Predicting The Occurrence of Weather And Volume Related Ground Delay Programs

Eugene Mangortey*, Olivia J. Pinon†, Tejas G. Puranik‡, and Dimitri N. Mavris§
Georgia Institute of Technology, Atlanta, GA, 30332

Traffic Management Initiatives (TMI) such as Ground Delay Programs (GDP) are instituted by traffic management personnel to address and reduce the impacts of constraints in the National Airspace System. Ground Delay Programs are initiated whenever demand is projected to exceed an airport's acceptance rate over a lengthy period of time. Such instances occur when an airport is affected by conditions such as inclement weather, aircraft congestion, runway-related incidents, equipment failures, and other causes that do not fall in these categories. Over the years, efforts have been made to reduce the impact of Ground Delay Programs on airports and flight operations by predicting their occurrence. However, these efforts have largely focused on weather-related Ground Delay Programs, primarily due to a lack of access to comprehensive Ground Delay Program data. There has also been limited benchmarking of Machine Learning algorithms to predict the occurrence of Ground Delay Programs. Consequently, this research 1) fused data from the Traffic Flow Management System (TFMS), Aviation System Performance Metrics (ASPM), and Automated Surface Observing Systems (ASOS) datasets, and 2) leveraged supervised Machine Learning algorithms to develop prediction models as a means to predict the occurrence of weather and volume-related Ground Delay Programs. The Kappa Statistic evaluation metric revealed that Boosting Ensemble was the best suited algorithm for predicting the occurrence of weather and volume-related Ground Delay Programs.

I. Nomenclature

AAR	=	Airport Arrival Rates
$ASPM$	=	Aviation System Performance Metrics
$ASOS$	=	Automated Surface Observing Systems
$CASSIE$	=	Computing Analytics and Shared Services Integrated Environment
CSV	=	Comma-Separated Value
$EDCT$	=	Expected Departure Clearance Times
FAA	=	Federal Aviation Administration
$FIXM$	=	Flight Information Exchange Model
FN	=	False Negative
FP	=	False Positive
GDP	=	Ground Delay Program
NAS	=	National Airspace System
TAF	=	Terminal Aerodrome Forecast
$TFMS$	=	Traffic Flow Management System
TMI	=	Traffic Management Initiative
TN	=	True Negative
TP	=	True Positive

*Graduate Research Assistant, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Student Member
†Senior Research Engineer, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Senior Member
‡Research Engineer II, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Member
§Regents Professor for Advanced Systems Analysis, School of Aerospace Engineering, AIAA Fellow

II. Introduction and Motivation

The Federal Aviation Administration (FAA) is responsible for regulating, and maintaining the safety and efficiency of the National Airspace System (NAS) [1]. The National Airspace System is comprised of a vast network of radars, airports and landing areas, aeronautical charts, information, services, rules and regulations, procedures, technical information, and manpower [1]. Whenever conditions at an airport or an area in the airspace require action(s) to be taken to maintain safety or ease congestion in the airspace, traffic management personnel analyze demand on the NAS and assess if Traffic Management Initiatives (TMI) such as Ground Delay Programs (GDP) should be initiated [2].

A. Ground Delay Program (GDP)

A Ground Delay Program is initiated at an affected airport whenever aircraft demand is projected to exceed the airport's acceptance rate for a lengthy period of time [3]. An airport's acceptance rate may be lower than demand due to conditions affecting the airport, such as inclement weather, aircraft congestion, runway-related incidents, equipment failures etc. Figure 1 shows a breakdown of the different causes of Ground Delay Programs in 2017.

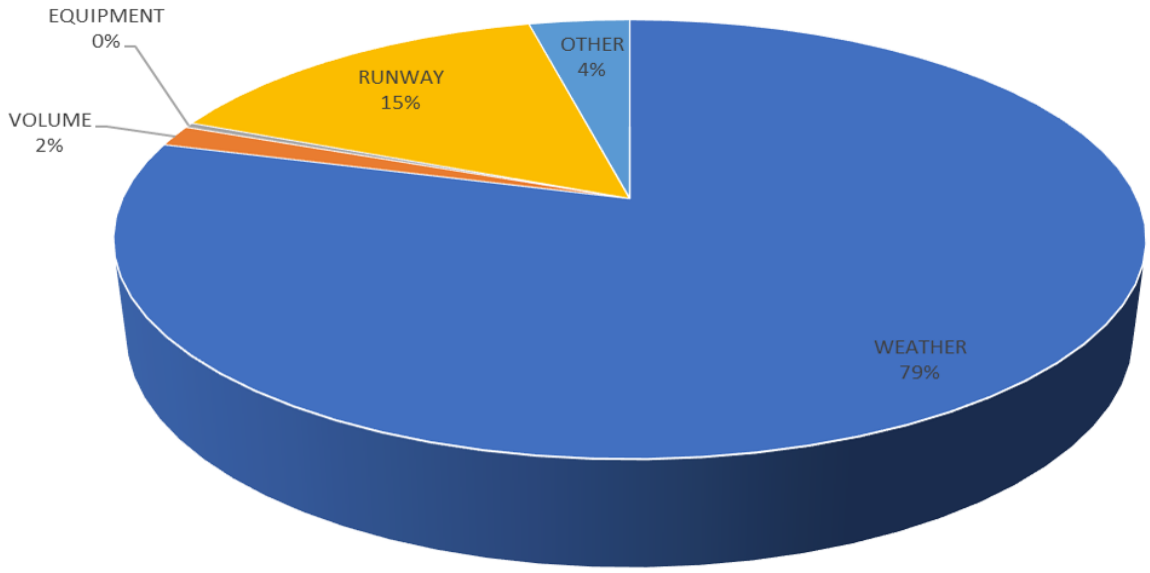

Fig. 1 Causes of Ground Delay Programs across the National Airspace System (NAS) in 2017 [4]

Whenever Ground Delay Programs are issued, traffic management personnel use the Enhanced Traffic Management System (ETMS) to predict, on national and local scales, traffic surges, gaps, and volume based on current and anticipated airborne aircraft [5]. This is done by evaluating the projected flow of traffic into airports and sectors, then implementing the least restrictive action necessary to ensure that traffic demand does not exceed system capacity. During Ground Delay Programs, ETMS issues Expected Departure Clearance Times (EDCT) to affected flights. EDCT is the runway release time ("Wheels Off") assigned to aircraft due to Traffic Management Initiatives (TMI) that require holding aircraft on the ground at the departure airport [6]. EDCT are updated whenever conditions improve to reduce delay durations.

B. Review of prior research related to Ground Delay Programs (GDP)

Smith et al. [7] used Terminal Aerodrome Forecast (TAF) weather data, Airport Arrival Rate (AAR) data from Aviation System Performance Metrics (ASPM), delay data from the Bureau of Transportation Statistics website and Support Vector Machines to predict Airport Arrival Rates (AAR), which were then used to predict weather-related Ground Delay Program program rates, duration and passenger delays. The authors pointed out that the limitations of Support Vector Machines such as its inability to predict rare occurrences impacted the performance of the prediction model. The performance of the prediction model can also be improved by benchmarking Machine Learning algorithms to identify the best suited algorithm for the prediction models.

Hansen et al. [8] outlined the Dynamic Stochastic Ground Holding (DSGH) algorithm, which was used to plan

and control a Ground Delay Program at the San Francisco International Airport under uncertainty in airport capacity. The algorithm revised departure delays, if necessary, by dynamically adapting to weather forecasts. Results obtained showed that the algorithm reduced expected delays by 7%. However, the scope of this research can be expanded by including other causes of Ground Delay Programs such as aircraft congestion in planning and controlling a Ground Delay Program as attempted by Hansen et al.

Mukherjee et al. [9] also predicted the occurrence of Ground Delay Programs based on weather conditions and traffic demand using the Logistic Regression and Decision Tree Machine Learning algorithms. Results showed that the Logistic Regression model performed better than the Decision Tree model in predicting the occurrence of Ground Delay Programs at the Newark and San Francisco International Airports. It is important to note that the prediction models were developed using actual weather conditions instead of weather forecasts. Forecasted weather conditions are used by traffic management personnel to implement Ground Delay Programs. This limitation can be addressed by using weather forecast data from a database such as the Automated Surface Observing Systems (ASOS).

Mangortey et al. [10] predicted the occurrence of weather-related Ground Delay Programs (GDP) at the Newark (EWR), La Guardia (LGA), and Boston Logan (BOS) International Airports using Decision Tree Machine Learning algorithm, which performed well. Expanding the scope of this research to include volume-related Ground Delay Programs, and benchmarking different techniques will provide much more information to aviation stakeholders.

The review of prior research highlights a few limitations and/or gaps. First, prior work has largely focused on weather-related Ground Delay Programs. Other important causes such as volume constraints have been largely ignored, primarily due to a lack of access to data. This research addresses this limitation by predicting the occurrence of Ground Delay Programs caused by inclement weather and volume constraints.

Second, a lack of benchmarking to evaluate and compare the performances of different Machine Learning algorithms in predicting Ground Delay Programs has led to the development of poorly performing prediction models [11, 12]. Consequently, this research focuses on benchmarking Machine Learning algorithms to identify a suitable algorithm for the prediction model.

