Report on Flight Delay Prediction
2024F-T3 BDM 3603 - Big Data Framework 01 (DSMM Group 1)
Assignment -1

Submitted By:
Aishwarya Karki
Student ID: c0903073

Submitted to : Dr. Ishant Gupta

1. Dataset Description and Preprocessing Steps

Dataset Overview:

The dataset includes records related to flights and their associated delay times. It aims to predict whether a flight will be delayed based on attributes such as flight timings, airline, and weather conditions.

This dataset was extracted from Kaggle. https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022

The fields in the dataset provide a detailed view of each flight's schedule, airline, operational details, and more. Key attributes include:

1. **Flight and Airline Information**:
    - **FlightDate**: Date of the flight.
    - **Airline**: Code of the airline operating the flight.
    - **Origin & Dest**: Origin and destination airport codes.
    - **Flight_Number_Marketing_Airline**: Flight number assigned by the marketing airline.
    - **Marketing_Airline_Network**: Network of the marketing airline.
    - **Operating_Airline**: Airline operating the flight.
    - **Tail_Number**: Tail number of the aircraft.
2. **Schedule and Timing Information**:
    - **CRSDepTime & CRSArrTime**: Scheduled departure and arrival times.
    - **DepTime & ArrTime**: Actual departure and arrival times.
    - **DepDelay & ArrDelay**: Actual delays in minutes for departure and arrival.
    - **DepDelayMinutes & ArrDelayMinutes**: Delay durations in minutes.
    - **DepDel15 & ArrDel15**: Binary indicators of significant delays (15+ minutes).
    - **DepTimeBlk & ArrTimeBlk**: Categorical time blocks for scheduled times, helpful for peak-time analysis.
3. **Location and Route Information**:
    - **OriginAirportID & DestAirportID**: Unique IDs for origin and destination airports.
    - **OriginCityName & DestCityName**: Names of the cities.
    - **OriginState & DestState**: States for origin and destination airports.

    - **OriginWac & DestWac**: World area codes representing geographic regions for origin and destination.
4. **Operational Details**:
    - **Cancelled & Diverted**: Binary indicators for cancellation and diversion.
    - **TaxiOut & TaxiIn**: Taxi times on departure and arrival.
    - **WheelsOff & WheelsOn**: Times when the plane left the ground and touched down.
    - **ActualElapsedTime & CRSElapsedTime**: Actual and scheduled flight times.
    - **AirTime**: Time spent in the air.
5. **Distance and Group Information**:
    - **Distance**: Distance in miles between origin and destination.
    - **DistanceGroup**: Categorical grouping based on distance.
    - **DivAirportLandings**: Number of landings at diversion airports, useful for emergency rerouting analysis.
6. **Date and Time Periods**:
    - **Year, Quarter, Month, DayofMonth, DayOfWeek**: Attributes that allow for analysis of seasonal and day-of-week patterns affecting delays.

The dataset contains historical flight data records focused on predicting delays. Key attributes include flight timings, airlines, and other relevant details. This report uses PySpark DataFrames and MLlib for efficient handling and processing of large datasets

2. Data Cleaning and Preprocessing Steps

Data Cleaning

- **Handling Missing Values**: Imputation was used for critical columns, and rows with extensive missing values were removed.
- **Duplicate Removal**: Duplicates were dropped to maintain unique entries.

Feature Engineering

- **Time-Based Features**: Extracted features like **Hour of Day**, **Day of Week**, and **Month** to analyze delay patterns.
- **Binary Delay Indicator**: Transformed delay durations into a binary label (Delayed/Not Delayed) for classification.

- **Delay Pattern**: Added interaction features to assess how weather and time affect delays.

