# Abstract

Flight delays are on a steady rise, presenting significant financial challenges and fostering dissatisfaction among airline companies' customers. Addressing this pressing issue, the study employed supervised machine learning models to forecast flight delays. The data set that records information of flights departing from JFK airport for one year was used for the prediction. Seven algorithms (Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree) were trained and tested to complete the binary classification of flight delays. The evaluation of algorithms was fulfilled by comparing the values of four measures: accuracy, precision, recall, and f1-score. These measures were weighted to adjust the imbalance of the selected data set. The comparative analysis showed that the Decision Tree algorithm has the best performance with an accuracy of 0.9777, and the KNN algorithm has the worst performance with an f1-score of 0.8039. Tree-based ensemble classifiers generally have better performance over other base classifiers.

# Introduction

As people increasingly choose to travel by air, the number of flights that fail to take off on time also increases. This growth exacerbates the crowded situation at airports and causes financial difficulties within the airline industry. Air transportation delay indicates the lack of efficiency of the aviation system. It is a high cost to both airline companies and their passengers. The overall economic impact of airline delays is a critical aspect of the aviation industry, as highlighted in several **studies**. The annual economic impact of airline delays was estimated to be **$31.2 billion** in 2010 [6], while other recent studies estimated these expenditures to be **$40.2 billion** [7]. Hence, forecasting flight delays holds the potential to enhance airline operations and elevate passenger satisfaction, thereby bringing a positive impact on the economy.

In this study, our main objective is to compare the performance of machine learning algorithms in predicting flight delays. Delving into the realm of Airline Delay prediction, we aim to harness the power of ML algorithms to distinguish and evaluate critical factors by uncovering patterns and relationships among them, which significantly impact the success of delay prediction. Our goal is to gauge the effectiveness of these algorithms in predicting and optimizing the success of airline prediction.

# Data

The dataset is sourced from Kaggle and encompasses a vast repository of airline delay and cancellation data, offering a comprehensive perspective on the operational challenges faced by the aviation industry from 2009 to 2018. For our study, we have specifically focused on data from the years 2015 and 2016, providing insights into the performance and reliability of air transportation services during this period. The dataset consists of a total of 11,437,737 rows and 28 columns, encompassing data from the years 2015 and 2016. This includes a broad spectrum covering temporal dynamics, Operational details, Geospatial Information, and delay classifications. The dataset comprises Integer, Categorical, and Binary data types, reflecting various essential information crucial for investigating Airline delays. This extensive dataset presents with abundant opportunities for thorough examination and analysis, facilitating a detailed exploration of airline delays and cancellations.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Name** | **Datatype** | **Sample Values** | **Feature Description** |
| FL\_DATE | Object | 2015-01-01 | The Date of the Flight |
| OP\_CARRIER | Object | 'NK', 'MQ', 'OO', 'EV', ‘HA' | The Name of the Carrier |
| OP\_CARRIER\_FL\_NUM | Int64 | 195, 197, 198 | Flight Number of the Carrier |
| ORIGIN | Object | 'MCO', 'LGA', 'FLL', 'IAH' | Origin Airport |
| DEST | Object | 'FLL', 'MCO', 'LAS', 'ORD' | Destination airport |
| CRS\_DEP\_TIME | Int64 | 2147, 1050, 700 | Schedule Departure Time (HHMM) |
| DEP\_TIME | float64 | 2147, 1050, 700 | Actual Departure Time (HHMM) |
| DEP\_DELAY | float64 | -4., 14., 12. | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| TAXI\_OUT | float64 | 15., 20., 19., 8. | Taxi Out Time, in Minutes; The time elapsed between departure from the origin airport gate and wheels off. |
| WHEELS\_OFF | float64 | 2158., 1124., 731. | Wheels Off Time (local time) in HHMM |
| WHEELS\_ON | float64 | 2158., 1124., 731. | Wheels On Time (local time) in HHMM |
| TAXI\_IN | float64 | 7., 9., 10., 4., 5. | Wheels down and arrival at the destination airport gate, in minutes |
| CRS\_ARR\_TIME | Int64 | 2250, 1404, 757 | Scheduled Arrival time (HHMM) |
| ARR\_TIME | float64 | 2245., 1403., 813. | Actual Arrival time (HHMM) |
| ARR\_DELAY | float64 | -5.0, -1.0, 16.0 | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| CANCELLED | float64 | 0., 1. | Cancelled Flight Indicator (1=Yes); was the flight cancelled? |
| CANCELLATION\_CODE | Object | 'A', 'B', 'C', 'D' | Reason for cancellation (A = carrier, B = weather, C = NAS, D = security |
| DIVERTED | float64 | 0., 1. | Diverted Flight Indicator (1 = Yes) |
| CRS\_ELAPSED\_TIME | float64 | 63., 194., 57., 196. | Estimated Elapsed Time of Flight, in Minutes |
| ACTUAL\_ELAPSED\_TIME | float64 | 63., 194., 57., 196. | Elapsed Time of Flight, in Minutes |
| AIR\_TIME | float64 | 40., 150., 32., 164. | Flight time in Minutes |
| DISTANCE | float64 | 177., 1076., 1222. | Distance between airports (miles) |
| CARRIER\_DELAY | float64 | 1., 15., 127., 174. | Carrier Delay, in Minutes |
| WEATHER\_DELAY | float64 | 31., 17., 24., 61. | Weather Delay, in Minutes |
| NAS\_DELAY | float64 | 16., 18., 25., 19. | National Air System Delay, in Minutes |
| SECURITY\_DELAY | float64 | 8., 21., 6., 14. | Security Delay, in Minutes |
| LATE\_AIRCRAFT\_DELAY | float64 | 8., 29., 21., 10. | Late Aircraft Delay, in Minutes |

# Methodology