Optimizing Aviation Infrastructure: Flight Delay Prediction to Support SDG 9

1. Introduction

Air transportation plays a crucial role in global economic development and connectivity. However, flight delays remain a persistent challenge, affecting passenger satisfaction, airline operational costs, fuel consumption, and overall infrastructure efficiency. With the increasing availability of large-scale aviation data, machine learning techniques provide an effective approach for analyzing complex patterns and supporting data-driven decision-making.

This project aligns with **United Nations Sustainable Development Goal 9 (Industry, Innovation, and Infrastructure)** by leveraging machine learning to improve the efficiency and resilience of aviation infrastructure. Accurate prediction of flight delays can assist airlines and airport authorities in optimizing resource allocation, improving scheduling, and reducing operational inefficiencies.

2. Problem Statement

Flight delays impact millions of passengers annually and result in significant economic losses to airlines and airports. Traditional analysis methods often struggle to capture the complex relationships between multiple factors such as flight schedules, airline operations, and temporal patterns. As a result, there is a need for an intelligent, data-driven solution that can analyze historical flight data and provide reliable delay predictions before flights depart.

3. Objectives

The objectives of this project are:

- To preprocess and clean a real-world aviation dataset.
- To perform feature transformation and feature selection to improve data quality.
- To train and evaluate multiple machine learning models for flight delay prediction.
- To compare model performance using appropriate evaluation metrics.
- To support SDG 9 by deriving insights that contribute to more efficient aviation infrastructure planning.

4. Dataset Description

- **Dataset Name:** Flight Delay Prediction Dataset (2018–2022)
- **Source:** Kaggle
- **Dataset Link:** https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022
- **Dataset Size:** Approximately 200,000+ flight records
- **Task Type:** Binary classification

Key Input Features:

- Month, DayofMonth, DayOfWeek (temporal features)
- UniqueCarrier (airline operator)
- Origin and Dest (departure and destination airport codes)
- DepTime (scheduled departure time)
- Distance (flight distance)

Target Variable:

- **DepDelay:** A binary target variable created as:
    - 0 = On-time flight (≤ 15 minutes delay)
    - 1 = Delayed flight (> 15 minutes delay)

5. Proposed Methodology

5.1 Data Preprocessing Plan

Data preprocessing will be conducted to ensure data quality and suitability for machine learning models. The planned steps include:

- **Data Cleaning:** Handling missing values, removing duplicate records, and excluding cancelled flights.
- **Data Transformation:**
    - Encoding categorical variables (e.g., UniqueCarrier, Origin, Dest) using One-Hot Encoding or Label Encoding.
    - Scaling numerical features such as Distance and DepTime using Min-Max Scaling.
- **Data Reduction:** Feature selection using correlation analysis to remove redundant or highly correlated features, improving model efficiency and training speed.

5.2 Machine Learning Models Plan

The following machine learning algorithms will be implemented and compared:

1. Logistic Regression (baseline model)
2. Decision Tree (for interpretability)
3. Random Forest (to model non-linear relationships)
4. Gradient Boosting / XGBoost (ensemble learning for improved performance)
5. k-Nearest Neighbors (kNN)

Hyperparameter tuning will be performed using Grid Search or Random Search techniques. Model performance will be evaluated using **Accuracy, F1-Score**, and **ROC-AUC**, with particular attention given to class imbalance in delayed versus on-time flights.

6. Preliminary Results

Preliminary experiments were conducted to validate the feasibility of the proposed approach. After completing data preprocessing, including missing value handling, encoding, scaling, and feature selection, a baseline Logistic Regression model was trained. The model achieved high accuracy and F1-score, indicating that the dataset contains meaningful patterns and is suitable for machine learning-based delay prediction. These results are preliminary, and further model refinement and evaluation will be conducted in later stages of the project.

7. Expected Outcomes

The expected outcomes of this project include:

- Identification of the most effective machine learning model for predicting flight delays.
- Improved understanding of key factors contributing to flight delays.
- Actionable insights that support more efficient aviation infrastructure management in line with SDG 9.

8. Project Deliverables and Timeline

- **Week 9:** Proposal submission and GitHub repository initialization
- **Week 10–12:** Data preprocessing pipeline development and initial model training
- **Week 13:** Hyperparameter tuning and comparative model analysis
- **Week 14:** Final report submission and code finalization

9. GitHub Repository

Project collaboration and version control will be managed using GitHub:

- **Repository Link:** https://github.com/Hasti-fgh/Flight-Delay-ML-Assignment

10. Conclusion

This proposal presents a feasible and well-structured machine learning project aligned with Sustainable Development Goal 9. By utilizing a real-world aviation dataset, applying systematic data preprocessing, and evaluating multiple machine learning models, the project is achievable within the given timeframe and is expected to contribute meaningful insights toward improving aviation infrastructure efficiency.