ISyE 6740 – Spring 2023
Final Report

Team Member Names:
Maynard Emmanuel B. Miranda (GT ID: 903652021)
Alfonzo Miguel Sabado (GT ID: 903656898)

Project Title: U.S. Flight Delay and Cancellation Prediction

Problem Statement

Flight delays and cancellations are a significant problem for both travelers and airline companies. For travelers, it can cause inconvenience and missed events. For airline companies, it may cause profit loss and reputation damage. Predicting flight delays and cancellation is challenging due to various situations or events that can happen before the flight such as weather problems, aircraft downtime/mechanical problems, queuing issues, system problems, crew scheduling issues, and air traffic congestion. The goal of this project is to help address this problem, and improve an airline's ability to cancel flights in advance in order to improve:

(1) Customer experience. Canceling flights 4 hours before the flight, allows customers to avoid traveling to the airport only to find their flights canceled or moved. It also enables customers to rebook flights earlier, and find better routes.

(2) Planning of route recovery. By being more intentional on the flights/routes that are canceled, airlines can better plan how to get back to 100% operational capacity, with the least amount of delays.

The study focuses on the 10 busiest U.S. airports - LAX, ORD, ATL, DEN, DFW, SFO, SEA, MCO, EWR, and BOS. The team performed data analysis to determine weather patterns that cause flights to: arrive on-time or earlier, be slightly delayed by 30 minutes or less, face moderate delays of more than 30 minutes but less than 2 hours, encounter long delays of more than 2 hrs, or be canceled altogether

The team used k-nearest neighbors, Naive Bayes, Decision Trees, and Random Forest classification models to predict the flight results. Hyperparameter experimentation was performed, and various feature engineering techniques discussed in the class were explored to find the optimal feature combination that should be used in the model.

Dataset

The data used for this project is the Airline On-Time Performance dataset created and maintained by the Bureau of Transportation Statistics. The data contains the United States domestic flight information from 1987 to 2023 and has multiple features such as airline carrier, airport departure and destination, arrival time, delay times, cancellation information, etc. In addition to flight data, the team also used the airport's local weather data from National Centers for Environmental Information (NCEI).

For this study, the flight data used is for the top 10 busiest U.S airports from the period 2021 to 2022. This dataset was then merged with weather data gathered at the specific airport at the time of boarding, or about 1 hour before the scheduled flight, to give us additional features: temperature (dry bulb, wet bulb), precipitation, visibility, humidity, sea level pressure altimeter, dry bulb temperature, precipitation, sea level pressure, wind speed and direction, to predict a flight's final result.

During the data merging process, we employed data cleansing processes such as removing irrelevant data and null data, fixing feature data types, and grouping data based on flight time delay, to make the final dataset ready for analysis.

Methodology

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to identify patterns and trends in the data. Based on the result which can be found on Figure 1, the percentage of On Time/Early flights was slightly higher in 2021 than in 2022, while the reverse is true for Delayed and Canceled flights.

It must be noted that airline companies and airports reduced the number of employees in 2021 due to Covid-19 pandemic, and may have been one of the reasons when the number of flights began to return to normal in 2022.

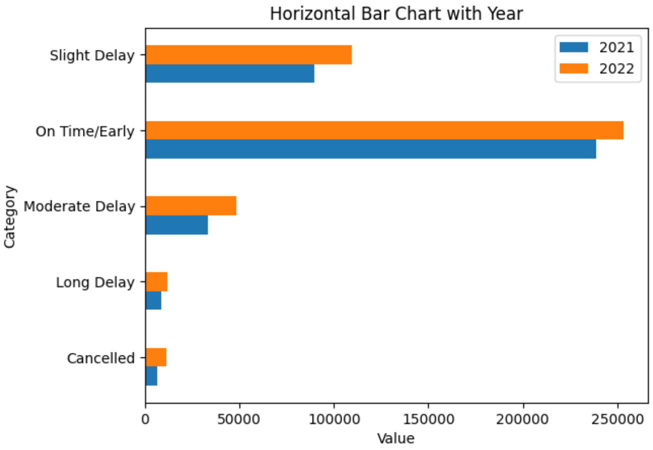


Figure 1: Flight Results vs Year

Figure 2 illustrates the number of flights per month and their corresponding flight results distribution. Based on the chart, it can be inferred that the number of flights increases from January to December, with October having the highest number of flights and February the lowest. The flight results distribution looks consistent having just a small increase or decrease overtime.

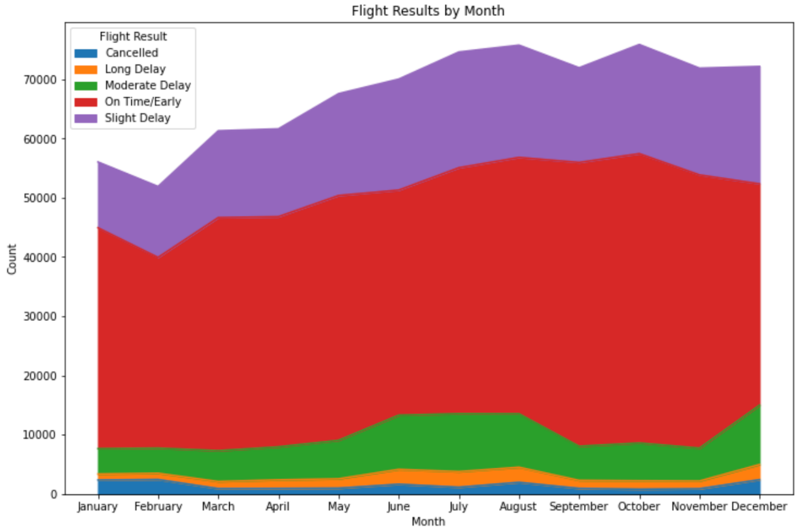


Figure 2: Flight Results vs Month

Figure 3 shows the number of flights per boarding hour and flight result distribution per boarding hour. The chart does not have 'Cancelled' flights since the flights were cancelled and they never had a chance to board and the expected boarding time was not provided on the raw dataset. The top 3 peak hours are 7 am, 6 am, and 9 am. The hours can be categorized into the following:

