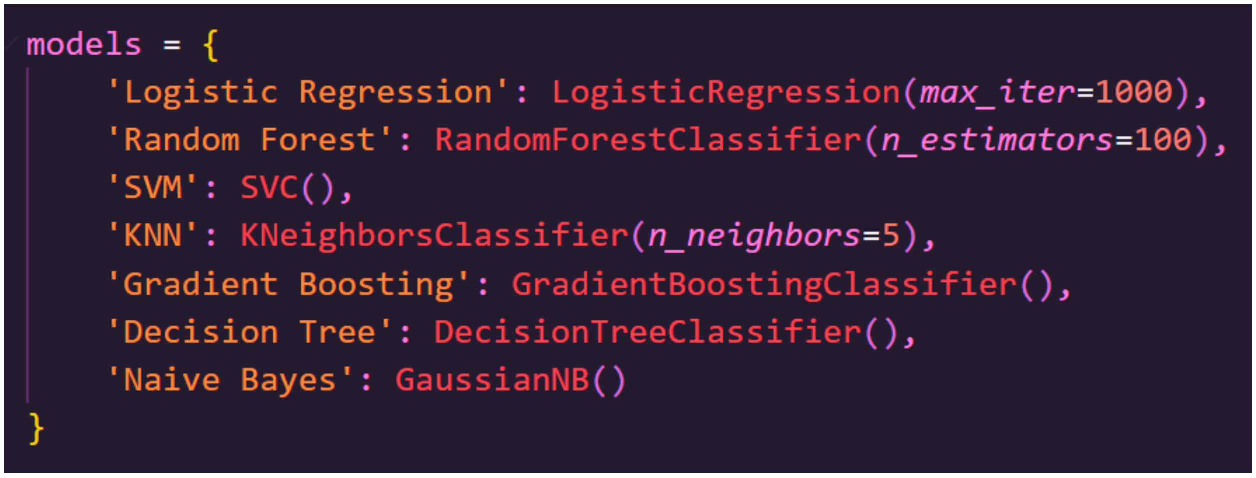


```
Mean Absolute Error (MAE): 5.855884787286874
Root Mean Squared Error (RMSE): 8.404336446981045
```

---

5. Hyperparameter Tuning

To optimize model performance, hyperparameter tuning was performed using RandomizedSearchCV. Various hyperparameters were tested, including the number of estimators, maximum depth, and minimum samples for splits. After identifying the best parameters, the models were retrained with these optimal settings.



```
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'Random Forest': RandomForestClassifier(n_estimators=100),
    'SVM': SVC(),
    'KNN': KNeighborsClassifier(n_neighbors=5),
    'Gradient Boosting': GradientBoostingClassifier(),
    'Decision Tree': DecisionTreeClassifier(),
    'Naive Bayes': GaussianNB()
}
```

---

6. Model Evaluation and Results

6.1 Model Performance on Test Data

After training the models with the optimal hyperparameters, the models were evaluated on the test dataset. The binary classifier, multiclass classifier, and regressor were all assessed on their performance.

The performance of the models was compared based on metrics like accuracy, precision, recall, F1 score, and mean squared error. Additionally, the Kaggle score was checked for the final model to evaluate its competitiveness in real-world applications.