C. Research Objective

The objective of this research is to benchmark Machine Learning algorithms, to identify the best suited algorithm for predicting the occurrence of weather and volume related Ground Delay Programs. In order to achieve this objective, there was a need to analyze Ground Delay Programs and their incidence across the largest airports in the United States. Figure 2 shows that the Newark (EWR), San Francisco (SFO), La Guardia (LGA), and Los Angeles (LAX) International Airports had the highest incidence of Ground Delay Programs in 2017. It can also be seen that Los Angeles International Airport had the best distribution of the different types of Ground Delay Programs compared to the other airports. Thus, the prediction model was developed for Los Angeles International Airport. However, it is worth noting that the methodology developed and used for this research can be re-implemented for other airports. The remainder of this paper highlights the datasets and methodology used for this research, and discusses the results obtained from this research.

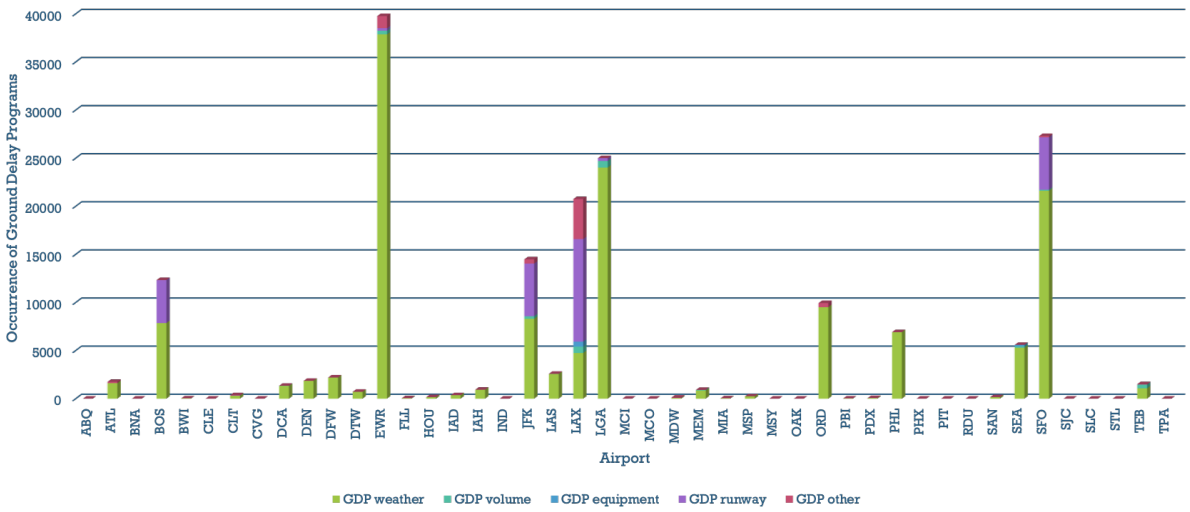

Fig. 2 Breakdown of Ground Delay Programs by airport (2017) [4]

III. Datasets

In order to achieve the research objective highlighted in the previous section, there was a need to identify and leverage datasets that contain information about Ground Delay Programs, airports, and forecasted weather conditions. The datasets used are:

- Traffic Flow Management System (TFMS)
- Aviation System Performance Metrics (ASPM)
- Automated Surface Observing System (ASOS)

A. Traffic Flow Management System (TFMS)

The Traffic Flow Management System (TFMS) is used by air traffic management personnel to plan and execute traffic flow management initiatives to ensure that constrained areas in the National Airspace System remain safe and operate optimally [13]. TFMS is comprised of two message streams: TFMS Flight and TFMS Flow. The TFMS Flight message stream provides initial flight plan messages, amended flight plan messages, departure and arrival time notifications, flight cancellation messages, boundary crossing messages, and track position reports. The TFMS Flow message stream on the other hand, provides data on traffic flow management initiatives such as Ground Stops, Reroutes, Airspace Flow Programs etc [13].

The TFMS datasets were obtained from the FAA's Computing Analytics and Shared Services Integrated Environment (CASSIE). This collaborative environment brings FAA divisions, partners, and stakeholders together in a shared services environment consisting of Big Data, computing power, and analytical tools. CASSIE utilizes the open-source software framework, Hadoop Hortonworks, for data storage and handling. In particular, the Hadoop Distributed File System (HDFS) allows for computer clusters to be linked robustly for high performance storage and computation [14]. Another component of Hadoop is NiFi, which automates the movement of data between disparate data sources and systems, making data ingestion fast, easy and secured [15].

B. Aviation System Performance Metrics (ASPM)

The Aviation System Performance Metrics database provides data from flights operating at 77 airports in the United States referred to as "ASPM airports" [16], flight data from 27 air carriers referred to as "ASPM carriers" [17], airport weather and runway data, and airport arrival and departure rates [18]. The ASPM data used for this research was obtained from the online ASPM database in csv format [19]. This database provides a comprehensive overview of air traffic for these airports and air carriers, and is composed of five modules [18]:

1) Metric module: This provides a comparison of actual flight departure and arrival times, and flight plan times, at

"ASPM airports" and between city pairs, actual and unimpeded taxi times for "ASPM airports", a comparison of actual flight departure and arrival times, and flight plan times for individual flights, and data regarding cancelled flights and completion rates
2) Efficiency module: This provides Terminal and System Airport Efficiency data for airports, and actual airport throughput (number of departures and arrivals) during a specified period of time
3) Enroute module: This provides average distance and time data for city pairs of 300 miles or more, and average distance and time data from all flights 300 miles or more from their arrival airport
4) Dashboard module: This provides limited next day airport information
5) Other module: This provides information on flight diversions, a summary of Traffic Management Initiatives and other aviation-related advisories, and the severity of weather factors with regards to their impact on flight delays at airports

C. Automated Surface Observing Systems (ASOS)

The Automated Surface Observing Systems (ASOS) dataset provides forecasted weather conditions which are widely used by meteorologists, climatologists, hydrologists, and aviation weather experts [20, 21]. This data provides a summary of airport weather conditions such as the date and time that the conditions were recorded as well as weather attributes such as ambient temperature, sea level pressure, visibility, wind speed, wind direction, wind gusts, dew point temperature, precipitation accumulation, cloud height and amount, etc. ASOS data used for this research was obtained online and in csv format [22].

IV. Methodology

In order to achieve the objective highlighted in Section II, the following served as a comprehensive methodology:

1) Step #1: Data processing
2) Step #2: Data fusion
3) Step #3: Model generation, validation, and testing
4) Step #4: Model evaluation

A. Step #1: Data Processing

In order to utilize the data required for this research, there was a need to not only understand the formats and contents of the raw datasets but to also parse them into useful formats, when needed. This section will cover steps taken to parse the Traffic Flow Management System (TFMS) datasets into a format suitable for analytical purposes.

1. Traffic Flow Management System (TFMS)

The Traffic Flow Mangement System (TFMS) datasets are stored in Flight Information Exchange Model (FIXM) [23] format, which is widely used for storing and transmitting aviation data. Consequently, there was a need to parse the TFMS datasets from FIXM format to csv format. These datasets are stored as hourly files comprised of all messages generated during that hour in the FAA's CASSIE environment, and have schema files which dictate the structure of the files. The TFMS files were parsed using their respective schema to ensure that all required fields were extracted in their correct format. This was done using a Python [24] parser developed by Mangortey et al. [10] which follows the process highlighted in Figure 3 and is described below:

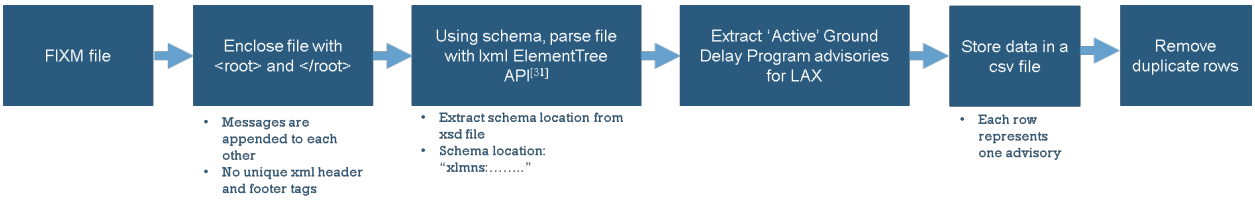

Fig. 3 XML/FIXM to JSON conversion process

1) Since the datasets are comprised of messages generated within the hour, there is no way to distinguish between the beginning of the file and the end of the file. Thus, it is important to enclose each file with a header and footer such as <root > and <\root > respectively to ensure that each file has unique starting and end points.
2) Extract the schema location from the xsd file. The schema location is typically of the format "xlmns:....."
3) Parse the FIXM file using the ElementTree [25] Application Program Interface (API)
4) Extract "Active" weather and volume-related Ground Delay Program messages for LAX
5) Store each Ground Delay Program message as a row in a csv file

After parsing the TFMS datasets, duplicate rows were removed, and the data was analyzed to ensure that the datasets were parsed correctly. Parameters extracted for "Active" Ground Delay Programs at the Los Angeles International Airport from January to August 2017 were the start and end dates and times of Ground Delay Programs, and their causes.