- Busiest: 5 am to 6 pm
- Busy: 4 am and 7 pm to 11 pm
- Not busy: 12 am to 3 am

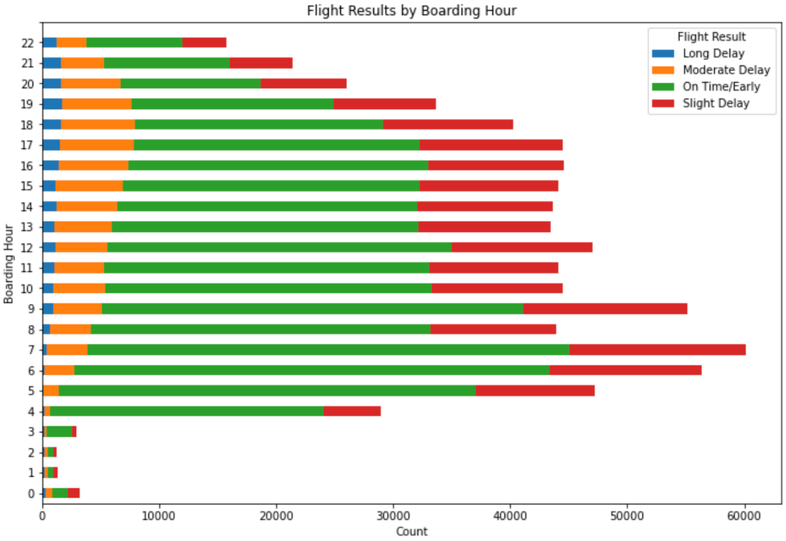


Figure 3: Flight Results vs Boarding Hour

Figure 4 provides insights into the number of flights based on their distance and their corresponding flight result distribution. It is evident that the majority of the flights cover a distance between 500 to 1249 miles. The flight result distribution shows variations, but the differences are minimal.

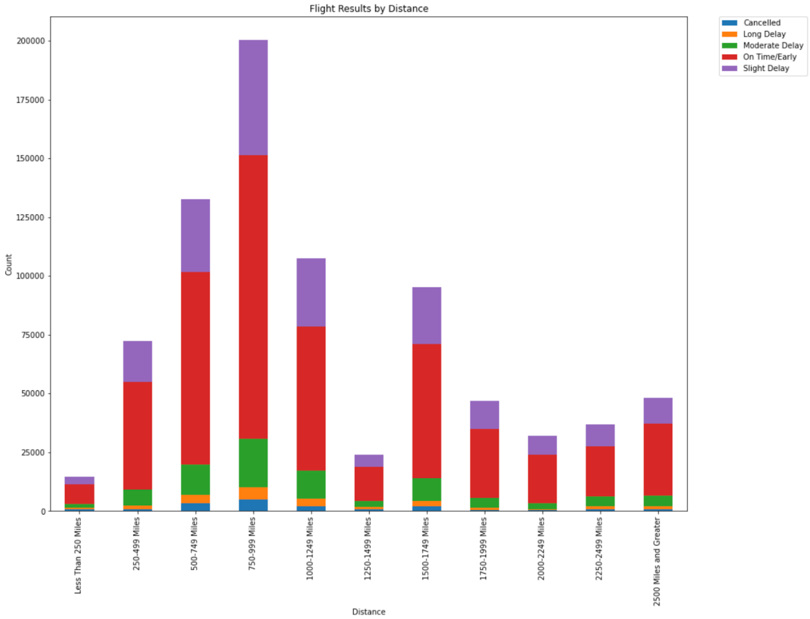


Figure 4: Flight Results vs Distance

Figure 5 illustrates the number of flights per day and their corresponding flight result distribution. The chart reveals that the aviation industry is consistently busy, with a relatively constant flight result distribution across all days. Although there are slight variations in the distribution, the changes appear to be minor and within a narrow range.

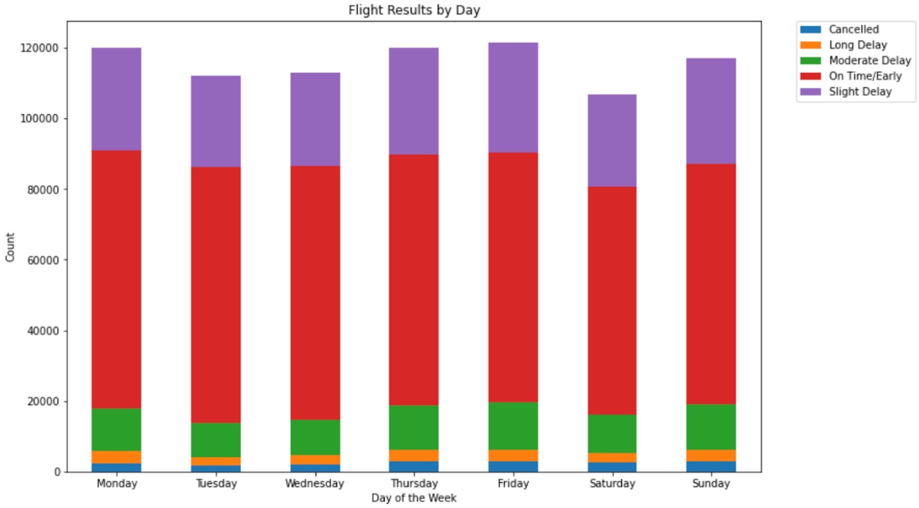


Figure 5: Flight Results vs Day

Figure 6 displays the number of flights for each airline company and their respective flight result distribution. United Airlines, Delta Airlines, and American Airlines are the major companies that operate in the 10 busiest U.S. airports. Based on the chart, Delta Airlines and American Airlines have almost the same number of flights but American Airlines has more flight cancellations and delays compared to Delta Airlines.

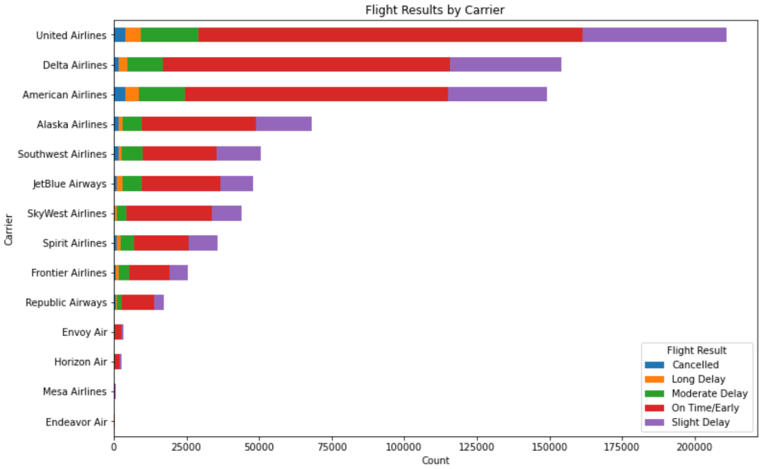


Figure 6: Flight Results vs Carrier

Figure 7 displays the number of flights for each airport and their respective flight result distribution. The chart highlights that LAX, ORD, ATL, and DEN are among the top airports in terms of flight volume. However, a closer look at the delay and cancellation distribution reveals that EWR and MCO are leading in this aspect.