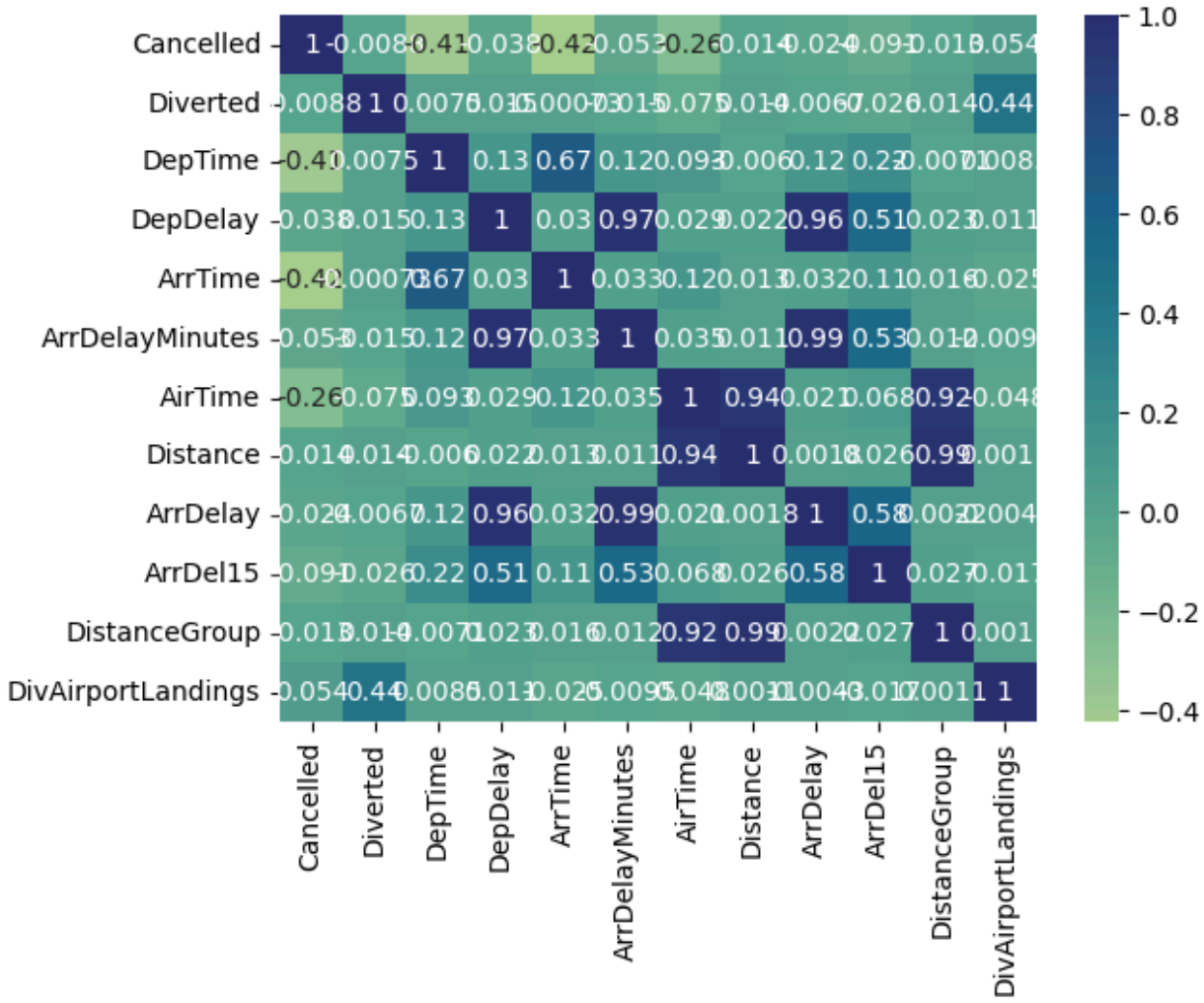
Encoding and Scaling

- String Indexer: Applied to Airline categorical for Airline Index feature.
- **One Hot Encoding**: Applied to categorical variables such as Airline Index.
- **Vector Assembler**: Standardized numerical features for model compatibility.

3. Key Insights from Data Analysis

Exploratory data analysis revealed:

- **Departure and Arrival Delay Correlation**: Departure delays are highly correlated with arrival delays, indicating that delays often propagate.



- **Time Patterns**: Delays are common during seasons and which month.
- **Seasonal and Weather Trends**: Winter months and adverse weather conditions are associated with more frequent delays.
- **Airline and Route-Specific Delays**: Certain airlines and routes have consistently higher delay frequencies.

4. Machine Learning Model Description

The models explored included logistic regression, random forest, and gradient boosting, implemented in a PySpark MLlib pipeline to handle preprocessing, feature engineering, and model training.

Models and Pipeline Configuration

1. **Logistic Regression**: Used as a baseline classifier.