2. Aviation Systems Performance Metrics (ASPM)

Airport data for Los Angeles International Airport from January to August 2017 was extracted online in csv format from the Aviation System Performance Metrics database. The number of scheduled arrivals per hour and the actual number of arrivals per hour at LAX were extracted and used for this research.

3. Automated Surface Observing Systems (ASOS)

Automated Surface Observing Systems data was extracted online in csv format. The following parameters were extracted for Los Angeles International Airport from January to August 2017:

- Date and time
- Air Temperature (Fahrenheit)
- Dew Point Temperature (Fahrenheit)
- Relative Humidity (%)
- Wind Direction (Degrees)
- Wind Speed (Knots)
- Precipitation Accumulation (Inches)
- Pressure Altimeter (Inches)
- Sea level pressure (Millibars)
- Visibility (Miles)
- Wind Gusts (Knots)
- Cloud Coverage Type
- Cloud Altitude (Feet)

In order to ensure that the ASOS dataset was complete and appropriate for Machine Learning, the dataset was analyzed for missing values. The cloud coverage and altitude parameters particularly had a lot of missing values which meant that no clouds were present. These missing values were replaced with “M" representing missing values.

B. Step #2: Data Fusion

In order to develop a prediction model using Machine Learning algorithms, there was a need to fuse the datasets together. Data Fusion is a method of data analysis that involves fusing data from different sources to produce more consistent and useful information than that obtained from a single data source [26]. The datasets were fused by date and time, and the cause of the Ground Delay Programs served as the target of this model."Normal" was indicated as the cause whenever a Ground Delay Program did not occur. Predictors for this model were number of actual arrivals, number of scheduled arrivals, weather conditions, month, and hour.

C. Step #3: Model Generation, Validation and Testing

Lantz [27] defines Machine Learning as "the field of study interested in the development of computer algorithms to transform data into intelligent actions". Machine Learning has been widely used over the years. Examples of Machine Learning applications include forecasts of weather behavior and long-term climate changes [28], identification of fraudulent credit card transactions [29], prediction of popular election outcomes, [30] etc.

Machine Learning has been beneficial to many industries. However, it has its limitations as it has little flexibility to extrapolate outside of the strict parameters it learned. Thus, it is important for the model to be trained accurately and comprehensively to avoid over-fitting or under-fitting. Machine Learning algorithms and their applications also rely on various assumptions. It is thus critical for analysts to have a clear understanding of the assumptions associated with each Machine Learning algorithm.

Machine Learning algorithms are divided into three categories: supervised learning, unsupervised learning, and meta-learners/ensembles. Understanding the categories of Machine Learning algorithms is an integral step towards developing accurate prediction models. For the scope of this research, supervised learning algorithms were used to develop prediction models. Supervised learning is the process of training a Machine Learning model to predict value(s) using other values in the dataset. In particular, supervised learning algorithms attempt to discover and model the relationship between the value(s) being predicted and other values (predictors). These models are known as predictive models. Predictive models can be used to predict previous and real-time events, as well as future events [31]. Supervised learning algorithms that were benchmarked to identify the best suited algorithm for predicting the occurrence of Ground Delay Programs were Bagging Ensemble, Naive Bayes, Decision Trees, Boosting Ensemble, Classification Rule Learners, Random Forests, and Support Vector Machines [27].

In order to develop the prediction models, the fused datasets were partitioned into three sets: training, validation and testing. This process is known as the holdout method [32]. From Figure 4, it can be seen that half of the data was assigned to the training set, which is used to generate the model, one-fourth of the data was assigned to the validation set, which was used to iterate and refine the model, and one-fourth of the data was assigned to the test set, which was used to generate predictions for evaluations. The fused data was randomly divided between the three sets to ensure that the training, validation and test data do not have systematic differences. The performance of the test data alone should never be allowed to influence the performance of the model. Thus, it was important to include the validation set to ensure that a truly accurate estimate of future performances was obtained.

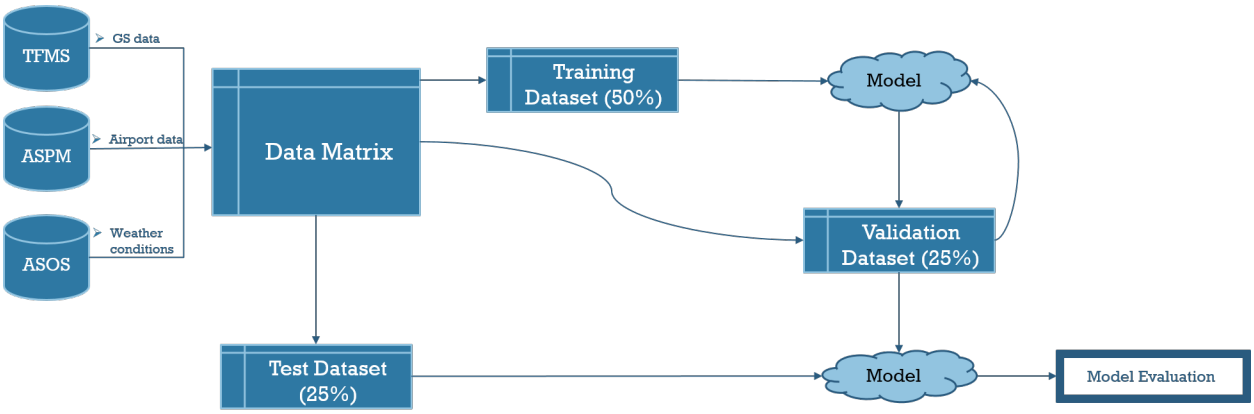

Fig. 4 Model Generation, Validation and Testing Process

D. Step #4: Evaluation

Evaluating the performance of prediction models is an important step as it informs as to how the model will perform on future data. The models were evaluated using results obtained from a confusion matrix. A confusion matrix as seen in Table 1 is a table that categorizes predictions according to whether they match the actual value. Performance metrics such as Accuracy, Kappa Statistic, Sensitivity, Specificity, Precision, and Recall were computed to assess model performance [27].

True Positive (TP) refers to the correct classification of the class of interest. True Negative (TN) refers to the correct classification of the class that is not of interest. False Positive (FP) refers to the incorrect classification of the class of interest. False Negative (FN) refers to the incorrect classification of the class that is not of interest [27].

Table 1 Confusion Matrix

	Actual: No	**Actual: Yes**
Predicted: No	True Negative (TN)	False Positive (FP)
Predicted: Yes	False Negative (FN)	True Positive (TP)

1. Accuracy

This refers to the ratio of the number of true positives and negatives, to the total number of predictions and is specified as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Sensitivity

This refers to the proportion of true positives that were correctly classified and is specified as:

$$Sensitivity = \frac{TP}{TP + FN}$$

3. Specificity

This refers to the proportion of negative examples that were correctly classified and is specified as:

$$Specificity = \frac{TN}{FP + TN}$$

4. Precision

This refers to the proportion of positive examples that were truly positive and is specified as:

$$Precision = \frac{TP}{FP + TP}$$

5. Recall

This refers to the ratio of true positives to the total number of positives and is specified as:

$$Recall = \frac{TP}{TP + FN}$$

6. Kappa Statistic

A model might have high accuracy because it correctly predicts the most frequent class, particularly when the dataset is unbalanced. Kappa Statistic adjusts accuracy by accounting for the probability of a correct prediction by chance alone, and is appropriate for unbalanced datasets. Kappa Statistic is specified below where P_0 is the observed value and P_E is the expected value. Table 2 provides an interpretation of Kappa Statistic values [27].

$$K = \frac{P_0 - P_E}{1 - P_E}$$

Table 2 Interpretation of Kappa Statistic values

Kappa Statistic	Interpretation
< 0	Poor Agreement
0 - 0.2	Slight Agreement
0.2 - 0.4	Fair Agreement
0.4 - 0.6	Moderate Agreement
0.6 - 0.8	Substantial Agreement
0.8 - 1	Almost Perfect Agreement

V. Results

As mentioned previously, the objective of this research is to predict the occurrence of weather and volume-related Ground Delay Programs. Seven Machine Learning algorithms were benchmarked to identify a suitable algorithm for the prediction model: Decision Trees, Naive Bayes, Classification Rule Learners, Support Vector Machines, Bagging Ensemble, Boosting Ensemble, and Random Forests. This section highlights the steps taken to develop and tune the models using these algorithms, and provides an analysis of their performance with the validation and testing sets using R. In order to ensure that the algorithms were assessed accurately, the data was randomly divided into three categories: training, validation, and testing sets. The training, validation, and testing sets had **2940, 981, and 980** data points respectively. It is worth noting that the data is heavily imbalanced since the number of no Ground Delay Program events greatly outnumbers the number of weather and volume related Ground Delay Program events. Since the data is heavily imbalanced, the kappa statistic metric was used to evaluate the performance of the algorithms as it accounts for the unbalanced nature of datasets. The remainder of this section will focus on summarizing how the different algorithms were used to train, validate, and test the models, and how they performed with the validation and testing tests.