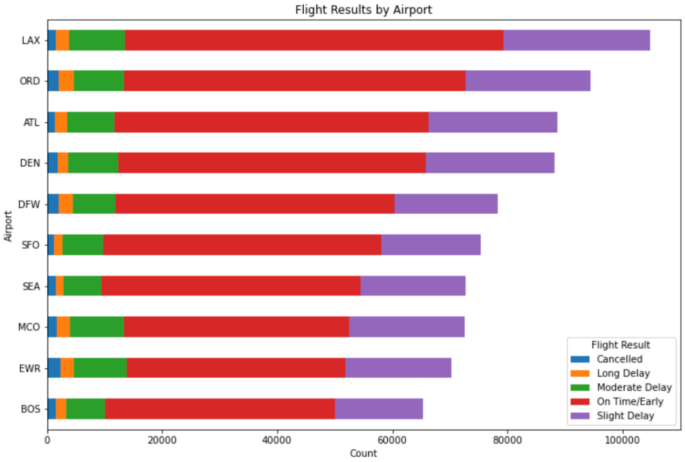


Figure 7: Flight Results vs Airport

There are several weather measurements available on the dataset but the listed weather measurements below show a significant relationship from the response variable. For example refer to Figure 8,, for departure airport wind speed, if it goes above a certain value, the flight is likely to become moderate delay, long delay, or canceled.

- Altimeter setting: This is the barometric pressure at a specific location, adjusted to sea level. It is used in aviation to determine the altitude of an aircraft above sea level.
- Dew point temperature: This is the temperature at which air becomes saturated with water vapor and condensation begins to form. It is a measure of the amount of moisture in the air.
- Dry bulb temperature: This is the ambient air temperature, measured using a thermometer that is not affected by moisture.
- Relative humidity: This is the amount of moisture in the air compared to the maximum amount it could hold at a given temperature.
- Wet bulb temperature: This is the temperature of a thermometer that is wrapped in a wet cloth and exposed to moving air. It is a measure of the humidity and air movement in the environment.
- Wind speed: This is the rate at which air is moving horizontally past a fixed point on the ground.

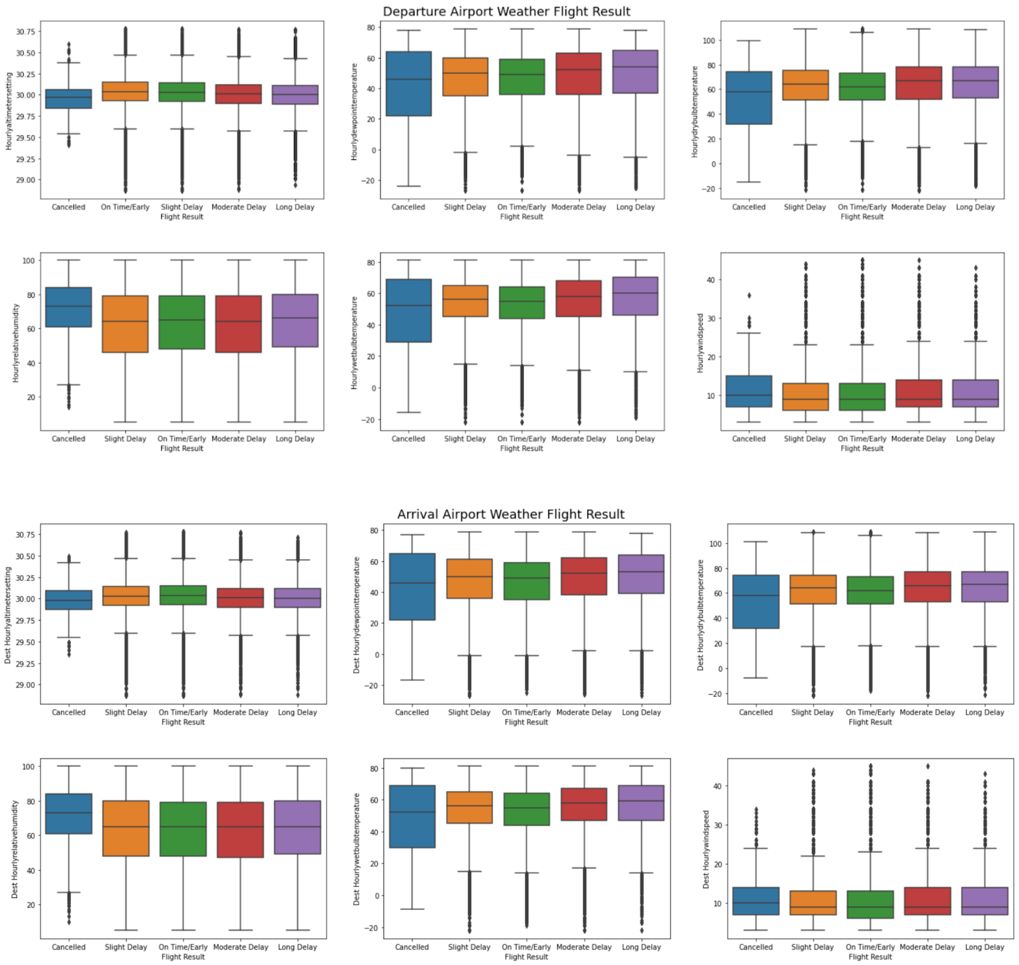


Figure 8: Flight Results vs Weather

Pareto analysis is a technique used to identify and prioritize the most important factors that contribute to a certain outcome. The Figure 9 shows that weather and carrier together contribute to 95% of the flight cancellations. This means that the majority of flight cancellations are caused by weather and carrier issues. Weather conditions can be unpredictable and can impact the safety of flights, resulting in cancellations or delays. Carrier issues may include equipment malfunctions, staffing shortages, or other operational problems that can affect the ability of airlines to operate their flights as scheduled.