This model underwent hyperparameter tuning with cross-validation.

5. Evaluation Metrics and Model Performance

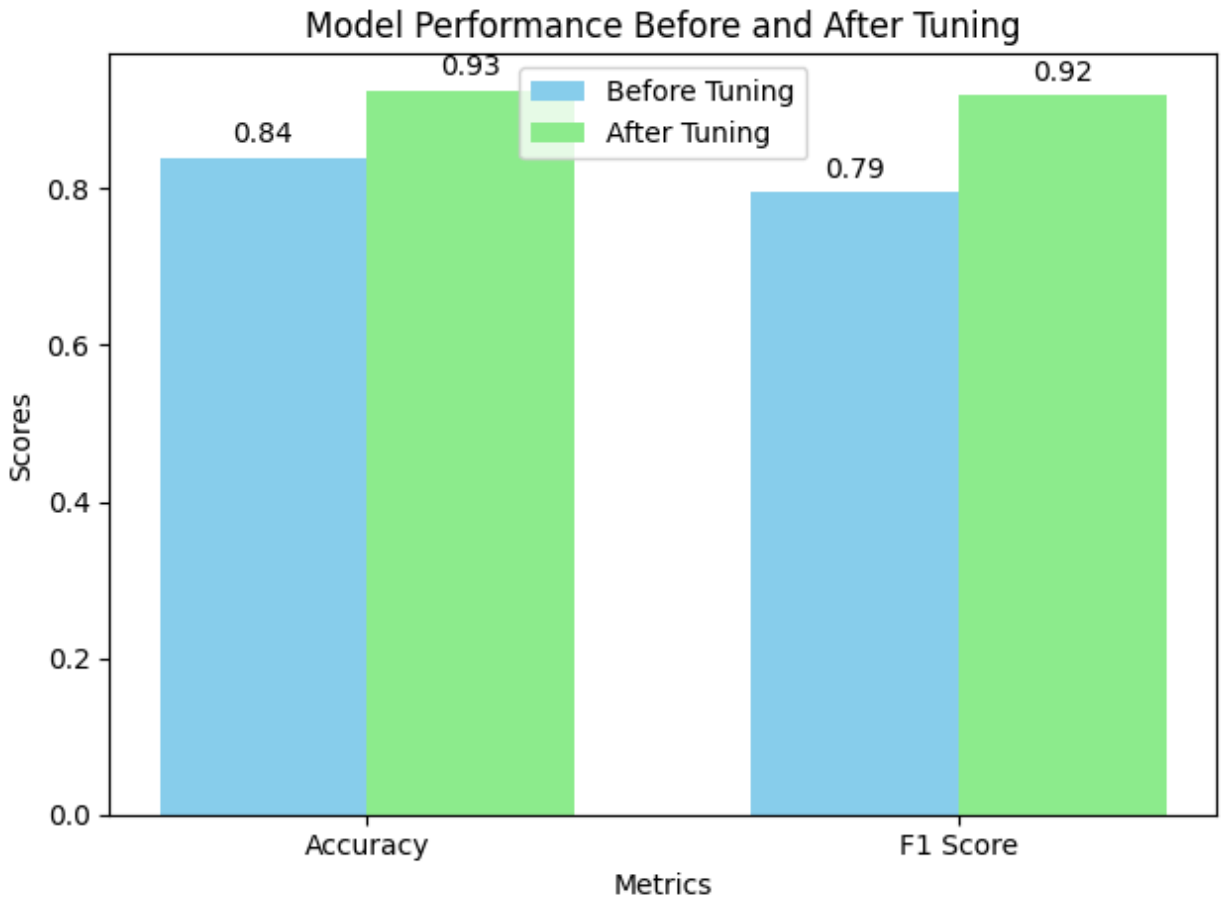
Initial Evaluation

- **Accuracy Before Tuning**: 84%
- **F1 Score Before Tuning**: 0.79

Final Evaluation After Tuning

After tuning, Gradient Boosting achieved the best results:

- **Accuracy After Tuning**: 93%
- **F1 Score After Tuning**: 0.92
- **AUC : 0.93**



A performance comparison plot showed the improvement in accuracy and F1 score after tuning, demonstrating the tuning's positive impact on model effectiveness.