A. Decision Trees

The model was trained using the "C50" function [27, 33] and the training dataset. The model's performance was improved using the validation dataset and adaptive boosting, "where multiple decision trees are built and the trees vote for the best class for each example" [27]. This involved adding a "trials" parameter when using the "C50" function. The optimal number of "trials" produced the lowest number of incorrect predictions. Finally, the model's performance was evaluated using the testing dataset and the "confusionMatrix" function [34]. Analysis of the Decision Tree algorithm revealed that the model had an average tree size of 72.9. Figure 5 shows that the month, altimeter pressure, dew point, sea level pressure, and visibility were the highest weighted predictors for this model, each contributing 6.254% as seen in Figure 5.

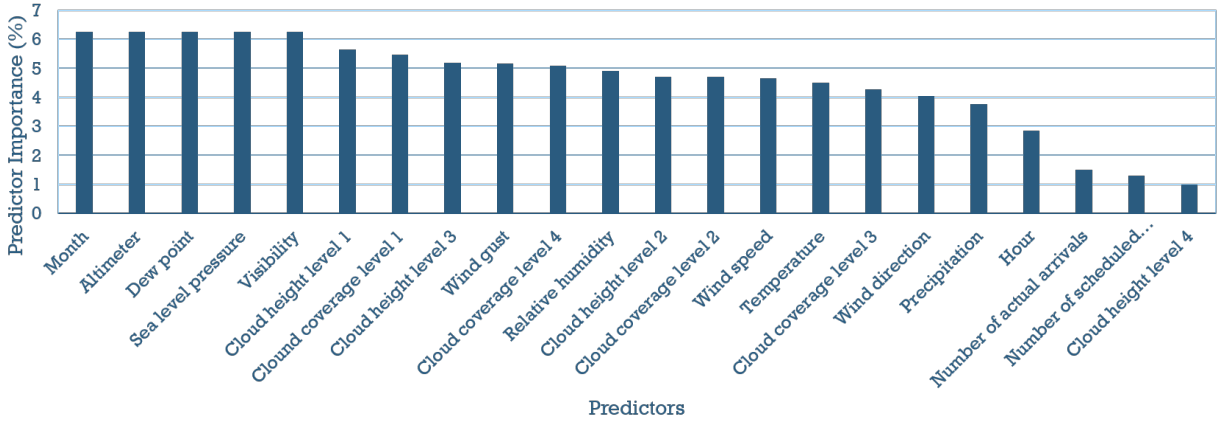

Fig. 5 Predictor importance for Decision Tree algorithm

1. Validation Dataset

Table 3 shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events, respectively. The model had an accuracy of 0.942, kappa statistic of 0.58, and a 95% Confidence Interval between 0.925 and 0.956, which is the range that the probability of a correct prediction lies within.

Table 3 Confusion matrix from Decision Tree algorithm using the validation dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	46	11	57
Predicted No GDP	41	883	924
Actual Total	87	894	981

Since the dataset is unbalanced, there was a need to further expand the evaluation of the model by analyzing how the model predicted volume-related Ground Delay Program events, weather-related Ground Delay Program events, and no Ground Delay Program events. Table 4 shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 4 Detailed confusion matrix from Decision Tree algorithm using the validation dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	11	1	1	13
Predicted Weather GDP	4	30	10	44
Predicted No GDP	11	30	883	924
Actual Total	26	61	894	981

From Table 4, it can be seen that the model accurately predicted 11 volume-related Ground Delay Program events, and incorrectly predicted 1 weather-related Ground Delay Program event and no Ground Delay Program event as volume-related Ground Delay Program events. The model also accurately predicted 30 weather-related Ground Delay Program events, and incorrectly predicted 4 volume-related Ground Delay Program events and 10 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 883 no Ground Delay Program events, and incorrectly predicted 11 volume-related Ground Delay Program events and 30 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 5 summarizes the detailed evaluation of the Decision Tree algorithm's performance with the validation dataset. Moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited when predicting volume and weather-related Ground Delay Programs. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 5 Detailed evaluation of the Decision Tree algorithm with the validation dataset

Metric	**Volume-related GDP**	**Weather-related GDP**	**No GDP**
Sensitivity	0.423	0.492	0.988
Specificity	0.997	0.985	0.529
Precision	0.846	0.681	0.956
Recall	0.985	0.967	0.807

2. Testing Dataset

Table 6 shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.922, kappa statistic of 0.531, and a 95% Confidence Interval between 0.903 and 0.938, which is the range that the probability of a correct prediction lies within.

Table 6 Confusion matrix from Decision Tree algorithm using the testing dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	51	18	69
Predicted No GDP	54	857	911
Actual Total	105	875	980

Table 7 shows the detailed confusion matrix for the testing dataset.

Table 7 Detailed confusion matrix from Decision Tree algorithm using the testing dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	9	2	2	13
Predicted Weather GDP	2	38	16	56
Predicted No GDP	22	32	857	911
Actual Total	33	72	875	980

From Table 7, it can be seen that the model accurately predicted 9 volume-related Ground Delay Program events, and incorrectly predicted 2 weather-related Ground Delay Program and 2 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 38 weather-related Ground Delay Program events, and incorrectly predicted 2 volume-related Ground Delay Program and 16 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 857 no Ground Delay Program events, and incorrectly predicted 22 volume-related Ground Delay Program and 32 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 8 summarizes the detailed evaluation of the Decision Tree algorithm's performance with the testing dataset. Low/moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Programs. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 8 Detailed evaluation of the Decision Tree algorithm with the testing dataset

Metric	**Volume-related GDP**	**Weather-related GDP**	**No GDP**
Sensitivity	0.272	0.530	0.979
Specificity	0.996	0.980	0.486
Precision	0.692	0.678	0.941
Recall	0.975	0.963	0.739

3. Summary

Overall, with kappa statistic values of 0.580 and 0.531 from the validation and testing datasets respectively, the Decision Tree algorithm had a moderate performance which can be attributed to the unbalanced nature of the dataset.

B. Naive Bayes

The model was trained using the "naiveBayes" function [27, 35] and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34].

1. Validation Dataset

Table 21 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.448, kappa statistic of 0.0709, and a 95% Confidence Interval between 0.416 and 0.479, which is the range that the probability of a correct prediction lies within. Table 22 in the appendix also shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 9 summarizes the detailed evaluation of the Naive Bayes algorithm's performance with the validation dataset. Low/moderate sensitivity and high specificity for volume-related Ground Delay Programs and no Ground Delay Program events show that the model's performance is limited in predicting volume-related Ground Delay Program and no Ground Delay Program events. However, the high sensitivity and moderate specificity show that the model predicted the majority of weather-related Ground Delay Program events.

Table 9 Detailed evaluation of the Naive Bayes algorithm with the validation dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.038	0.902	0.428
Specificity	0.963	0.465	0.828
Precision	0.028	0.101	0.962
Recall	0.974	0.986	0.124

2. Testing Dataset

Table 23 in the appendix shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.449, kappa statistic of 0.0708, and a 95% Confidence Interval between 0.418 and 0.481, which is the range that the probability of a correct prediction lies within. Table 24 in the appendix also shows the detailed confusion matrix for the testing dataset.

Table 10 summarizes the detailed evaluation of the Naive Bayes algorithm's performance with the testing dataset. Low.moderate sensitivity and high specificity for volume-related Ground Delay Program and no Ground Delay Program events show that the model's performance is limited in predicting volume-related Ground Delay Program and no Ground Delay Program events. However, high sensitivity and low specificity show that the model predicted the majority of weather-related Ground Delay Program events.

Table 10 Detailed evaluation of the Naive Bayes algorithm with testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.121	0.833	0.429
Specificity	0.959	0.476	0.762
Precision	0.093	0.112	0.938
Recall	0.969	0.973	0.138

3. Summary

Overall, with kappa statistic values of 0.0709 and 0.0708 from the validation and testing datasets respectively, the Naive Bayes algorithm performed poorly.

C. Classification Rule Learners

The model was trained using the "JRip" function [27, 36] and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34].

1. Validation Dataset

Table 25 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.918, kappa statistic of 0.5, and a 95% Confidence Interval between 0.899 and 0.934, which is the range that the probability of a correct prediction lies within. Table 26 in the appendix shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 11 summarizes the detailed evaluation of the Classification Rule Learners algorithm's performance with the validation dataset. Moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 11 Detailed evaluation of the Classification Rule Learners algorithm with the validation dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.615	0.459	0.959
Specificity	0.991	0.965	0.552
Precision	0.640	0.467	0.956
Recall	0.989	0.964	0.565

2. Testing Dataset

Table 27 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.895, kappa statistic of 0.444, and a 95% Confidence Interval between 0.874 and 0.913, which is the range that the probability of a correct prediction lies within. Table 28 in the appendix shows the detailed confusion matrix for the testing dataset.

