

# Flight Delay Predictions Using Machine Learning Classification by LightGBM Model

**Abstract**—Flight delays are a crucial issue in aviation that impact passenger fulfillment, airline operations, and airport efficiency. This study utilizes the dataset 'US 2023 Civil Flights, Delays, Meteo and Aircrafts' which contains over 6 million civil flight data in the United States throughout the year 2023. This dataset contains information on departure and arrival delays, meteorological variables (wind speed, rainfall, air pressure), and aircraft attributes (age, type, airline). The aim of this study is to identify the key factors affecting flight delays using a data-driven approach. The methodology employed includes stages of data preprocessing, visual exploration, and the application of predictive models based on machine learning, such as Random Forest, Gradient Boosting, and Lightgbm to measure the contribution of each variable. Evaluation is conducted using accuracy, precision, recall, and F1-score metrics. The results indicate that weather factors and aircraft age have a significant relationship with delays, which can be used for strategic decision-making by airport operators and airlines.

**Keywords**—Flight Delays, Machine Learning, LightGBM, Prediction.

## I. INTRODUCTION

Flight delays in the United States have become a remarkable issue. According to the Bureau of Transportation Insights (BTS), it's reported that the annual total cost of flight delays was over \$30 Billion, which makes a significant challenge to the airlines [1]. In addition, the delays are caused by limited airspace, some of them are late arriving aircraft. This can cause subsequent flights to be delayed. Since aircraft follow scheduled routes, an early delay can also significantly impact later flight [1]. As an example, if an aircraft departs one hour late, it will likely arrive late at the next airport, this can cause further delays in some departures, the worst scenario is chaining reaction of delay propagation across multiple flights. Delay propagation is innate with the National Airspace System (NAS), involving interconnected resources, such as aircraft, crew, passengers, and gates. Increasing air traffic demand forces to reduce the buffer times, making NAS more vulnerable to delays. Hence, modeling delay propagation is needed for an accurate prediction.[1]

Weather is one of the factors in both extreme-weather and aviation-system delays, as well as in delays caused by late-arriving aircraft. Some airlines often don't specify the exact weather-related cause in category [2]. BTS uses data from airlines to decide the rate of flight delays caused by weather, including delays caused under extreme-weather, late-arriving, and National Aviation System delays. In general, flight delays are attributed to multiple interconnections such as operational inefficiencies, technical issues, and air traffic congestion [2]. Analyzing impacts of each factor in flight delays is challenging due to many variations such as regions, time, and operational contexts. However, with the increasing availability of open-access data, there's a growing opportunity to use data-driven methods to understand the delay patterns. Such as the "US 2023 Civil Flights, Delays, Meteo and Aircrafts" dataset, it provides detailed information on over 6 million domestic flights in the U.S during 2023, including flight schedules, weather conditions, and aircraft specifications.

This study aims at how weather and aircraft specifications related to flight delays using machine learning. Following a structured methodology that starts from data cleaning and integration to exploring analysis and predictive modeling such as Random Forest, Gradient Boosting, Lightgbm. The performance is assessed through accuracy, precision, recall, and F1-Score. The results are expected to offer valuable predictions to give insights that support the airline's decision making for better delay mitigation.

## I. RELATED WORKS

Research on flight delay prediction has gained significant attention in recent years, both from academic institutions and industry practitioners. These studies generally utilize a combination of historical flight data and weather data, applying machine learning (ML) technique.

### A. Hybrid Optimization with Random Forest

The COA-Optimized Weighted Random Forest (COWRF) model, which combines the DBSCAN clustering algorithm with the Coyote Optimization Algorithm (COA) to enhance the performance of Random Forest in predicting flight delays. The model was tested on a large dataset from JFK Airport and achieved an accuracy of 97.2%, significantly outperforming standard Random Forest models [3][4].