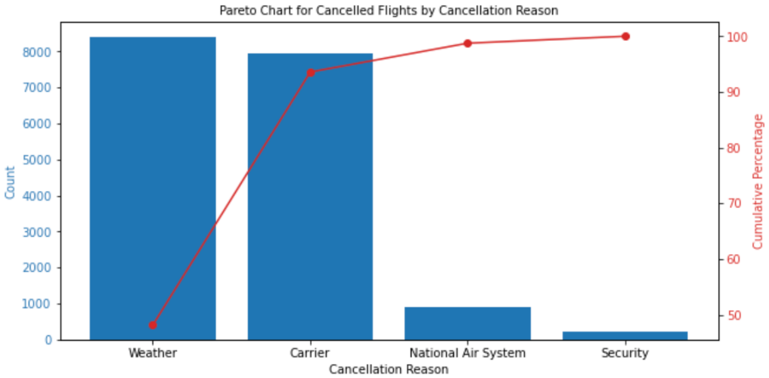
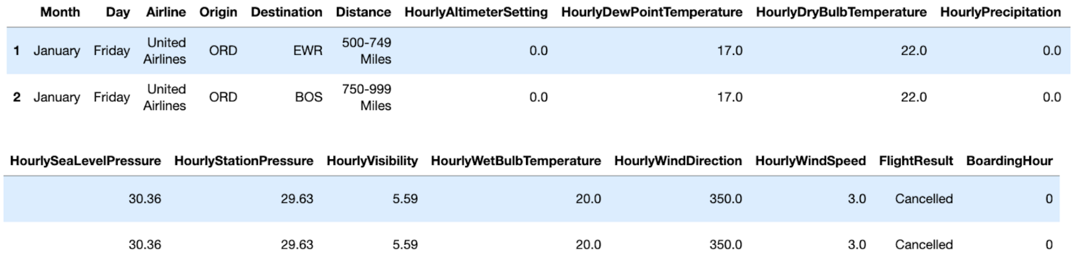


Figure 9: Flight Cancellation Pareto Chart

Cleaned and Final Dataset

After completing the data cleaning process, the final dataset was composed of the following columns: Month, Day, Airline, Origin, Distance, Hourly Altimeter Setting, Hourly Dew Point Temperature, Hourly Dry Bulb Temperature, Hourly Relative Humidity, Hourly Sea Level Pressure, Hourly Wet Bulb Temperature, Hourly Wind Direction, Hourly Wind Speed, Boarding Hour, and Flight Result. The Flight Result column represents the response variable while the other columns serve as feature variables. Figure 10 shows the first two rows of the final data.

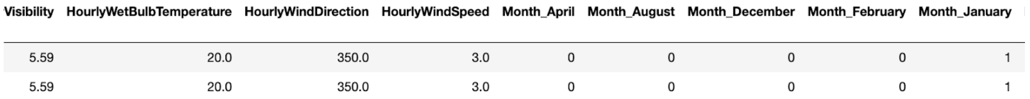


	Month	Day	Airline	Origin	Destination	Distance	HourlyAltimeterSetting	HourlyDewPointTemperature	HourlyDryBulbTemperature	HourlyPrecipitation
1	January	Friday	United Airlines	ORD	EWR	500-749 Miles	0.0	17.0	22.0	0.0
2	January	Friday	United Airlines	ORD	BOS	750-999 Miles	0.0	17.0	22.0	0.0

HourlySeaLevelPressure	HourlyStationPressure	HourlyVisibility	HourlyWetBulbTemperature	HourlyWindDirection	HourlyWindSpeed	FlightResult	BoardingHour
30.36	29.63	5.59	20.0	350.0	3.0	Cancelled	0
30.36	29.63	5.59	20.0	350.0	3.0	Cancelled	0

Figure 10: Cleaned Dataset

One Hot Encoding was used to transform categorical variables into a numerical format by creating dummy variables with binary values of 0 and 1. This allows the machine learning algorithms to better understand the relationship between categorical variables and the response variable, thus improving the accuracy of the models. Figure 11 shows the first two rows of the feature matrix.



Visibility	HourlyWetBulbTemperature	HourlyWindDirection	HourlyWindSpeed	Month_April	Month_August	Month_December	Month_February	Month_January
5.59	20.0	350.0	3.0	0	0	0	0	1
5.59	20.0	350.0	3.0	0	0	0	0	1

Figure 11: One Hot Encoding

Lastly, as shown in figure 11, while categorical variables have been successfully converted into 0 and 1 format, numerical variables still have varying ranges. In order to achieve better classification results, the feature matrix was standardized by scaling all columns to a range of 0 to 1.

Principal Component Analysis

Principal Component Analysis (PCA) is a widely used statistical technique used for reducing the dimensionality of large datasets while retaining as much information as possible. PCA works by finding the linear combinations of the original variables that explain the maximum amount of variance in the dataset. These new variables are known as principal components, and they can be used in place of the original variables to reduce the dimensionality of the dataset.

PCA can improve the classification model by reducing the dimensionality of the feature space. This can help to remove noise, redundancy, and other unwanted features from the data, making it easier for the classification algorithm to find the underlying patterns in the data. By reducing the dimensionality of the data, PCA can also help to overcome the curse of dimensionality, which refers to the difficulty in finding meaningful patterns in high-dimensional data.

The first step in PCA is to compute the covariance matrix of the dataset. This matrix contains the variances of each variable along the diagonal, and the covariances between each pair of variables in the off-diagonal elements. Once the covariance matrix is computed, the eigenvectors and eigenvalues of the matrix are calculated. The eigenvectors are the directions in which the data varies the most, and the eigenvalues represent the variance of the data along each eigenvector. The eigenvectors and eigenvalues are used to transform the original dataset into a new coordinate system that maximizes the variance in the data.

For this analysis, two principal components were used. This means that the two eigenvectors with the highest eigenvalues were retained. These two eigenvectors were used to transform the original dataset into a new two-dimensional coordinate system that maximizes the variance in the data.

By reducing the data's dimensionality to two, it is now possible to easily visualize the flight response against the feature variables. Figure 12 displays the flight results, plotted using the first and second principal components. It contains 100 data points randomly selected from the training dataset.