Table 12 summarizes the detailed evaluation of the Classification Rule Learners algorithm's performance with the testing dataset. Moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 12 Detailed evaluation of the Classification Rule Learners algorithm with the testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.578	0.542	0.936
Specificity	0.989	0.947	0.571
Precision	0.655	0.448	0.948
Recall	0.985	0.963	0.517

3. Summary

Overall, with kappa statistic values of 0.5 and 0.444 with the validation and testing datasets respectively, the Classification Rule Learners algorithm had a moderate performance which can also be attributed to the unbalanced nature of the dataset.

D. Support Vector Machines

The model was trained using the "ksvm" function [27, 37], the "rbfdot" kernel (radial-based kernel), and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34].

1. Validation Dataset

Table 29 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.910, kappa statistic of 0.0173, and a 95% Confidence Interval between 0.891 and 0.927, which is the range that the probability of a correct prediction lies within. Table 30 in the appendix shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 13 summarizes the detailed evaluation of the Support Vector Machine algorithm's performance with the validation dataset. Extremely low sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, high sensitivity and extremely low specificity of no Ground Delay Program predictions shows that the model predicted the majority of no Ground Delay Program events accurately.

Table 13 Detailed evaluation of the Support Vector Machines algorithm with the validation dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0	0.016	0.998
Specificity	1	0.998	0.011
Precision	N/A	0.330	0.912
Recall	0.974	0.939	0.333

2. Testing Dataset

Table 31 in the appendix shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.897, kappa statistic of 0.081, and a 95% Confidence Interval between 0.876 and 0.915, which is the range that the probability of a correct prediction lies within. Table 32 in the appendix shows the detailed confusion matrix for the testing dataset.

Table 14 summarizes the detailed evaluation of the Support Vector Machines algorithm's performance with the testing dataset. Low sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, high sensitivity and low specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 14 Detailed evaluation of the Support Vector Machines algorithm with the testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0	0.069	0.999
Specificity	1	0.999	0.048
Precision	N/A	0.833	0.897
Recall	0.966	0.931	0.833

3. Summary

Overall, with kappa statistic values of 0.0173 and 0.0811 from the validation and testing datasets respectively, the Support Vector Machine algorithm performed poorly.

E. Bagging Ensemble

The model was trained using the "bagging" function [27, 38] and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34]. Analysis of the Bagging Ensemble algorithm revealed that altimeter pressure, dew point, and sea level pressure were the highest weighted predictors for this model as seen in Figure 6.

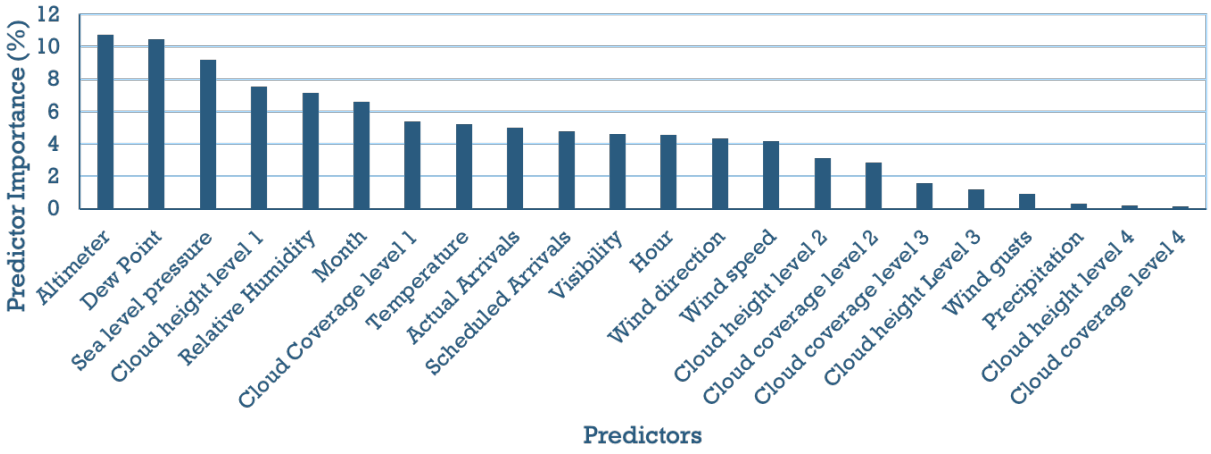

Fig. 6 Predictor importance for Bagging Ensemble algorithm

1. Validation Dataset

Table 33 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.937, kappa statistic of 0.474, and a 95% Confidence Interval between 0.919 and 0.951, which is the range that the probability of a correct prediction lies within. Table 34 in the appendix shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 15 summarizes the detailed evaluation of the Bagging Ensemble algorithm's performance with the validation dataset. Moderate/low sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and low specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 15 Detailed evaluation of the Bagging Ensemble algorithm with the validation dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.423	0.295	0.996
Specificity	1	0.992	0.368
Precision	1	0.720	0.942
Recall	0.985	0.955	0.889

2. Testing Dataset

Table 35 in the appendix shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.901, kappa statistic of 0.268, and a 95% Confidence Interval between 0.881 and 0.919, which is the range that the probability of a correct prediction lies within. Table 36 in the appendix shows the detailed confusion matrix for the testing dataset.

Table 16 summarizes the detailed evaluation of the Bagging Ensemble algorithm's performance with the testing dataset. Low sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and low specificity of no Ground Delay Program predictions show that the model predicted majority of no Ground Delay Program events accurately.

Table 16 Detailed evaluation of the Bagging Ensemble algorithm with the testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.303	0.139	0.986
Specificity	0.998	0.988	0.200
Precision	0.833	0.476	0.911
Recall	0.976	0.935	0.634

3. Summary

Overall, with kappa statistic values of 0.474 and 0.268 from the validation and testing datasets respectively, the Bagging Ensemble had a fair performance.

F. Boosting Ensemble

The model was trained using the "boosting" function [27, 39] and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34]. Analysis of the Boosting Ensemble algorithm revealed that month, dew point, altimeter pressure, and sea level pressure were the highest weighted predictors for this model, as seen in Figure 7.

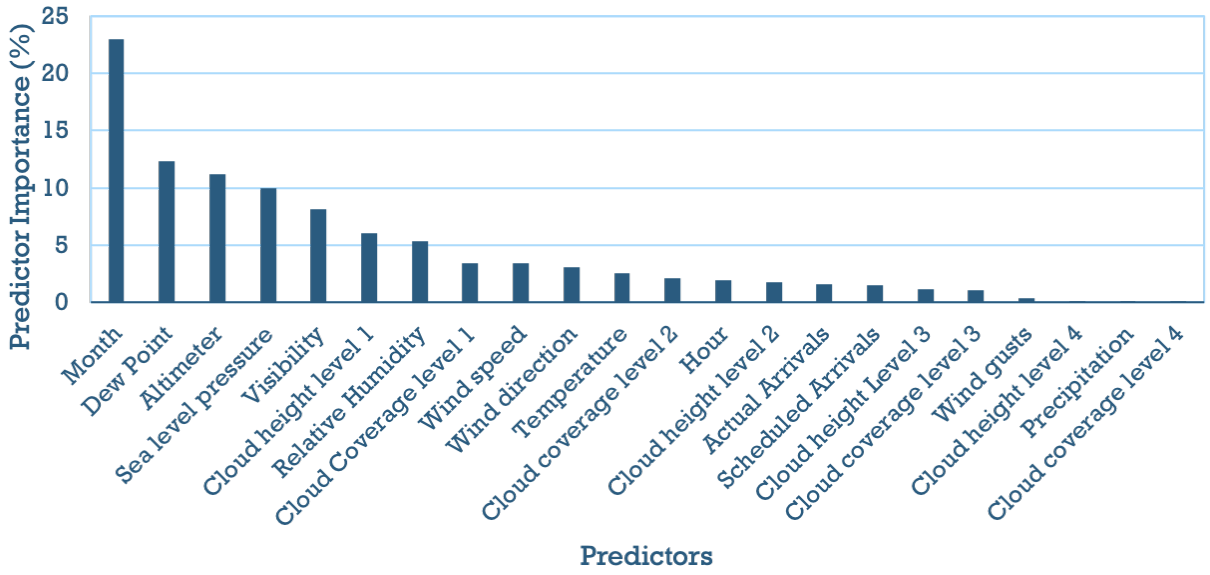

Fig. 7 Predictor importance for Boosting Ensemble algorithm

1. Validation Dataset

Table 37 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.948, kappa statistic of 0.629, and a 95% Confidence Interval between 0.932 and 0.961, which is the range that the probability of a correct prediction lies within. Table 38 in the appendix shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 17 summarizes the detailed evaluation of the Boosting Ensemble algorithm's performance with the validation dataset. Moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 17 Detailed evaluation of the Boosting Ensemble algorithm with the validation dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.538	0.508	0.989
Specificity	0.999	0.986	0.576
Precision	0.933	0.705	0.959
Recall	0.988	0.968	0.847

2. Testing Dataset

Table 39 in the appendix shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.943, kappa statistic of 0.657, and a 95% Confidence Interval between 0.926 and 0.957, which is the range that the probability of a correct prediction lies within. Table 40 shows the detailed confusion matrix for the testing dataset.

