Predicting Flight Arrival Delays & On-Time Performance

Background

In the airline industry, on-time performance (OTP) is a critical measure of service quality and operational efficiency. Airlines strive to minimize delays, as they impact passenger satisfaction, operational costs, and regulatory compliance. Despite sophisticated scheduling systems, flights frequently face delays due to various factors, including weather conditions, air traffic congestion, airline operational inefficiencies, and airport constraints.
Arrival delays not only cause inconvenience to passengers but also have a cascading effect on subsequent flights, leading to increased costs and reduced efficiency for airlines. Understanding and predicting flight delays is essential for optimizing flight schedules, improving resource allocation, and enhancing passenger experience.

In this project, you need to develop machine learning models to:

- Predict the arrival delay of a flight (Regression)
- Classify whether a flight will be on time or delayed (Classification)

3 datasets have been provided:

- Airlines Data
- Airports Data
- Flights Data

Refer to the data dictionary for more information.

By analyzing these datasets, you will extract key features and apply machine learning techniques to gain insights into the factors influencing flight delays.

Project Goals

1. Arrival Delay Prediction (Regression Task)

You will build regression models to predict arrival-delay based on relevant features.

Evaluation Metrics:

- Mean Absolute Error (MAE)
- R^2 Score (Coefficient of Determination)

Compare different regression models:

- Linear Regression
- Ensemble Techniques (Boosting and Bagging)

2. On-Time Performance (OTP) Classification

The Federal Aviation Administration (FAA) considers a flight "on time" if it arrives within 15 minutes of its scheduled arrival time. In this part of the project, you need to convert ARRIVAL_DELAY into a binary classification problem:

- On-Time (OTP) = 1 → ARRIVAL_DELAY ≤ 15 minutes
- Delayed = 0 → ARRIVAL_DELAY > 15 minutes

You will build classification models using:

- Logistic Regression
- Decision Trees
- Random Forest
- K-Nearest Neighbors (KNN)
- Ensemble Techniques (Bagging, Boosting)

Evaluation Metrics:

- Accuracy
- Precision, Recall, and F1-Score
- ROC Curve & AUC Score
- Confusion Matrix

Project Guidelines & Steps

Data Preprocessing & Feature Selection

- Handle missing values in flight records
- Encode categorical variables using appropriate methods (one-hot encoding, ordinal encoding, target encoding(mean encoding), Frequency encoding, using domain information).
- Perform feature selection to determine the most impactful variables.

Exploratory Data Analysis (EDA)

- Identify patterns and trends in flight delays.
- Visualize delays by airline, airport, time of day, and day of the week.
- Examine correlations between departure delay and arrival delay.

Model Training & Evaluation

- Train regression and classification models using different ML techniques.
- Compare performance using evaluation metrics.
- Use stratified k-fold cross-validation to assess generalizability.

Visualization & Interpretation

- Plot ROC curves and precision-recall curves for classification models.
- Use feature importance analysis to understand which factors contribute most to flight delays.

Comparison & Insights

- Compare the strengths and weaknesses of different models.
- Identify the best-performing models for both tasks and discuss trade-offs.

Expected Outcomes

- A well-tuned regression model that predicts arrival delays with minimal error.
- A classification model that accurately classifies on-time flights vs. delayed flights using performance metrics such as ROC-AUC and F1-score.
- A comparative analysis of different machine learning techniques applied to real-world aviation data.
- Visualizations & Insights that explain the key factors affecting flight delays.

Final Deliverable

Each learner/team will submit:

1. A Python notebook/script with:

a) Data preprocessing and feature engineering
b) Model training and evaluation
c) Performance comparison

2. A Presentation/Report summarizing:

a) Key findings
b) Model comparison results
c) Business implications of the predictions