### B. The Impact of Extreme Weather on Flight Delay

After analyzing over 2 million U.S. flights using a difference-in-difference econometric approach. Some studies found that extreme weather events such as heavy rain and strong winds can increase average delay times by up to 23 minutes [5]. This highlights the importance of incorporating real-time weather data in predictive models.

Complementary findings have been reported in various regions globally. For example, Skultety et al. [6] studied en-route flight delays in European airspace and highlighted that dangerous weather phenomena, particularly thunderstorms, significantly affect flight punctuality, especially during summer months. Similarly, Rodríguez-Sanz et al. [7] demonstrated that adverse weather conditions such as fog, crosswinds, and snowstorms at European airports reduce runway throughput and increase arrival delays. These conditions constrain both landing and take-off operations, leading to operational bottlenecks.

In a more recent study, Hsu et al. [8] utilized location-based mobile tracking data to analyze the impact of Winter Storm Elliott in the United States, showing how passenger delay times and airport congestion can dramatically rise during extreme weather events. Moreover, machine learning models developed by Dalmau et al. [9] show how adverse weather affects airport peak service rates by degrading operational efficiency in terms of aircraft movements per hour.

Taken together, these studies highlight the critical importance of incorporating real-time, high-resolution weather data into predictive delay models. Such integration not only improves forecast accuracy but also aids

stakeholders-airlines, airport operators, and passengers in making better-informed operational decisions.

### C. Long-Term Delay Prediction Using LSTM

With the utilization in 10 years of historical and weather data from three major airports (ICN, JFK, MDW). Several ML algorithms were compared such as: Random Forest, Lightgbm, and SVM with a deep learning model (LSTM) for predicting delays up to 48 hours (about 4 days) in advance. The LSTM model achieved the highest accuracy, reaching 85.2% at JFK Airport [10].

### D. Open Literature Review on Machine Learning for Delay Prediction

There are some comprehensive reviews of various delay prediction approaches, including statistical methods, classification, regression, and machine learning. Some study also introduced a taxonomy of flight delay research, covering key aspects such as feature selection, preprocessing, and evaluation methods, which can serve as a foundation for developing robust predictive pipelines [11].

### E. Open Literature Review on Machine Learning for Delay Prediction

From the four studies above, we can conclude the following:

- Model optimization (e.g., COWRF) significantly improves prediction accuracy.
- Weather variables are essential predictors and must be integrated as key features.
- Deep learning models like LSTM are highly effective for long-term predictions.
- A well-structured end-to-end pipeline should include preprocessing, feature engineering, modeling, and evaluation.

Accordingly, this project adopts a hybrid ML approach, incorporating meteorological and time-based features, while implementing explainable AI techniques to enhance the transparency of the prediction process.

## III. METHODOLOGY

Biarin w

### A. Datasets Introduction

This research utilizes three main datasets obtained from the Kaggle repository [1], sdapecifically designed to analyze factors influencing flight delays in the United States throughout 2023. Each dataset contributes unique information including flight schedules, weather conditions, and airport geolocation as in Table 1.

TABLE I. DATASETS DIMENTIONALLY SIZE AND DESCRIPTION

Dataset Name	Row (Data)	Columns (Features)	Description
US_flights_2023.csv	6,743,404	24	Main dataset containing domestic U.S. flight schedules, delays, and aircraft information.
weather_meteo_by_airport.csv	132,860	10	Daily weather information (temperature, wind, pressure, etc.) for each airport in 2023.
airports_geolocation.csv	132,860	10	Contains airport identification (IATA), city/state, and precise latitude and longitude details.