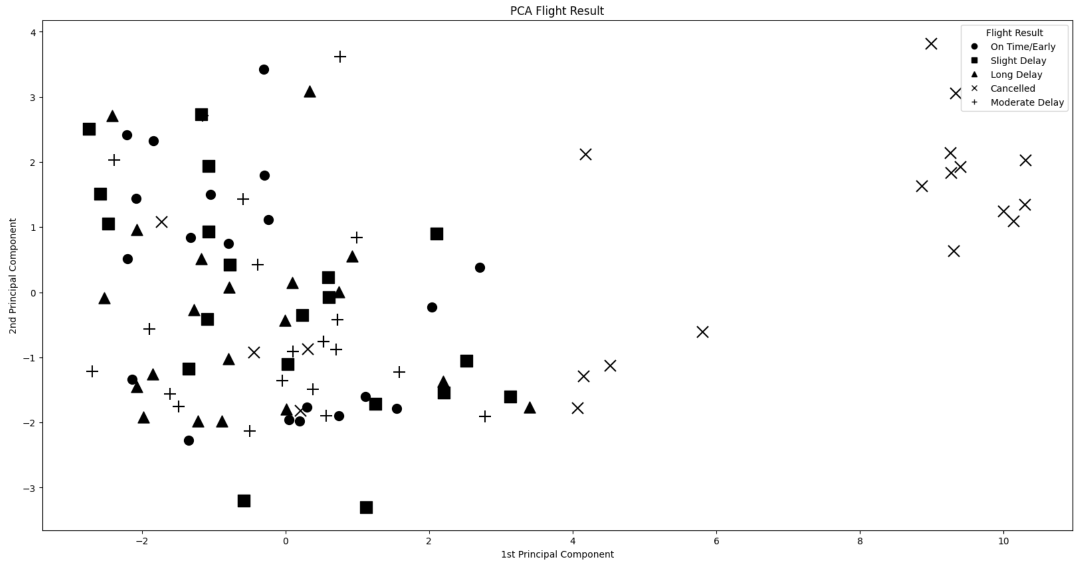


Figure 12: PCA

By looking at Figure 12, it can be noticed that ‘Cancelled’ flights are on the right side of the plot and the ‘On Time/’Early,’ ‘Slight Delay,’ ‘Long Delay,’ and ‘Moderate Delay’ are all scattered on the left side.

K-Nearest Neighbor

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that can be used for classification and regression tasks. In KNN, a new data point is assigned a label based on the class label of its nearest neighbors in the training set. The value of K, which represents the number of neighbors to consider, is a hyperparameter that must be determined before training the model.

To determine the optimal value of K, KNN was executed for different values of K, ranging from 0 to 40, and the accuracy of each model was computed. As shown in Figure 13, the optimal value of K is 39, indicating that considering the 39 nearest neighbors provides the best trade-off between model complexity and accuracy for this dataset.

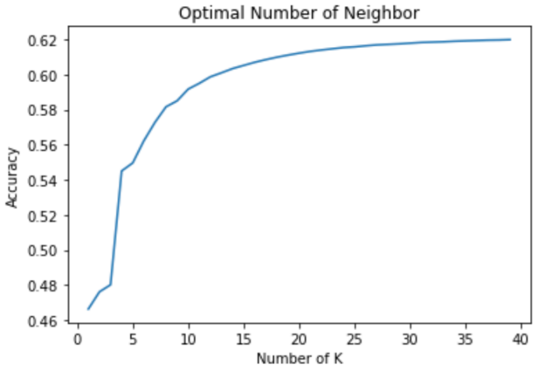


Figure 13: Optimal K

KNN algorithm with the optimal parameter (K=39) was utilized to predict flight results. Figure 14 illustrates the comparison between the training data and the predicted flight results. The black shapes represent the training data while the red shapes represent the predicted results. The plot contains 100 data points from each train and test dataset which were selected at random.

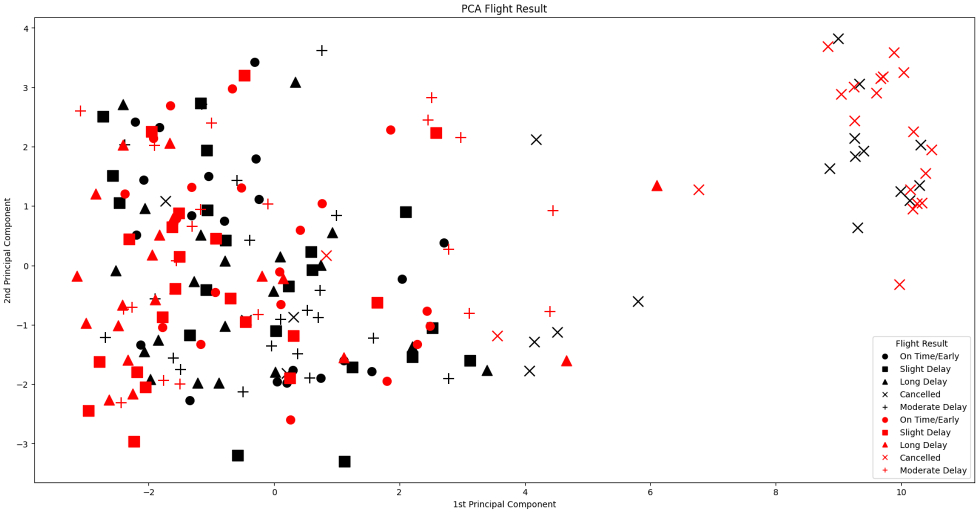


Figure 14: KNN Result

The model accuracy is 61.99%. Table 1 shows the other classification measures.

Flight Result	Precision	Recall	F1 Score	Support
Canceled	0.78	0.76	0.77	8726
Long Delay	1.00	0.00	0.00	10249
Moderate Delay	0.24	0.00	0.00	40747
On Time/Early	0.62	0.99	0.76	245998
Slight Delay	0.27	0.01	0.01	99600

Table 1: KNN Classification Measures

Table 1 shows the performance of the KNN flight classification model. Precision measures the proportion of correctly classified flights for a particular class, while Recall measures the proportion of correctly classified flights out of all the actual flights. F1 Score is the harmonic mean of Precision and Recall. Support is the number of actual flights that belong to each class. The model performed well for some classes, such as On Time/Early and Cancelled, but not so well for others, such as Long Delay and Moderate Delay.

Moreover, as can be noticed Figure 14, cancelled flights are on the right side and delayed flights are all scattered on the left side. To further improve the model, Flight results were re-categorized as “Arrived” or “Cancelled” only, combining “On Time/Early,” “Slight Delay,” “Moderate Delay,” and “Long Delay” as “Arrived.”

