ADANA ALPARSLAN TÜRKEŞ
BİLİM VE TEKNOLOJİ ÜNİVERSİTESİ

FLIGHT ROUTE OPTIMIZATION

Gizem EROL, Selvinaz Zeynep KIYIKCI, Zeynep Sude KIRMACI

Introduction

This project aims to determine the most optimized route in flight data using data analytics and machine learning techniques. Within the scope of the project, different machine learning algorithms such as SVM, Linear Regression, Decision Tree, Logistic Regression and Naive Bayes were evaluated. As a result of the experiments and result analysis, it was observed that Linear Regression and Decision Tree algorithms performed the best in determining the optimised route.

Dataset and Features

1.Source: The primary data source for this project is the "**flights.csv**" dataset available on GitHub.

2.Key Features Used:

- **DistanceKilometers:** Flight distance in kilometers.
- **AvgTicketPrice:** Average ticket price.
- **FlightTimeHour/Min:** Flight time.
- **Weather Data:** Origin and destination weather conditions.
- **FlightDelayMin:** Actual delay times.

3.Preprocessing Steps:

- Handling missing data (e.g., filling delays with 0).
- Normalization for numerical features.
- One-hot encoding for categorical features like weather.

Best Model 1 : Linear Regression

Linear Regression was chosen as one of the key models for this project due to its simplicity, interpretability, and effectiveness in predicting numerical outcomes. It is particularly well-suited for analyzing the relationship between a continuous dependent variable (e.g., flight delay time) and one or more independent variables (e.g., flight distance). In this case, the model predicts the likelihood of flight delays using **DistanceKilometers** as the primary predictor.

Despite its limitation in capturing non-linear relationships or complex interactions among features, Linear Regression's strong performance in terms of low error rates (MSE: 0.17) and high accuracy (80.61%) demonstrated its suitability for identifying optimized flight routes.

Analysis and Discussion

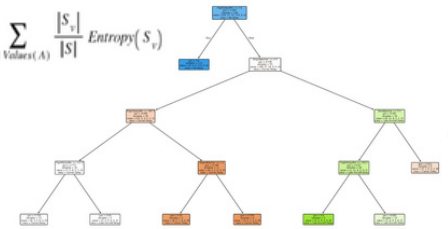
Linear Regression: By modelling the linear relationships of flight routes, it achieved a test accuracy of 80.61% and a low error rate (MSE: 0.17), which is good for optimised route identification.
Logistic Regression: The categorical classification approach performed relatively poorly with a test accuracy of 54.08% and an F1 score of 0.58.
Naive Bayes: Under the assumption of independent variables in the data, it achieved a moderate level of success with a test accuracy of 56.12% and an F1 score of 0.58.
SVM (Support Vector Machine): The model, which aims to discriminate between the data, showed a reasonable performance with a test accuracy of 56.12% and an F1 score of 0.59.
Decision Tree: It was the highest performing model with 92.31% test accuracy and 89.22% cross validation success.

Model	Train Accuracy (%)	Test Accuracy (%)	Cross Validation Accuracy (%)	MSE	F1 Score
Linear Regression	76.55	80.61	77.78	0.17	-
Logistic Regression	53.12	54.08	53.40	0.24	0.58
Naïve Bayes	55.97	56.12	52.85	0.25	0.58
SVM	57.67	56.12	57.65	0.69	0.59
Decision Tree	90.35	92.31	89.22	-	-

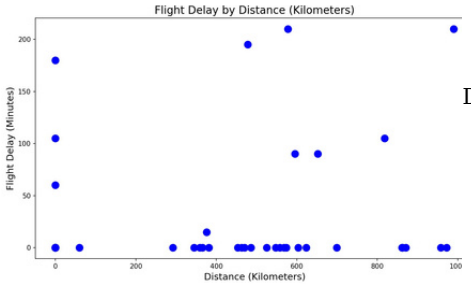
Best Model 2 : Decision Tree

The Decision Tree algorithm was chosen for its impressive test accuracy of 92.31% and cross-validation accuracy of 89.22% in classifying flight delays. It effectively analyzes various features like distance and weather, providing clear and interpretable results. Its high performance and ease of use made it one of the best models for flight route optimization

$$G(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$



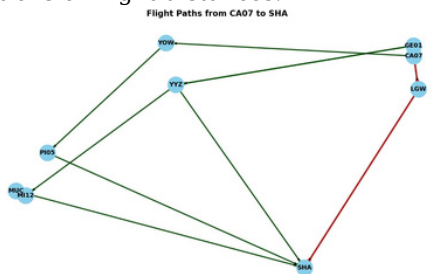
Flight distance vs Flight Delay Scatter Plot



This plot visualizes the relationship between flight distance (x-axis: DistanceKilometers) and delay time (y-axis: FlightDelayMin). Each point represents a flight. Patterns in the plot help analyze how delays vary with distance, aiding in delay prediction and route optimization.

Result

This study successfully implemented predictive models to estimate flight delays and classify delay types, addressing key challenges in air travel management. The Linear Regression model proved effective for predicting delay durations based on numerical features such as distance and flight time, while the Decision Tree model provided accurate categorization of delay types. The integration of these models into a directed flight graph further allowed the calculation of optimal routes based on travel efficiency metrics, such as minimizing delay durations or flight distances.




```
Origin Airport ID: CA07
Destination Airport ID: SHA
Optimal Route: ['CA07', 'LGW', 'SHA']
Total Distance for Optimal Route: 10749.59 km
Predicted Delay (min) for Optimal Route: 0.25
Delay Status for Optimal Route: No Delay

Alternative Routes:
Alternative Route 1: ['CA07', 'GE01', 'YYZ', 'SHA']
  Total Distance: 18728.76 km
  Total Delay: 0.00 minutes
```