These datasets were merged based on shared identifiers such as airport\_id and Tail\_Number, allowing for integrated analysis combining flight activity, weather, and geospatial context. To ensure accurate modeling and reduce bias, canceled and diverted flights were excluded from the main dataset

### B. Long-Term Flight Delay Risk Assessment

To assess the long-term risk of flight delays, the data was split into three parts: training, validation, and test sets. The target variable is binary:

- $c = 1$  (the flight was delayed)
- $c = 0$  (the flight was not delayed)

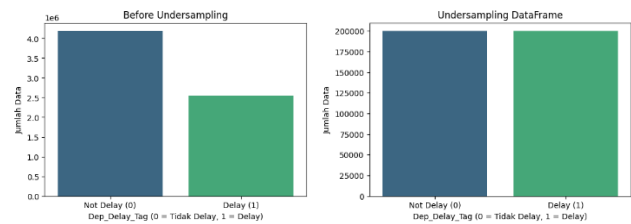
The goal of this analysis is to build a machine learning model with high recall and AUC, so it can effectively detect flights that are truly delayed, even when the data is imbalanced.”.

### C. Undersampling Data

The dataset used in this study comes from three main sources, such U.S. flight data from 2023 (US\_flights\_2023.csv), weather data by airport (weather\_df), and airport geolocation data (airports\_geolocation.csv). These datasets were merged to support further flight delay prediction.

After merging, the combined dataset contained over 6 million rows. In addition to its large size, there was a significant class imbalance in the target variable between class 1 and class 0.

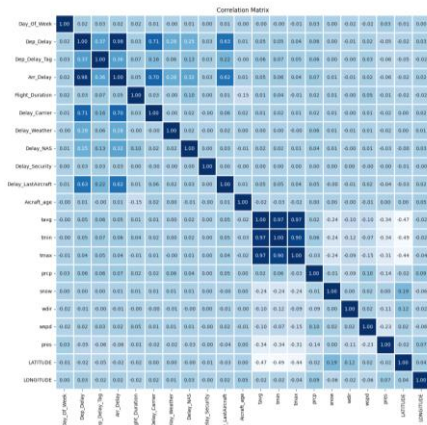
To address this imbalance and reduce the computational load, random undersampling was applied. This method reduces the number of records from the majority class without losing important information, helping to balance the class distribution. In this study, both classes were limited to 200,000 rows, resulting in a balanced training dataset of 400,000 records.



### D. Feature Selections

To enhance model performance and reduce overfitting, a feature selection process was conducted. This involved statistical analysis on both numerical and categorical attributes [12] to identify and remove redundant or irrelevant features.

#### a) Correlation Analysis on Numerical Features



A Pearson correlation matrix was generated to identify linear relationships between numerical features and the target variable Dep\_Delay\_Tag. Highly correlated features such as Dep\_Delay (0.98), Arr\_Delay (0.62), and Delay\_LastAircraft (0.62) were removed to avoid data leakage. Additional features like tmin, tmax, Aircraft\_age, LATITUDE, and LONGITUDE were also discarded due to low correlation (less than 0.1), indicating limited predictive value.

Numerical features removed: Arr\_Delay, Dep\_Delay, Delay\_Carrier, Delay\_NAS, Delay\_Security, Delay\_Weather, Delay\_LastAircraft, tmin, tmax, Aircraft\_age, wdir, wpgt, pres, LATITUDE, LONGITUDE

#### b) Association Analysis on Categorical Geatures

The Chi-Square test of independence was used to assess the statistical significance of categorical features. All tested features had p-values < 0.05. Cramér's V was applied to measure association strength with the target, with Dep\_Delay\_Type (0.59) and Arr\_Delay\_Type (0.51) showing the strongest associations. Despite their significance, features such as Tail\_Number, Dep\_CityName, and airport\_id were removed due to high cardinality, multicollinearity, or risk of overfitting.

Categorical features removed: airport\_id, IATA\_CODE, AIRPORTY, CITY, Tail\_Number, Dep\_CityName, STATE, COUNTRY, Model

#### c) Dropped Features

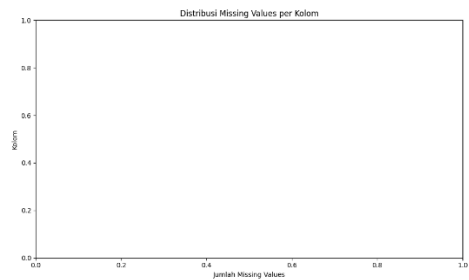
A total of 25 features were excluded based on correlation analysis, association metrics, and domain relevance.