Using the same procedure as before, the optimal value of k was determined to be k=9.

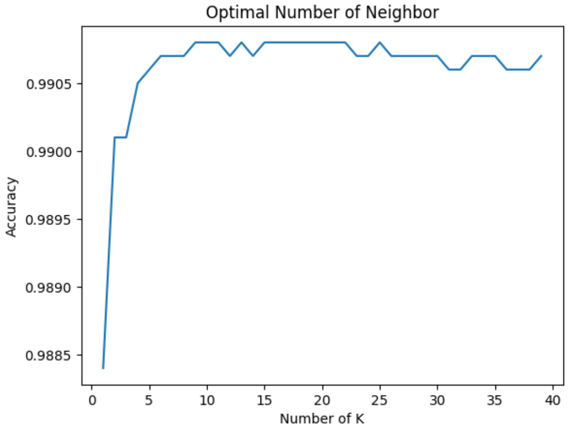


Figure 15: Optimal K

Figure 16 shows the train and predicted flight results. The black shapes represent the training data while the red shapes represent the predicted result. The plot contains 100 data points from each train and a predicted set selected at random.

It can be noticed on the plot, that there is a clear separation between “Arrived” and “Cancelled” flights.

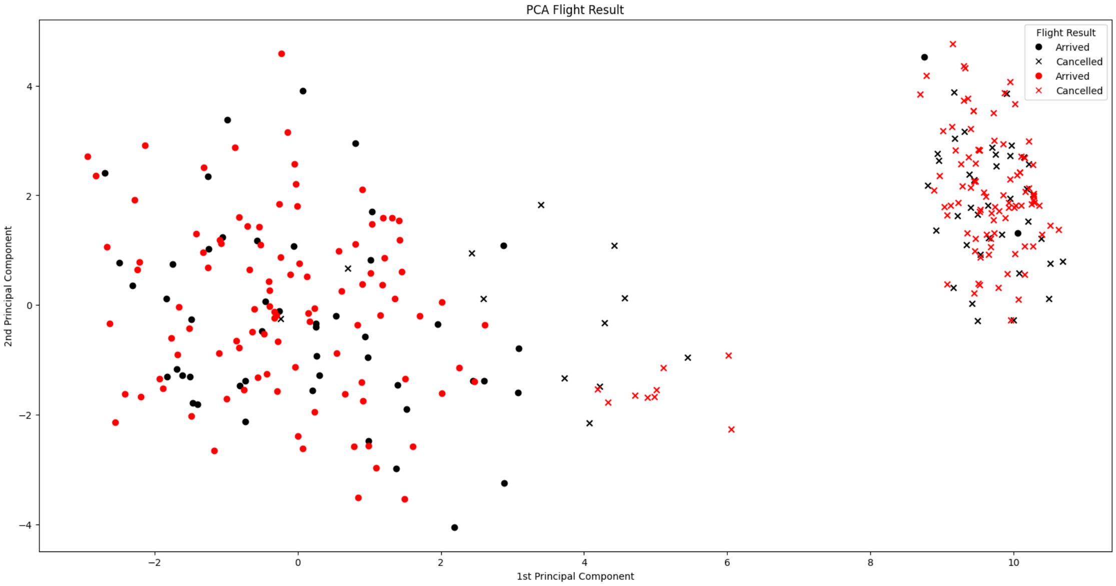


Figure 16: KNN Result

Github Repository: https://github.com/mebmiranda/U.S.-Flight-Delay-and-Cancellation-Prediction

The model accuracy is 99.08%. Table 2 shows the other classification measures.

Flight Result	Precision	Recall	F1 Score	Support
Arrived	0.99	1.00	1.00	396593
Canceled	0.84	0.70	0.77	8727

Table 2: KNN Classification Measures

In general, KNN is proficient in predicting "Arrived" and "Cancelled" flights, but not as effective in forecasting "Delay" flights. This could be attributed to the possibility that some relevant features that are crucial for predicting delays may be missing from the dataset.

Naive Model

Gaussian Naive Bayes is a popular probabilistic algorithm used for classification tasks. It is based on Bayes' theorem and assumes that the probability distribution of each feature is Gaussian. The algorithm works by calculating the conditional probability of a given input belonging to each class, and then choosing the class with the highest probability as the predicted class.

Using Gaussian Naive Bayes to predict flight results, the model accuracy is 62.13%. Table 3 shows the other classification measures.

Flight Result	Precision	Recall	F1 Score	Support
Canceled	0.77	0.77	0.77	8726
Long Delay	0.00	0.00	0.00	10249
Moderate Delay	0.00	0.00	0.00	40747
On Time/Early	0.62	1.00	0.76	245998
Slight Delay	0.00	0.00	0.00	99600

Table 3: Naive Bayes Classification Measures

Table 3 shows the performance of the Naive Bayes flight classification model. Precision measures the proportion of correctly classified flights for a particular class, while Recall measures the proportion of correctly classified flights out of all the actual flights. F1 Score is the harmonic mean of Precision and Recall. Support is the number of actual flights that belong to each class. The model performed well for some classes, such as On Time/Early and On Time/Early, but not so well for others, such as Long Delay and Moderate Delay.

To improve the model, similar to KNN, flight results were re-categorized as “Arrived” or “Cancelled” only, combining “On Time/Early,” “Slight Delay,” “Moderate Delay,” and “Long Delay” as “Arrived.” The new model accuracy is 99.00% and Table 4 shows the other classification measures.

Flight Result	Precision	Recall	F1 Score	Support
Arrived	0.99	1.00	0.99	396593
Canceled	0.78	0.75	0.76	8727

Table 4: Naive Bayes Classification Measures

In general, Naive Bayes is proficient in predicting "Arrived" and "Cancelled" flights, but not as effective in forecasting "Delay" flights. This could be attributed to the possibility that some relevant features that are crucial for predicting delays may be missing from the dataset.

Decision Tree and Random Forest

Decision Trees are fairly easy to implement, do not require much data preparation such as scaling and normalization, and are intuitive and can be easily explained to non-technical stakeholders. Random Forest is a combination of multiple decision trees that helps provide a more accurate result compared to a single tree, and avoids overfitting. This, however, comes at a cost, and usage of Random Forest would require more computing time and power.