Table 18 summarizes the detailed evaluation of the Boosting Ensemble algorithm's performance with the testing dataset. Moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions

show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 18 Detailed evaluation of the Boosting Ensemble algorithm with the testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.485	0.583	0.989
Specificity	0.996	0.991	0.581
Precision	0.800	0.840	0.952
Recall	0.982	0.968	0.871

3. Summary

Overall, with kappa statistics values of 0.629 and 0.657 from the validation and testing datasets respectively, the Boosting Ensemble performed well.

G. Random Forests

The model was trained using the "randomForest" function [27, 40] and the training dataset. The model's performance was tuned and evaluated using the validation and testing datasets, respectively, and the "confusionMatrix" function [34]. The analysis of the Random Forests algorithm revealed that altimeter pressure, sea level pressure, the month, and dew point were the highest weighted predictors for this model, as seen in Figure 8.

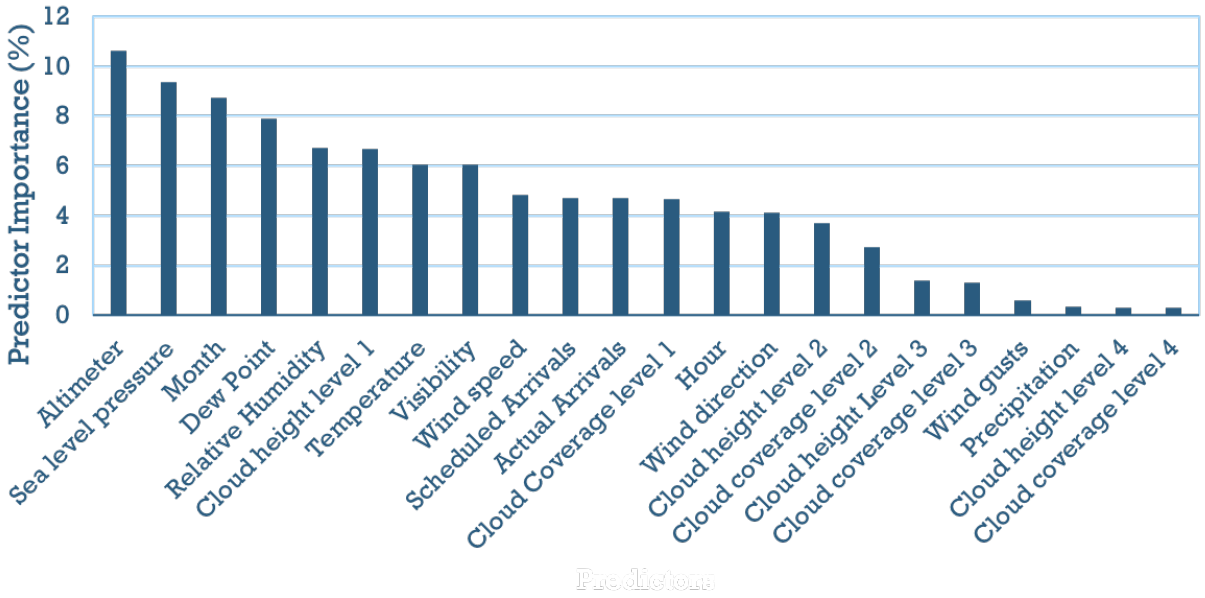

Fig. 8 Predictor importance for Random Forests algorithm

1. Validation Dataset

Table 41 in the appendix shows the confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.944, kappa statistic of 0.559, and a 95% Confidence Interval between 0.928 and 0.957, which is the range that the probability of a correct

prediction lies within. Table 42 in the appendix also shows the detailed confusion matrix for the validation dataset, where the last column and row represent the sum of predicted and actual events respectively.

Table 19 summarizes the detailed evaluation of the Random Forest Ensemble algorithm's performance with the validation dataset. Low/moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and low specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 19 Detailed evaluation of the Random Forest algorithm with the validation dataset for predicting the occurrence of Ground Delay Programs

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.385	0.424	0.996
Specificity	0.999	0.992	0.459
Precision	0.909	0.778	0.949
Recall	0.984	0.963	0.909

2. Testing Dataset

Table 43 in the appendix shows the confusion matrix for the testing dataset, where the last column and row represent the sum of predicted and actual events respectively. The model had an accuracy of 0.927, kappa statistic of 0.508, and a 95% Confidence Interval between 0.908 and 0.942, which is the range that the probability of a correct prediction lies within. Table 44 in the appendix also shows the detailed confusion matrix for the testing dataset.

Table 20 summarizes the detailed evaluation of the Random Forests algorithm's performance with the testing dataset. Low/moderate sensitivity and high specificity for volume and weather-related Ground Delay Program predictions show that the model's performance is limited in predicting volume and weather-related Ground Delay Program events. However, the high sensitivity and moderate specificity of no Ground Delay Program predictions show that the model predicted the majority of no Ground Delay Program events accurately.

Table 20 Detailed evaluation of the Random Forest algorithm with the testing dataset

Metric	Volume-related GDP	Weather-related GDP	No GDP
Sensitivity	0.212	0.458	0.992
Specificity	0.996	0.993	0.409
Precision	0.636	0.846	0.933
Recall	0.973	0.958	0.860

3. Summary

Overall, with kappa statistic values of 0.559 and 0.508 from the validation and testing datasets respectively, the Random Forests algorithm had a moderate performance.

H. Comparison of algorithms

Since the dataset is heavily imbalanced, accuracy is an inaccurate measure of the performance for these techniques. Kappa statistic, on the other hand, is appropriate for evaluating imbalanced datasets as it adjusts accuracy by accounting for the possibility of a correct prediction by chance alone [27]. The performance of the seven Machine Learning Techniques was thus compared using the Kappa statistic evaluation metric. Figure 9 shows that the **Boosting Ensemble**

had the highest kappa statistic value for both validation and testing datasets. Thus, it was identified as the best suited algorithm for predicting the occurrence of Ground Delay Programs.

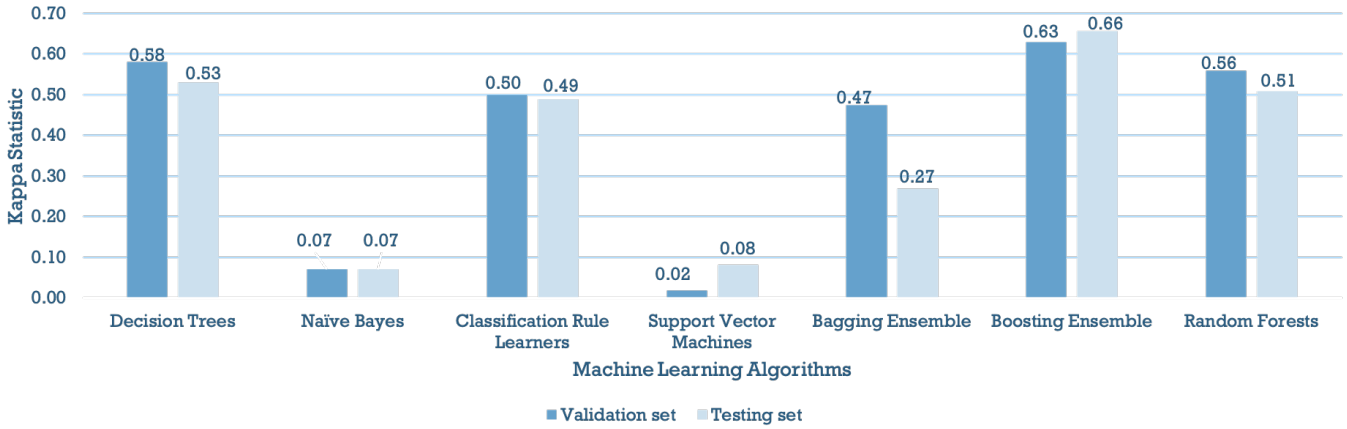

Fig. 9 Comparison of Machine Learning techniques for predicting the occurrence of Ground Delay Programs using Kappa Statistic

VI. Conclusion

Delays associated with the National Airspace System represent one of the two most common delays seen between June 2003 and July 2018. Whenever National Airspace System-related delays occur, Traffic Management Initiatives (TMI) such as Ground Delay Programs (GDP) may be issued at affected airports. Ground Delay Programs are implemented to control air traffic volume to specific airports where the projected traffic demand is expected to exceed the airport's acceptance rate over lengthy periods of time. Ground Delay Programs are caused by inclement weather, volume constraints, runway-related incidents, equipment failures etc. Over the years, efforts have been made to reduce the impact of Ground Delay Programs on airport and flight operations by predicting their occurrence. However, these efforts have largely focused on weather-related Ground Delay Programs, primarily due to a lack of access to comprehensive Ground Delay Program data. There has also been limited benchmarking of Machine Learning algorithms to predict the occurrence of Ground Delay Programs. Consequently, this research 1) fused data from the Traffic Flow Management System (TFMS), Aviation System Performance Metrics (ASPM), and Automated Surface Observing Systems (ASOS) datasets, and 2) leveraged supervised Machine Learning algorithms to develop prediction models as a means to predict the occurrence of weather and volume-related Ground Delay Programs at Los Angeles International Airport. The kappa statistic evaluation metric revealed that Boosting Ensemble was the best suited algorithm for predicting the occurrence of weather and volume-related Ground Delay Programs. Even though this methodology was applied to Los Angeles International Airport, it can be re-implemented at predict the occurrence of weather and volume related Ground Delay Program events at other airports.