List of dropped features: ["airport\_id", "IATA\_CODE", "AIRPORTY", "CITY", "Tail\_Number", "Dep\_CityName", "STATE", "COUNTRY", "Arr\_Delay", "Delay\_Carrier", "Delay\_NAS", "Delay\_Security", "Delay\_Weather", "Delay\_LastAircraft", "tmin", "tmax", "Aircraft\_age", "wdir", "wpgt", "pres", "LATITUDE", "LONGITUDE", "Dep\_Delay", "Model", "FlightDate"]

### E. Data Preprocessing

#### a) Missing Values

As shown in the diagram, there are no missing values in the current dataset. All columns show zero missing data. This likely happened because the random undersampling process indirectly removed rows with missing values. Since the sampling was random from both classes, it's possible that only complete rows were selected.



#### b) Duplicate Values

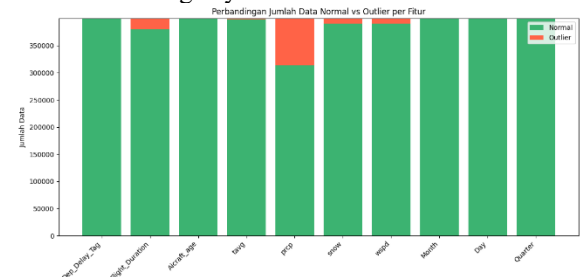
The diagram shows the top 10 duplicated row combinations in the dataset. The Y-axis shows combined feature values, and the X-axis shows how often each combination appears. There are 33 total rows involved 4 unique combinations and 29 duplicate rows. Only the second and later occurrences are counted as duplicates, while the first is treated as the original. Removing duplicates is important to avoid bias in model training, as duplicates can cause the model to "memorize" specific patterns and reduce its ability to generalize to new data.



#### c) Outlier Analysis

The diagram compares the number of normal data (green) and outliers (red) for each numeric feature. The prep feature has the most outliers, followed by Flight\_Duration, both of which contain extreme values that could impact model performance if not handled. Other weather features like snow and wspd also have some outliers, though fewer. Features such as Dep\_Delay\_3hr, Aircraft\_Age, tavg, Month, Day, and Quarter appear stable with few or no outliers, making them safe for modeling.

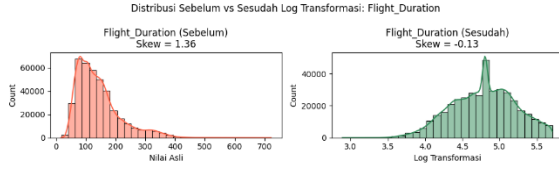
To handle this, outliers were replaced with the median of each feature. Median is more robust to extreme values than the mean, helping reduce bias without removing any rows.



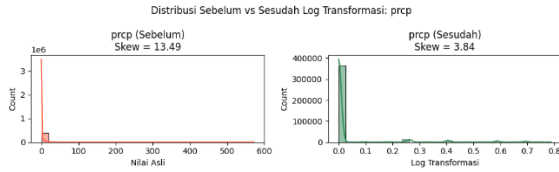
#### d) Skewness Analysis

The Flight\_Duration feature originally had a right-skewed distribution (skewness = 1.36), meaning most flights were short, with a few very long flights.

long ones. This can affect models sensitive to data distribution. After log transformation, the distribution became more symmetric (skewness = -0.13), indicating better normalization.



The **prcp** (precipitation) feature had a very high right skew (skewness = 13.49), with extreme values acting as noise or outliers. After log transformation, skewness dropped to 3.84—an improvement, though still not fully normal. Further steps like **capping outliers**, **Box-Cox transformation**, or removing extreme values can be considered for better results.



