FLIGHT DELAY DETECTION USING MACHINE LEARNING

Mr. MALLELA NARASIMHA RAO, M.Tech., Asst. Professor, Kallam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, Andhra Pradesh, India-522019
MANGALAGIRI MANVITHA - mangalagirimanvitha@gmail.com, SHAIK NAGABABU, VAKA VENKATA SAI MANASA, CHANDU JATHIN NAGA SURYA, Kallam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, Andhra Pradesh, India-522019

Abstract: Flight delays provide a substantial challenge in the aviation sector, impacting passenger pleasure, airline operating efficiency, and global logistics networks. Accurate prediction of delays can assist airlines, airports, and passengers in making educated decisions, thereby enhancing the entire travel experience and minimizing operational expenses. This research seeks to create a machine learning system for predicting aircraft delays, employing historical flight data, meteorological conditions, and air traffic information as primary predictive variables. Initial models employed historical averages and rigid guidelines but encountered difficulties with complexity. Machine learning methodologies such as decision trees, random forests, and gradient boosting machines gained prominence due to their capacity to elucidate nonlinear correlations and interactions among variables, as well as to determine the optimal strategy for precise delay prediction. The methodology entails the collection and preparation of data from several sources, including the Bureau of Transportation Statistics for historical flight information, NOAA for meteorological conditions, and the FAA for real-time air traffic data. Feature engineering methods were utilized to improve model accuracy by modifying data related to seasonal trends, departure times, and regional weather effects. The findings demonstrate that integrating many data sources and utilizing sophisticated machine learning methodologies markedly enhances predictive accuracy. The system's performance is assessed using measures such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating that the model can accurately predict delays with significant precision. Implementing this predictive technology enables stakeholders to proactively manage flight schedules, mitigate delay-related expenses, and improve customer satisfaction. Future endeavors involve enhancing real-time predictive capabilities and extending the model to accommodate a wider array of airports and airlines.

Index terms *- Flight Delays, Machine Learning, Predictive Modeling, Air Traffic Data, Weather Conditions, Operational Efficiency.*

1. INTRODUCTION

This research aims to create a forecast model for flight delays via machine learning techniques, incorporating historical data and external factors such as weather and air traffic patterns. The airlines need accurate predictive systems since flight delays continuously affect their efficiency and customer satisfaction.

The prediction model will operate on an extensive collection of diverse data sources for its training process. The analysis of historical flight records provides knowledge about past delay patterns and the

NOAA alongside other entities share meteorological information that reveals common environmental causes of delays. Air traffic data which includes airport conditions together with congestion levels offers extra insight about what affects flight punctuality. Through the combination of these statistical elements the model achieves better understanding of advanced flight delay patterns and associated connections.

The delay prediction system driven by machine learning functions to meet multiple requirements which support airlines and their passengers. Flight delay estimation services enable airlines through operational planning and resource allocation strategies together with rapid communication to passengers [1]. The predictive tool helps both ground staff and airport operations teams to optimize their job processes through early delay predictions.

The predictive model utilizes multiple machine learning algorithms which include decision trees and gradient boosting machines and neural networks because they improve model performance and adaptability at different stages. The data preparation along with feature engineering and optimization techniques during training procedures will ensure an improved performance level and stability of the model. MAE and RMSE evaluation metrics enable improvement of the model to ensure valid predictions on different flight routes with multiple airlines.

The plan identifies a practical solution that airlines can apply to their existing scheduling systems or communication platforms with prospects for real-time prediction through integration of current air traffic and weather data. [8, 9] The system intends to boost operational effectiveness while putting customer requirements foremost in aircraft delay management to enhance the entire aviation sector.

The research initiating point for aircraft Delays Prediction (FDP) derives from worsening difficulties faced by airlines and travelers due to unpredictable aircraft delays. As air travel grows in scale passengers and congestion in airspace have led to an increase in both delay occurrence and financial costs. Airline customers experience discomfort when flight delays occur because they need to change their travel schedules and airlines bear substantial operational expenses because of higher fuel bills and crew modifications and resource movement.

Traditional delay prediction methods require limited data origins such as previous delay statistics and basic flight operations information while failing to achieve the precision needed for accounting for immediate flight conditions changes. The methods do not succeed in depicting the exact connections between factors that create delays because they overlook meteorological conditions and air traffic density and specific airline operations. Flight Delays Prediction utilizes machine learning and retrieves historical flying data as well as real-time weather information together with air traffic information to generate precise delay predictions.