In this study, the performance for both classification algorithms was compared, and 3 (Relative Humidity, Wind Direction, and Sea Level Temp) of the top 5 features for both methods overlap when both models are made to classify 5 different flight results: On-time/Early, Slight Delay, Moderate Delay, Long Delay, Canceled. For this scenario, Decision Tree even outperforms Random Forest.

Flight Result	Precision	Recall	F1 Score	Support
Canceled	0.81	0.73	0.77	8726
Long Delay	0.00	0.00	0.00	10249
Moderate Delay	0.00	0.00	0.00	40747
On Time/Early	0.62	1.00	0.76	245998
Slight Delay	0.50	0.00	0.00	99600

Table 5: Decision Tree Classification Measures

Flight Result	Precision	Recall	F1 Score	Support
Canceled	0.84	0.95	0.89	8726
Long Delay	0.18	0.08	0.11	10249
Moderate Delay	0.25	0.15	0.19	40747
On Time/Early	0.67	0.79	0.72	245998
Slight Delay	0.30	0.23	0.26	99600

Table 6: Random Forest Classification Measures

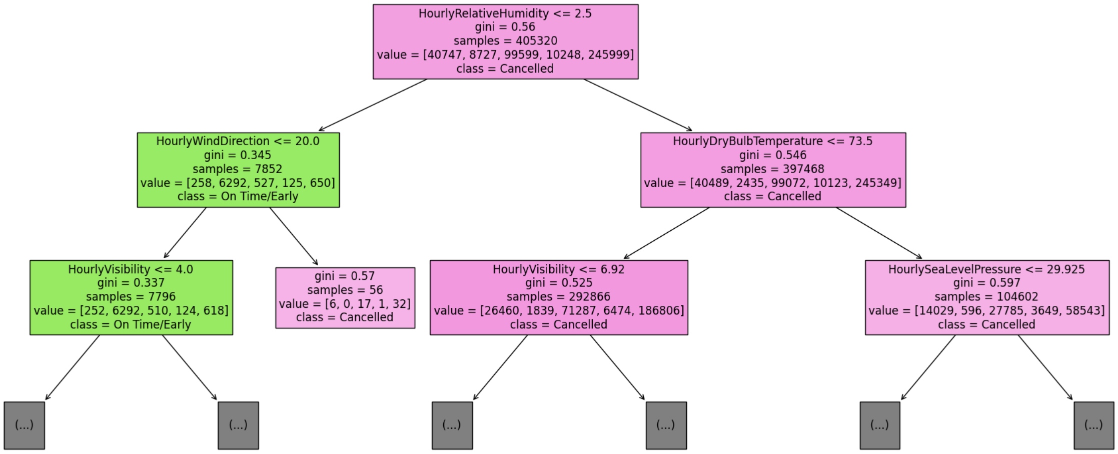


Figure 17. Decision Tree with max_depth = 3, min_samples_split = 5000

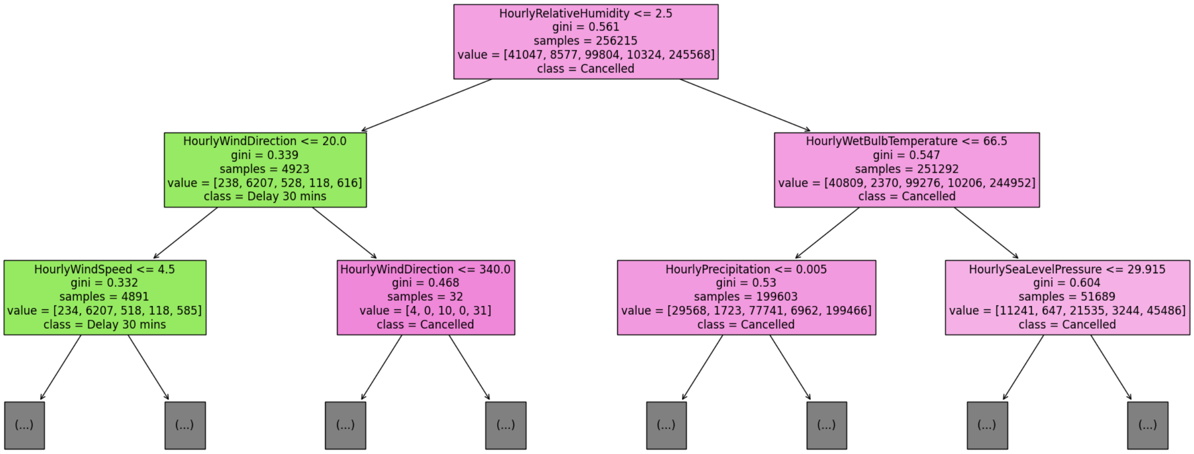

HourlyRelativeHumidity <= 2.5
gini = 0.561
samples = 256215
value = [41047, 8577, 99804, 10324, 245568]
class = Cancelled
HourlyWindDirection <= 20.0
gini = 0.339
samples = 4923
value = [238, 6207, 528, 118, 616]
class = Delay 30 mins
HourlyWetBulbTemperature <= 66.5
gini = 0.547
samples = 251292
value = [40809, 2370, 99276, 10206, 244952]
class = Cancelled
HourlyWindSpeed <= 4.5
gini = 0.332
samples = 4891
value = [234, 6207, 518, 118, 585]
class = Delay 30 mins
HourlyWindDirection <= 340.0
gini = 0.468
samples = 32
value = [4, 0, 10, 0, 31]
class = Cancelled
HourlyPrecipitation <= 0.005
gini = 0.53
samples = 199603
value = [29568, 1723, 77741, 6962, 199466]
class = Cancelled
HourlySeaLevelPressure <= 29.915
gini = 0.604
samples = 51689
value = [11241, 647, 21535, 3244, 45486]
class = Cancelled
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

Figure 18. Random Forest with n_estimators = 100

When the flight categorization was simplified and all non-Canceled flights (On-time/Early, Slight Delay, Moderate Delay, Long Delay) are combined into a single category named Arrived, while retaining the category Canceled, thereby reducing the classification problem to a binary classification, the accuracy increases to 99% for both models using the same the dataset. However, only 2 of top 5 features (Relative Humidity and Visibility) remain overlapping.