#### F. Data Encoding (One-Hot-Encoding)

Many features in the dataset were categorical (e.g., Airline, DepTime\_label) and needed to be converted into numeric form for machine learning. To do this, One-Hot Encoding was applied using `pd.get_dummies()`, turning each category into separate binary columns. This increased the number of features but made the data compatible with all models used, which require numeric input

#### G. Standardization and Principal Component Analysis (PCA)

To handle the large number of features caused by encoding categorical data, standardization and dimensionality reduction were applied before training the model. The data was first standardized using `StandardScaler`, then reduced using PCA (Principal Component Analysis) with `n_components=0.90`, meaning only the components that explain at least 90% of the data's variation were kept. This step helps simplify the features, speed up training, and reduce the risk of overfitting while keeping important information.[13]

#### H. Data Splitting

To ensure robust model evaluation and prevent data leakage, the dataset was partitioned into three distinct subsets: training, validation, and testing sets [14].

- 70% for training: 280,000 samples
- 10% for validation: 40,000 samples
- 20% for testing: 80,000 samples

Stratified sampling was used to maintain balanced class distribution across training, validation, and test sets. The validation set helped tune hyperparameters and monitor overfitting, while the test set evaluated the model's performance on unseen data. This setup is essential for

handling class imbalance and high variance, as in flight delay prediction.

#### I. Machine Learning Models

This study evaluates several machine learning algorithms to predict the likelihood of flight departure delays. The selected models cover different learning approaches, including ensemble methods, probabilistic classifiers, linear models, boosting techniques, and deep learning architectures [15]. The models used include.

- Ensemble Methods: Bagging Classifier, Random Forest, AdaBoost, Gradient Boosting, XGBoost, XGBRF, LightGBM
- Linear Model: Logistic Regression
- Probabilistic Classifier: Gaussian Naive Bayes (GaussianNB)
- Decision Tree-Based: Decision Tree Classifier

#### J. Evaluations Metrics

In evaluating flight delay prediction models, several common metrics were used: precision, recall, and F1-score [16]. Recall (or sensitivity) measures how many actual delayed flights were correctly predicted. Precision shows how many flights predicted as delayed were truly delayed. These metrics are especially important when the data is imbalanced, as in flight delay cases. F1-score, the harmonic mean of precision and recall, gives a balanced view of model performance, especially when prediction errors can have serious impacts.

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{Accuracy} = \frac{TN + TP}{TN + FP + TP + FN}$$

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

#### K. Hyperparameter Tuning

In this study, hyperparameter tuning for the LightGBM model was performed using `HalvingRandomSearchCV`, an efficient method for finding the best parameter combinations with less computational cost. The tuning focused on three main hyperparameters: `learning_rate`, `max_depth`, and `n_estimators`. A Stratified K-Fold cross-validation with 3 folds was used to maintain balanced class distribution in each split. The best model was selected based on the highest accuracy on training data, then evaluated on validation and test sets to ensure its performance generalizes well.

### IV. RESULT AND ANALYSIS

#### A. Experimental Setup

All experiments in this study were done using a personal computer with the following specs: a 12th Gen Intel® Core™ i7-12700H processor (14 cores, up to 4.7 GHz) and 16 GB of DDR4 RAM. The operating system was Windows 11 64-bit. All development, model training, and visualizations were carried out using Python 3.x in the Jupyter Notebook environment.