Acknowledgments

The authors wish to acknowledge the support of FAA analysts and researchers for spearheading this research. Particularly, they wish to thank Tom Tessitore, Mike Paglione, Anya Berges, and David Chong for helping shape this research and providing valuable feedback. The views and findings expressed in this document are those of the authors only, and do not represent those of the FAA.

References

[1] Dillingham, G., "National Airspace System: FAA Has Implemented Some Free Flight Initiatives, but Challenges Remain," *General Accounting Office, Washington DC*, 1998.

[2] Federal Aviation Administration, "Traffic Flow Management in the National Airspace System," , 2009. URL `https://www.fly.faa.gov/Products/Training/Traffic_Management_for_Pilots/TFM_in_the_NAS_Booklet_ca10.pdf`.

[3] Ball, M., and Guglielmo, L., "Ground Delay Programs: Optimizing over the Included Flight Set Based on Distance," *14th AIAA Aviation Technology, Integration, and Operations Conference*, 2014.

[4] Federal Aviation Administration, "OPSNET : Delays : EDCT/GS/TMI By Cause Report," , 2019. URL https://aspm.faa.gov/opsnet/sys/opsnet-server-x.asp.

[5] Federal Aviation Administration, "Enhanced Traffic Management System," , 2017. URL https://www.fly.faa.gov/Products/Information/ETMS/etms.html.

[6] Federal Aviation Administration, "Expected Departure Clearance Times," , 2017. URL http://aspmhelp.faa.gov/index.php/Expect_Departure_Clearance_Times_(EDCT).

[7] Smith, D., and Lance, S., "Decision Support Tool for Predicting Aircraft Arrival Rates from Weather Forecasts," *2008 Integrated Communications, Navigation and Surveillance Conference*, 2008.

[8] Hansen, M., Mukherjee, A., and Grabbe, S., "Ground Delay Program Planning under Uncertainty in Airport Capacity," *Transportation Planning and Technology, vol. 35, no. 6*, 2012.

[9] Mukherjee, A., Grabbe, S., and Sridhar, B., "Predicting Ground Delay Program At An Airport Based On Meteorological Conditions," *Air Traffic Control Quarterly, vol. 12, no. 1*, 2004.

[10] Mangortey, Eugene and Gilleron, Jerome and Dard, Ghislain and Pinon, Olivia and Mavris, Dimitri, "Development of a Data Fusion Framework to support the Analysis of Aviation Big Data," *AIAA Science and Technology Forum (AIAA Scitech)*, 2019.

[11] Asencio, M. A., *Clustering Approach for Analysis of Convective Weather Impacting the NAS*, 12th Integrated Communications, Navigation, and Surveillance Conference, Herndon, Virginia, 2012.

[12] Liu P-c, B., *Scenario-Based Air Traffic Flow Management: From Theory to Practice*, From Theory to Practice, Transportation Research - Part B, Vol.42, 2008, pp. 685-702, 2010.

[13] Federal Aviation Administration, *JAVA MESSAGING SERVICE DESCRIPTION DOCUMENT Traffic Flow Management Data Service (TFMData) Vol. 2.0.5*, Federal Aviation Administration, 2016.

[14] Apache Hadoop, "Welcome to Apache™ Hadoop®!" , 2018. hadoop.apache.org/.

[15] The Apache Software Foundation, "Apache Nifi," , 2018. https://nifi.apache.org/.

[16] Federal Aviation Administration, "ASPM 77," , 2019. https://aspmhelp.faa.gov/index.php/ASPM_77.

[17] Federal Aviation Administration, "ASPM Carriers," , 2019. https://aspmhelp.faa.gov/index.php/ASPM_Carriers.

[18] Federal Aviation Administration, "Aviation System Performance Metrics (ASPM)," , 2019. https://aspmhelp.faa.gov/index.php/Aviation_System_Performance_Metrics_(ASPM).

[19] Federal Aviation Administration, "Aviation System Performance Metrics (ASPM)," , 2019. https://aspm.faa.gov/.

[20] National Weather Service, "Automated Surface Observing Systems," , 2019. https://www.weather.gov/asos/asostech.

[21] Guttman, Nathaniel and Baker, Bruce, "Exploratory Analysis of the Difference between Temperature Observations Recorded by ASOS and Conventional Methods," *Bulletin of American Meteorological Society*, 1996.

[22] Iowa State University, "ASOS-AWOS-METAR Data Download," , 2019. https://mesonet.agron.iastate.edu/request/download.phtml.

[23] Lepori, Hubert, "Introduction to FIXM." , 2017. URL https://www.icao.int/MID/Documents/2017/SWIMInterregional/8.2IntroductiontoFIXM.pdf.

[24] Python.org, "Welcome to Python.org." , 2018. URL www.python.org/.

[25] Python Software Foundation, "19.7. Xml.etree.ElementTree - The ElementTree XML API," , 2018. URL docs.python.org/2/library/xml.etree.elementtree.html.

[26] Klein, Lawrence A, *Sensor and data fusion: a tool for information assessment and decision making*, SPIE, 2012.

[27] Lantz, Brett, *Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*, Packt Publishing, 2015.

[28] Tripathi, Shivam and Srinivas, V. and Nanjundiahb, Ravi S., "Downscaling of precipitation for climate change scenarios: A support vector machine approach," *Journal of Hydrology Volume 330, Issues 3–4*, 2006.

[29] Priya Ravindra Shimpi, "Survey on Credit Card Fraud Detection Techniques," *International Journal Of Engineering And Computer Science*, 2016.

[30] Nausheen,Farha and Begum, Sayyada Hajera, "Sentiment Analysis to Predict Election Results Using Python," *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018.

[31] Maglogiannis, Ilias and Wallace, Manolis and Karpouzis, Kostas, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies - Volume 160 Frontiers in Artificial Intelligence and Applications*, I O S Press, Incorporated, 2007.

[32] Ron Kohavi, *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection* , International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[33] RDocumentation, "C5.0 Decision Trees and Rule-Based Models," , 2019. https://www.rdocumentation.org/packages/C50/versions/0.1.2.

[34] RDocumentation, "confusionMatrix," , 2019. https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix.

[35] RDocumentation, "Naive Bayes Classifier," , 2019. https://www.rdocumentation.org/packages/e1071/versions/1.7-0.1/topics/naiveBayes.

[36] RDocumentation, "R/Weka Rule Learners," , 2019. https://www.rdocumentation.org/packages/RWeka/versions/0.4-40/topics/Weka_classifier_rules.

[37] RDocumentation, "Support Vector Machines," , 2019. https://www.rdocumentation.org/packages/kernlab/versions/0.9-27/topics/ksvm.

[38] RDocumentation, "Bagging Classification And Regression Trees," , 2019. https://www.rdocumentation.org/packages/ipred/versions/0.4-0/topics/bagging.

[39] RDocumentation, "boostings," , 2019. https://www.rdocumentation.org/packages/adabag/versions/4.2/topics/boosting.

[40] RDocumentation, "Classification And Regression With Random Forest," , 2019. https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest.

VII. Appendix

A. Naive Bayes

Table 21 Confusion matrix from Naive Bayes algorithm using the validation dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	72	511	583
Predicted No GDP	15	383	398
Actual Total	87	894	981

Table 22 Detailed confusion matrix from Naive Bayes algorithm using the validation dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	1	0	35	36
Predicted Weather GDP	16	55	476	547
Predicted No GDP	9	6	383	398
Actual Total	26	61	894	981

From Table 22, it can be seen that the model accurately predicted 1 volume-related Ground Delay Program event, and incorrectly predicted 35 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 55 weather-related Ground Delay Program events, and incorrectly predicted 16 volume-related Ground Delay Program events and 476 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 383 no Ground Delay Program events, and inaccurately predicted 9 volume-related Ground Delay Program and 6 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 23 Confusion matrix from Naive Bayes algorithm using the testing dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	80	499	579
Predicted No GDP	25	376	401
Actual Total	105	875	980

Table 24 Detailed confusion matrix from Naive Bayes algorithm using the testing dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	4	0	39	43
Predicted Weather GDP	16	60	460	536
Predicted No GDP	13	12	376	401
Actual Total	33	72	875	980

From Table 24, it can be seen that the model accurately predicted 4 volume-related Ground Delay Program events, and incorrectly predicted 39 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 60 weather-related Ground Delay Program events, and incorrectly predicted 16 volume-related Ground Delay Program and 460 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 376 no Ground Delay Program events, and incorrectly predicted 13 volume-related Ground Delay Program and 12 weather-related Ground Delay Program events as no Ground Delay Program events.

