CS 677 Final Project

Spring 2023

Zuowen Tang

Final Project Description

# Project Introduction

The objective of this dataset analysis project is to explore the factors contributing to flight delays and develop predictive models that can accurately forecast such delays. Flight delays have long been a significant issue, including causing passengers inconvenience, airline schedule disruption, and operational costs. Ideally, by understanding the underlying causes and patterns of flight delays, I expect to provide valuable thoughts that can assist airlines and airports in devising effective strategies to minimize delays and improve the overall travel experience.

# Dataset Introduction

There are a total of 539,383 instances and six key features that have been used in this project (see table below). The irrelevant data have been removed in order to increase the accuracy of the analysis and prediction.

|  |  |
| --- | --- |
| Features | Description |
| Airline | American Airlines, Delta Airlines, Alaska Airlines, United Airlines, etc.. |
| Airport From / To | BOS - Boston Logan International Airport, IAD - Dulles International Airport, JFK - John F. Kennedy International Airport, etc.. |
| Day Of Week | Monday to Sunday |
| Time | Duration of the flight |
| Length | Types of the aircraft |

# Methods and Algorithms

There are a total of eight different algorithms used in this project. They are Logistic Regression, k-NN, Naive Bayesian, LDA, QDA, Decision Tree, Random Forest, and SVM (Linear, Gaussian, Poly.)

# Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Accuracy** | **TPR** | **TNR** |
| Logistic Reg. | 66.02% | 61.28% | 64.24% |
| k-NN (k = 1) | 63.98% | 54.29% | 59.63% |
| k-NN (k = 3) | 65.25% | 56.98% | 61.73% |
| k-NN (k = 5) | 65.58% | 57.99% | 62.43% |
| Naive Bayesian | 63.98% | 54.29% | 59.63% |
| LDA | 65.25% | 56.98% | 61.73% |
| QDA | 65.58% | 57.99% | 62.43% |
| Decision Tree | 66.91% | 45.19% | 46.95% |
| Random Forest | 65.94% | 60.55% | 63.87% |
| Linear SVM | 63.07% | 65.26% | 59.32% |
| Gaussian SVM | 64.3% | 65.29% | 62.27% |
| Poly SVM | 64.85% | 64.91% | 64.68% |

# Conclusion

In conclusion, this dataset analysis project on flight delays has yielded limited results in the low correlations between different features. Most of the prediction algorithms used in this project demonstrated medium accuracy in forecasting delays, which suggests that the available data may not provide a comprehensive picture of the factors influencing flight delays.

Incorporating weather data into this analysis project could enhance the predictions’ accuracy and lead to more meaningful insights. If possible, future research should consider accessing weather datasets to explore further the possible impact of weather conditions on flight delays.