#### B. Performance Table

The performance table shows that some models, like Random Forest and Decision Tree, are very accurate on the training data but perform much worse on validation and test

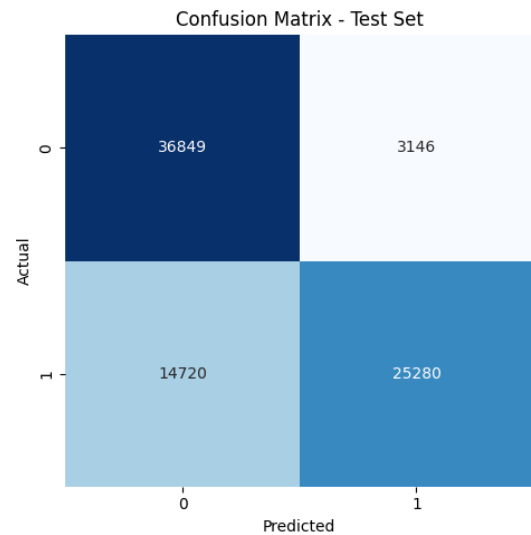
data clear signs of overfitting and poor generalization to new data. On the other hand, models like Logistic Regression, XGBoost, and especially LightGBM perform more consistently and stably across all stages. LightGBM is the best model in this experiment because it achieved the highest and most consistent accuracy, making it a good choice for flight delay prediction systems. Meanwhile, models like Gaussian Naive Bayes performed poorly because they are too simple for a complex dataset. Overall, boosting-based and regularized linear models are the most suitable options for this prediction task.

Model	Train Accuracy	Validation Accuracy	Test Accuracy
Bagging Classifier	0.9839	0.7572	0.7559
Logistic Regresison	0.7839	0.7820	0.7819
GaussianNB	0.5292	0.5305	0.5317
Random Forest	1.0000	0.7697	0.7682
Decision Tree	1.0000	0.6977	0.6952
Gradient Boosting	0.7755	0.7748	0.7737
AdaBoost	0.7489	0.7494	0.7459
XGBoost	0.8228	0.7799	0.7790
XGBRF	0.7616	0.7618	0.7575
LightGBM	0.7896	0.7827	0.7807

This table shows that XGBoost, LightGBM, and Logistic Regression are the best models because they have high F1-scores, meaning they are balanced in detecting both delays and non-delays. In terms of training time, Logistic Regression and LightGBM are also very efficient—much faster than models like Bagging or Gradient Boosting, which take a long time. Although GaussianNB is very fast and has high recall, its precision is low, which means it makes many wrong predictions. Overall, LightGBM and XGBoost are the most suitable because they are accurate, fast, and reliable for predicting flight delays.

Model	Precision	Recall	F1 Score	Training Time (s)
Bagging Classifier	0.8294	0.6476	0.7273	17609.474188
Logistic Regresison	0.8875	0.6459	0.7477	12.022280
GaussianNB	0.5179	0.8831	0.6529	6.426768
Random Forest	0.8518	0.6531	0.7393	1983.385960
Decision Tree	0.6950	0.7044	0.6997	1046.435800
Gradient Boosting	0.8799	0.6365	0.7378	14929.688928
AdaBoost	0.7882	0.6820	0.7313	3137.003448
XGBoost	0.8584	0.6704	0.7529	136.991289
XGBRF	0.8974	0.5911	0.7127	139.395139
LightGBM	0.8875	0.6474	0.7487	63.507046

### C. Confusion Tuning Result



### D. Hyperparameter Tuning Result

	Sebelum Hyperparameter Tuning	Setelah Hyperparameter Tuning
Validation Test	0.7827	0.7788834162562193
Test Set	0.7807	0.7766610413150822

Hyperparameter tuning was done to improve the model's performance by adjusting key settings in the algorithm. In this experiment, HalvingRandomSearchCV was used because it efficiently explores many parameter combinations with limited computing resources.

Before tuning, the model achieved a validation accuracy of 0.7827 and a test accuracy of 0.7807. After tuning, the best model found was a LightGBM Classifier with the following settings: learning\_rate=0.0219, max\_depth=8, and n\_estimators=258.