B. Classification Rule Learners

Table 25 Confusion matrix from Classification Rule Learners algorithm using the validation dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	48	37	85
Predicted No GDP	39	857	896
Actual Total	87	894	981

Table 26 Detailed confusion matrix from Classification Rule Learners algorithm using the validation dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	16	0	9	25
Predicted Weather GDP	4	28	28	60
Predicted No GDP	6	33	857	896
Actual Total	26	61	894	981

From Table 26, it can be seen that the model accurately predicted 16 volume-related Ground Delay Program events, and incorrectly predicted 9 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 28 weather-related Ground Delay Program events, and incorrectly predicted 4 volume-related Ground Delay Program and 28 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 857 no Ground Delay Program events, and incorrectly predicted 6 volume-related Ground Delay Program and 33 weather-related Ground Delay Program eventss as no Ground Delay Program events.

Table 27 Confusion matrix from Classification Rule Learners algorithm using the testing dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	60	56	116
Predicted No GDP	45	819	864
Actual Total	105	875	980

Table 28 Detailed confusion matrix from Classification Rule Learners algorithm using the testing dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	19	0	10	29
Predicted Weather GDP	2	39	46	87
Predicted No GDP	12	33	819	864
Actual Total	33	72	875	980

From Table 28, it can be seen that the model accurately predicted 19 volume-related Ground Delay Program events, and incorrectly predicted 10 no Ground Delay Program events as volume-related Ground Delay Programs. The model also accurately predicted 39 weather-related Ground Delay Program events, and incorrectly predicted 2 volume-related Ground Delay Program and 46 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 819 no Ground Delay Program events, and incorrectly predicted 12 volume-related Ground Delay Program and 33 weather-related Ground Delay Program events as no Ground Delay Program events.

C. Support Vector Machines

Table 29 Confusion matrix from Support Vector Machines algorithm using the validation dataset

	Actual GDP	Actual No GDP	Predicted Total
Predicted GDP	1	2	3
Predicted No GDP	86	892	978
Actual Total	87	894	981

Table 30 Detailed confusion matrix from Support Vector Machines algorithm using the validation dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	0	0	0	0
Predicted Weather GDP	0	1	2	3
Predicted No GDP	26	60	892	978
Actual Total	26	61	894	981

From Table 30, it can be seen that the model accurately predicted 1 weather-related Ground Delay Program event, and incorrectly predicted 2 no Ground Delay Program events as weather-related Ground Delay Programs events. The model also accurately predicted 892 no Ground Delay Program events, and incorrectly predicted 26 volume-related Ground Delay Program and 60 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 31 Confusion matrix from Support Vector Machines algorithm using the testing dataset

	Actual GDP	Actual No GDP	Predicted Total
Predicted GDP	5	1	6
Predicted No GDP	100	874	974
Actual Total	105	875	980

Table 32 Detailed confusion matrix from Support Vector Machines algorithm using the testing dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	0	0	0	0
Predicted Weather GDP	0	5	1	6
Predicted No GDP	33	67	874	974
Actual Total	33	72	875	980

From Table 32, it can be seen that the model accurately predicted 5 weather-related Ground Delay Program events, and incorrectly predicted 1 no Ground Delay Program event as a weather-related Ground Delay Program. The model also accurately predicted 874 no Ground Delay Program events, and incorrectly predicted 33 volume-related Ground Delay Program and 67 weather-related Ground Delay Program events as no Ground Delay Program events.

D. Bagging Ensemble

Table 33 Confusion matrix from the Bagging Ensemble algorithm using the validation dataset

	Actual GDP	Actual No GDP	Predicted Total
Predicted GDP	32	4	36
Predicted No GDP	55	890	945
Actual Total	87	894	981

Table 34 Detailed confusion matrix from the Bagging Ensemble algorithm using the validation dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	11	0	0	11
Predicted Weather GDP	3	18	4	25
Predicted No GDP	12	43	890	945
Actual Total	26	61	894	981

From Table 34, it can be seen that the model accurately predicted 11 volume-related Ground Delay Program events. The model also accurately predicted 18 weather-related Ground Delay Program events, and inaccurately predicted 3 volume-related Ground Delay Program and 4 no Ground Delay Program events as weather-related Ground Delay Program events. Finally, the model accurately predicted 890 no Ground Delay Program events, and incorrectly predicted

26 volume-related Ground Delay Program and 61 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 35 Confusion matrix from the Bagging Ensemble algorithm using the testing dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	21	84	105
Predicted No GDP	12	863	875
Actual Total	33	947	980

Table 36 Detailed confusion matrix from the Bagging Ensemble algorithm using the testing dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	10	0	2	12
Predicted Weather GDP	1	10	10	21
Predicted No GDP	22	62	863	947
Actual Total	33	72	875	980

From Table 36, it can be seen that the model accurately predicted 10 volume-related Ground Delay Program events, and incorrectly predicted 2 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 10 weather-related Ground Delay Program events, and inaccurately predicted 1 volume-related Ground Delay Program and 10 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 863 no Ground Delay Program events, and incorrectly predicted 22 volume-related Ground Delay Program and 62 weather-related Ground Delay Program events as no Ground Delay Program events.

E. Boosting Ensemble

Table 37 Confusion matrix from the Boosting Ensemble algorithm using the validation dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	50	9	59
Predicted No GDP	37	885	922
Actual Total	87	894	981

Table 38 Detailed confusion matrix from the Boosting Ensemble algorithm using the validation dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	14	1	0	15
Predicted Weather GDP	4	31	9	44
Predicted No GDP	8	29	885	924
Actual Total	26	61	894	981

From Table 38, it can be seen that the model accurately predicted 14 volume-related Ground Delay Program events, and incorrectly predicted 1 weather-related Ground Delay Program as a volume-related Ground Delay Program event. The model also accurately predicted 31 weather-related Ground Delay Program events, and inaccurately predicted 4 volume-related Ground Delay Program and 9 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 885 no Ground Delay Program events, and incorrectly predicted 8 volume-related Ground Delay Program and 29 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 39 Confusion matrix from the Boosting Ensemble algorithm using the testing dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	61	9	70
Predicted No GDP	44	866	910
Actual Total	105	875	980

Table 40 Detailed confusion matrix from the Boosting Ensemble algorithm using the testing dataset

	Actual Volume GDP	**Actual Weather GDP**	**Actual No GDP**	**Predicted Total**
Predicted Volume GDP	16	2	2	20
Predicted Weather GDP	1	42	7	50
Predicted no GDP	16	28	866	910
Actual Total	33	72	875	980

From Table 40, it can be seen that the model accurately predicted 16 volume-related Ground Delay Program events, and incorrectly predicted 2 weather-related Ground Delay Program and 2 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 42 weather-related Ground Delay Program events, and inaccurately predicted 1 volume-related Ground Delay Program and 7 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 866 no Ground Delay Program events, and incorrectly predicted 16 volume-related Ground Delay Program and 28 weather-related Ground Delay Program events as no Ground Delay Program events.

F. Random Forests

Table 41 Confusion matrix from the Random Forest algorithm using the validation dataset

	Actual GDP	**Actual No GDP**	**Predicted Total**
Predicted GDP	40	4	44
Predicted No GDP	47	890	937
Actual Total	87	894	981

Table 42 Detailed confusion matrix from the Random Forest algorithm using the validation dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	10	1	0	11
Predicted Weather GDP	3	26	4	33
Predicted No GDP	13	34	890	937
Actual Total	26	61	894	981

From Table 42, it can be seen that the model accurately predicted 10 volume-related Ground Delay Program events, and incorrectly predicted 1 weather-related Ground Delay Program event as a volume-related Ground Delay Program event. The model also accurately predicted 26 weather-related Ground Delay Program events, and inaccurately predicted 3 volume-related Ground Delay Program and 4 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 890 no Ground Delay Program events, and incorrectly predicted 13 volume-related Ground Delay Program and 34 weather-related Ground Delay Program events as no Ground Delay Program events.

Table 43 Confusion matrix from the Random Forest algorithm using the testing dataset

	Actual GDP	Actual No GDP	Predicted Total
Predicted GDP	43	7	50
Predicted No GDP	62	868	930
Actual Total	105	875	980

Table 44 Detailed confusion matrix from the Random Forest algorithm using the testing dataset

	Actual Volume GDP	Actual Weather GDP	Actual No GDP	Predicted Total
Predicted Volume GDP	7	2	2	11
Predicted Weather GDP	1	33	5	39
Predicted No GDP	25	37	868	930
Actual Total	33	72	875	980

From Table 44, it can be seen that the model accurately predicted 7 volume-related Ground Delay Program events, and incorrectly predicted 2 weather-related Ground Delay Program and 2 no Ground Delay Program events as volume-related Ground Delay Program events. The model also accurately predicted 33 weather-related Ground Delay Program events, and inaccurately predicted 1 volume-related Ground Delay Program and 5 no Ground Delay Program events as weather-related Ground Delayed Program events. Finally, the model accurately predicted 868 no Ground Delay Program events, and incorrectly predicted 25 volume-related Ground Delay Program and 37 weather-related Ground Delay Program events as no Ground Delay Program events.