Flight Result	**Precision**	**Recall**	**F1 Score**	**Support**
Arrived	0.99	1.00	1.00	396594
Canceled	0.81	0.73	0.77	8726

Table 7: Decision Tree Classification Measures

Flight Result	Precision	Recall	F1 Score	Support
Arrived	1.00	1.00	1.00	396594
Canceled	0.84	0.95	0.89	8726

Table 8: Random Forest Classification Measures

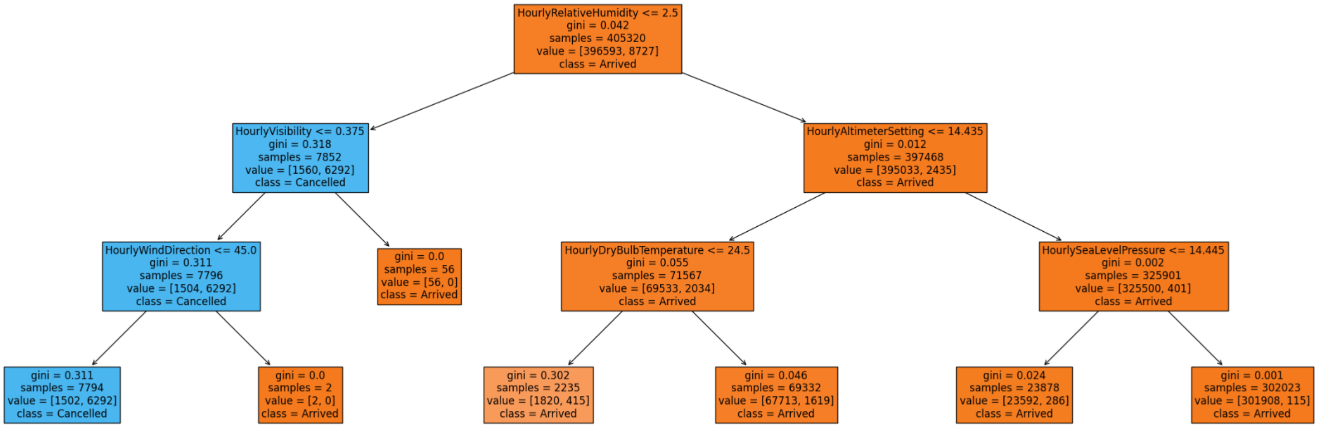


Figure 19. Decision Tree with max_depth = 3, min_samples_split = 5000

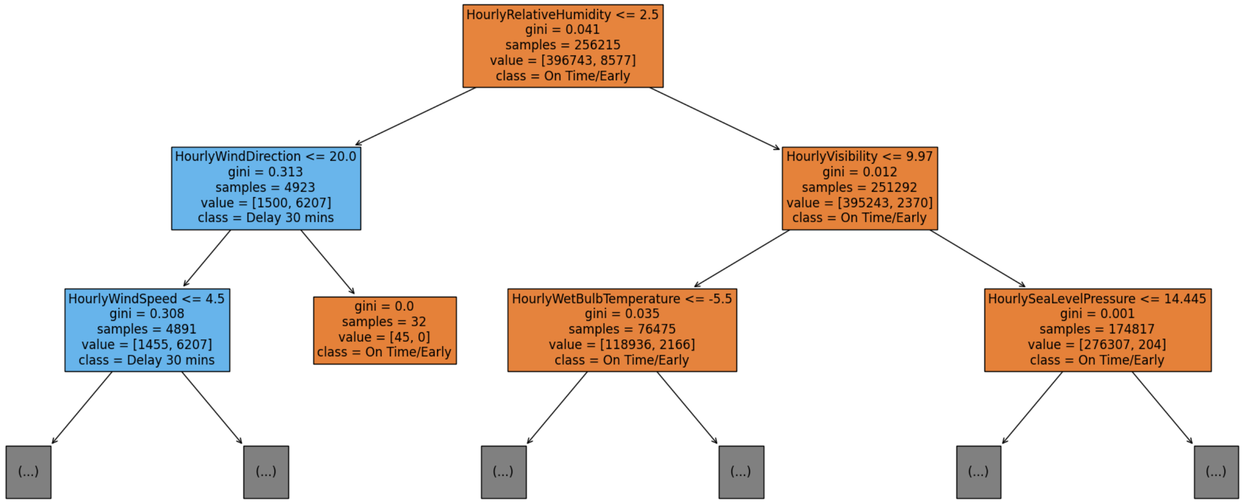


Figure 20. Random Forest with n_estimators = 100

Github Repository: https://github.com/mebmiranda/U.S.-Flight-Delay-and-Cancellation-Prediction

Evaluation and Final Results

Two scenarios were considered in the study. The first scenario (Table 7) categorizes flight results as On Time/Early, Slight Delay, Moderate Delay, Long Delay, or Cancelled. In the second scenario (Table 8), flight results were categorized as Arrived or Cancelled only.

Model	Accuracy	Features Used
KNN	61.99%	PCA ; components = 2
Naive Bayes	61.13%	PCA ; components = 2
Decision Tree	62.09%	Relative Humidity, Dry Bulb Temp, Visibility, Sea Level Temp, Wind Direction
Random Forest	57.07%	Relative Humidity, Wet Bulb Temp, Precipitation, Sea Level Temp, Wind Direction

Table 7: Classification Models Summary Scenario 1

Model	Accuracy	Features Used
KNN	99.08%	PCA ; components = 2
Naive Bayes	99.00%	PCA ; components = 2
Decision Tree	99.04%	Relative Humidity, Dry Bulb Temp, Visibility, Sea Level Temp, Wind Direction
Random Forest	99.50%	Relative Humidity, Wet Bulb Temp, Precipitation, Sea Level Temp, Wind Direction

Table 8: Classification Models Summary Scenario 2

While the four models performed poorly in scenario 1, the accuracy improved to 99% in scenario 2 for all models. Random Forest model appears to be the most effective, with an accuracy of 99.50%, utilizing features such as Relative Humidity, Wet Bulb Temperature, Precipitation, Sea Level Temperature, and Wind Direction.

Based on the results, weather measurements have high predictive capabilities in helping decide flight cancellations. Airlines can improve current operations by leveraging the departure airport's weather forecast during a flight's boarding time. If the expected weather conditions during boarding time lie within the parameters shown in this study, then airlines can announce flight cancellations to passengers earlier, and save them the time and effort of traveling to the airport.

For future studies, it may be necessary to consider additional relevant features that could further improve the model's ability to predict flight delays caused by factors not related to weather such as crew availability, airport congestion, airplane repairs.

Github Repository: https://github.com/mebmiranda/U.S.-Flight-Delay-and-Cancellation-Prediction