The tuned model got a validation accuracy of 0.7789 and a test accuracy of 0.7767. Although accuracy slightly dropped (by about 0.004), the performance remained stable. The small difference between validation and test accuracy shows that the model did not overfit and still performs well on new, unseen data. This consistent performance suggests the model configuration is already suitable for real-world use or deployment.

### E. Overfitting Analysis

The model's performance evaluation shows that accuracy is quite balanced across training, validation, and testing datasets 78.24%, 77.89%, and 77.67%, respectively. The small gap between training and validation accuracy (only 0.35%) and between validation and testing (just 0.22%) suggests that the model does not overfit and performs consistently on unseen data.

On the test set, the classification report shows balanced performance, with weighted average scores of precision: 0.80, recall: 0.78, and F1-score: 0.77. The model performs well in predicting both classes, although it's slightly less accurate at detecting delays (Class 1), with a recall of 0.63, compared to 0.92 for on-time flights (Class 0). Overall, the model generalizes well and has a low risk of overfitting.

### F. Result

Using the LightGBM classification model, out of a total of 79,995 records in the test set, the model correctly



predicted 62,129 departure delay labels (Dep\_Delay\_Tag). This shows that the model was able to accurately classify most of the data.

To better understand the model’s performance on key operational entities in the dataset, a detailed analysis was conducted on the top 10 airlines with the most data. The summary is presented in a table showing the number of predicted delays and non-delays for each airline, as well as the prediction accuracy for each one.

Airline	Total	Right	Wrong	Accuracy (%)
Southwest Airlines Co.	18127	12969	5158	71.545209
United Air Lines Inc.	8404	6391	2013	76.047120
Delta Air Lines Inc	11135	8636	2499	77.557252
American Airlines Inc.	11066	8774	2292	79.287909
Alaska Airlines Inc.	2914	2186	728	75.017159
Allegiant Air	1401	1093	308	78.015703
American Eagle Airlines Inc.	2548	2074	474	81.397174
Southwest Airlines Co.	18127	12969	5158	71.545209
Republic Airways	3013	2602	411	86.359111
Endeavor Air	2214	1873	341	84.598013

## V. CONCLUSION

### 1. Conclusion

This study demonstrates a comprehensive data-driven approach to predict flight departure delays using a combination of meteorological data, flight characteristics, and geolocation attributes. Through rigorous data preprocessing including feature selection, encoding, class balancing, and dimensionality reduction combined with a diverse set of machine learning models, we achieved robust predictive performance. Among the models tested, ensemble-based algorithms such as LightGBM exhibited superior results, particularly in terms of recall and F1-score, making them well-suited for handling class imbalance in delay classification tasks. Furthermore, the hyperparameter tuning process and overfitting analysis confirmed that the final models generalize well on unseen data, as evidenced by comparable performance across training, validation, and testing sets.

The feature reduction and PCA implementation effectively minimized model complexity while preserving 90% of data variance, contributing to better generalization and reduced computational cost. The overall pipeline from preprocessing to evaluation was carefully validated and aligned with state-of-the-art practices in predictive modeling. This ensures that the proposed system is not only accurate but also reliable and scalable for real-world applications.

### 2. Future Works

For future work, we aim to enhance user accessibility and interpretability by integrating a web-based User Interface (UI) that visualizes flight delay predictions in real time. Additionally, further improvements can be made by incorporating real-time data streams, exploring deep learning architectures (e.g., LSTM or GRU for time series), and conducting external validation using datasets from different years or countries to test model transferability. These extensions are expected to significantly enhance the usability and impact of the proposed system in operational airline environments.

