Project Report: Flight Delay Prediction Using Machine Learning

Introduction

Flight delays significantly impact airlines, passengers, and the aviation industry. The ability to accurately predict delays allows airlines to improve scheduling, reduce financial losses, and enhance customer satisfaction. This project aims to build and evaluate machine learning models using historical flight data to predict whether a flight will be delayed (result = 1) or not (result = 0).

Dataset Description

The dataset contains flight-level information with the following features: - MONTH (Numeric: 1–12) - ORIGIN_AIRPORT (Categorical) - DESTINATION_AIRPORT (Categorical) - SCHEDULED_DEPARTURE (Numeric, HHMM) - DEPARTURE_TIME (Numeric, HHMM) - SCHEDULED_TIME (Minutes) - AIR_TIME (Minutes) - DISTANCE (Miles) - SCHEDULED_ARRIVAL (Numeric, HHMM) - result (Target: 0 = on-time, 1 = delayed)

Data Preprocessing

1. Encoded categorical features (airports). 2. Converted HHMM times into hours of the day (SCHEDULED_HOUR, DEPARTURE_HOUR, ARRIVAL_HOUR). 3. Standardized numerical features (SCHEDULED_TIME, AIR_TIME, DISTANCE). 4. Train-test split: 80% training, 20% testing with stratification.

Models Implemented

The following classifiers were trained and evaluated: 1. Logistic Regression 2. K-Nearest Neighbors (KNN) 3. Multi-Layer Perceptron (MLP) 4. Decision Tree Classifier 5. Random Forest Classifier 6. LightGBM Classifier

Results Summary

- Logistic Regression: Accuracy ~88%, Precision high but Recall low (~33%). - KNN: Accuracy ~96%, good balance, but computationally heavy. - MLP: Accuracy ~95%, Recall ~80%, needs tuning. - Decision Tree: Accuracy ~100%, but overfits. - Random Forest: Accuracy ~99%, best balance, high generalization. - LightGBM: Accuracy ~97%, efficient, slightly lower recall than RF.

Model Comparison

See table below.

Conclusion

The Random Forest Classifier achieved the best performance with high precision and recall, making it the most suitable choice for deployment. LightGBM is also a strong candidate due to its efficiency. Logistic Regression and MLP underperformed in recall. Decision Tree overfitted, while

KNN is accurate but computationally expensive.

Recommendations

1. Deploy Random Forest or LightGBM. 2. Handle class imbalance using oversampling/SMOTE. 3. Explore feature importance (distance, scheduled hour). 4. Use ensembles (stacking RF + LightGBM). 5. Integrate into airline scheduling systems for real-time predictions.

Model	Accuracy	Precision	Recall	Remarks
Logistic Regression	~88%	~99%	~33%	Struggles with delays
KNN	~96%	~94%	~82%	Good, but slow on large data
MLP	~95%	~90%	~80%	Decent, needs tuning
Decision Tree	~100%	~100%	~100%	Overfitting
Random Forest	~99%	~99%	~97%	Best overall
LightGBM	~97%	~99%	~85%	Efficient alternative