### ACKNOWLEDGMENT

Link Github:

<https://github.com/ElizabethLauraHelvin/AOL-Data-Analytics>

### REFERENCES

- [1] Chen J, Li M. Chained predictions of flight delay using machine learning. AIAA Scitech 2019 Forum 2019. <https://doi.org/10.2514/6.2019-1661>;ISSUE:ISSUE:10.2514/MSCITECH 19.
- [2] Thiagarajan B, Srinivasan L, Sharma AV, Sreekanthan D, Vijayaraghavan V. A machine learning approach for prediction of on-time performance of flights. AIAA/IEEE Digital Avionics Systems Conference - Proceedings 2017;2017-September. <https://doi.org/10.1109/DASC.2017.8102138>.
- [3] Dai M. A hybrid machine learning-based model for predicting flight delay through aviation big data. Sci Rep 2024;14:1–16. <https://doi.org/10.1038/S41598-024-55217-Z>;SUBJMETA=1042,117,258,639,705;K WRD=COMPUTATIONAL+SCIENCE,C OMPUTER+SCIENCE,INFORMATION +TECHNOLOGY.
- [4] Sun D, Shi S, Wen H, Xu J, Zhou X, Wu J. A hybrid optimization method of factor screening predicated on GeoDetector and Random Forest for Landslide Susceptibility Mapping. Geomorphology 2021;379:107623. <https://doi.org/10.1016/J.GEOMORPH.2021.107623>.
- [5] Borsky S, Unterberger C. Bad weather and flight delays: The impact of sudden and slow onset weather events. Economics of Transportation 2019;18:10–26. <https://doi.org/10.1016/J.ECOTRA.2019.02.002>.
- [6] Škultéry F, Jarošová M, Rostáš J. Dangerous weather phenomena and their effect on en-route flight delays in Europe. Transportation Research Procedia 2021;59:174–82. <https://doi.org/10.1016/J.TRPRO.2021.11.109>.
- [7] Rodríguez-Sanz Á, Cano J, Fernández BR. Impact of weather conditions on airport arrival delay and throughput. IOP Conf Ser Mater Sci Eng 2021;1024. <https://doi.org/10.1088/1757-899X/1024/1/012107>.

- [8] Hsu C-W, Liu C, Liu Z, Mostafavi A. Unraveling Extreme Weather Impacts on Air Transportation and Passenger Delays using Location-based Data 2023.
- [9] Chan P-W, Chen F, Khattak A, Wu K, Dalmau R, Attia J, et al. Modelling the Impact of Adverse Weather on Airport Peak Service Rate with Machine Learning. *Atmosphere* 2023, Vol 14, Page 1476 2023;14:1476. <https://doi.org/10.3390/ATMOS14101476>.
- [10] Kim S, Park E. Prediction of flight departure delays caused by weather conditions adopting data-driven approaches. *J Big Data* 2024;11:1–25. <https://doi.org/10.1186/S40537-023-00867-5/TABLES/20>.
- [11] Sternberg A, Soares J, Carvalho D, Ogasawara E, Carvalho L, Sternberg A, et al. A Review on Flight Delay Prediction n.d. <https://doi.org/10.1080/01441647>.
- [12] Guyon I, De AM. An Introduction to Variable and Feature Selection André Elisseeff. *Journal of Machine Learning Research* 2003;3:1157–82.
- [13] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2016;374. <https://doi.org/10.1098/RSTA.2015.0202>.
- [14] Kuhn M, Johnson K. Applied predictive modeling. *Applied Predictive Modeling* 2013:1–600. <https://doi.org/10.1007/978-1-4614-6849-3/COVER>.
- [15] Naidu G, Zuva T, Sibanda EM. A Review of Evaluation Metrics in Machine Learning Algorithms. *Lecture Notes in Networks and Systems* 2023;724 LNNS:15–25. [https://doi.org/10.1007/978-3-031-35314-7\\_2](https://doi.org/10.1007/978-3-031-35314-7_2).
- [16] Guo T, Papadopoulos P, Mohammed P, Milanovic J V. Comparison of ensemble decision tree methods for on-line identification of power system dynamic signature considering availability of PMU measurements. 2015 IEEE Eindhoven PowerTech, PowerTech 2015 2015. <https://doi.org/10.1109/PTC.2015.7232364>.