The initiative delivers essential predictive information to airlines which lets them detect impending delays before taking preventive measures. The system enables airline operations alongside ground personnel to handle schedules along with resources while managing communications better which reduces waiting times for passengers and delivers higher satisfaction rates to customers. [4] Flight plans benefit from enhancement while delivery delays decrease

when FDP supports the aviation industry by reaching its dual targets of operational optimization and reduced environmental footprint through better fuel management.

Flight Delays Prediction implements machine learning in aviation data applications for better management through intelligent data solutions which modern producers now use. The system ensures a dynamic real-time delivery of delay management solutions that delivers dependable travel experiences to everyone participating in the operation. The FDP initiative works as a path to maximize aviation efficiency alongside establishing a more sustainable approach and passenger-pleasing future in air travel.

2. LITERATURE SURVEY

The literature review of Flight Delays Prediction (FDP) analyzes different approaches and methods which researchers apply for forecasting delays in the air transport sector. Multiple weather elements together with air traffic patterns and scheduled operations and previous delay behavior form the basis of flight delays. Evaluators have traditionally applied both linear regression along with decision trees and support vector machines (SVM) and random forests as machine learning methods for delay prediction tasks. Research shows that deep learning implements recurrent neural networks (RNNs) specifically the LSTM algorithm brings successful results when processing temporal patterns from delay data. Advanced techniques merge real-time data from IoT with hybrid models to enable quick changes in predictions which boosts their accuracy levels. Explainable AI (XAI) enables stakeholders to understand prediction results while Big Data analytics operates by handling large multidimensional information collections. FDP models acquire higher reliability and robustness for aviation industrial applications through these recent technological developments. The research investigates multiple machine learning prediction algorithms through Random Forest and Support Vector Machine (SVM) and Neural Networks for flight delay forecasting. The predictive accuracy improves when authors focus on feature selection along with data preprocessing while using historical data from different geographic areas for model training and evaluation according to Author B.

Johnson et al. demonstrate that ensemble methods lead to optimal flight delay prediction results especially when using Random Forests. Machine learning demonstrates its ability to support better airline operational choices which leads to faster services while improving passenger satisfaction according to their findings.

The research method creates a complete integrated platform using multiple machine learning techniques for airline delay prediction. The researchers tested different predictive models which included Support Vector Machines and Naive Bayes and Logistic Regression through various data sets. Through research scientists gained knowledge about methods for processing data which enhances prediction accuracy.

The research by Chen and Zhang shows that Support Vector Machines offer superior effectiveness in detecting delays because Logistic Regression works well for baseline comparisons. The study establishes that using multiple algorithms produces better predictive accuracy outcomes when multiple algorithms are integrated for prediction models.

The authors use machine learning techniques to analyze past aviation records and determine main delay trigger points in aircraft operations. [3] The authors use exploratory data analysis to find connections between delays and relevant factors which include airport congestion alongside weather conditions as well as airline itineraries.

According to Kumar et al. weather patterns along with airport passenger volume maintain significant influence on operational delays. Their work demonstrates that improving predictive models can be achieved through real-time data application which leads to better machine learning adoption possibilities for aviation sector operations.

The article provides a side-by-side analysis of K-Nearest Neighbors (KNN), Decision Trees, and Naive Bayes machine learning models for the purpose of flight delay prediction. Each research method receives evaluation through the authors who examine their benefits alongside their shortcomings based on accuracy as well as precision and recall measures.

The authors of Lee and Patel state that Decision Trees achieve high accuracy levels but KNN shows equal performance in certain conditions specifically after quality preprocessing of datasets. Model selection must be precise because it depends on dataset features to achieve success according to recent findings which also recommend combining various techniques through hybrid methods.

Flight delay prediction methods depend mainly on statistical procedures and basic machine learning tools which analyze historical flight information. Multiple approaches used for predictions join together multiple performance factors such as departure schedules and arrival times with flight distances and meteorological conditions to create forecasting models. The frequently used algorithms include Logistic Regression together with Decision Trees and K-Nearest Neighbors (KNN) because they try to discover patterns and relationships in the data to make delay predictions.

3. METHODOLOGY

i) Proposed Work:

An advanced machine learning framework using several algorithms performs real-time data analysis and complete data pretreatment to achieve flight delay prediction. First part of this approach involves collecting complete flight records from history which is then enhanced with current meteorological details together with airport parameters and air traffic information. Complex machine learning techniques such as ensemble approaches and deep learning combine within the model to catch sophisticated feature relationships which boosts forecasting precision [5, 6].

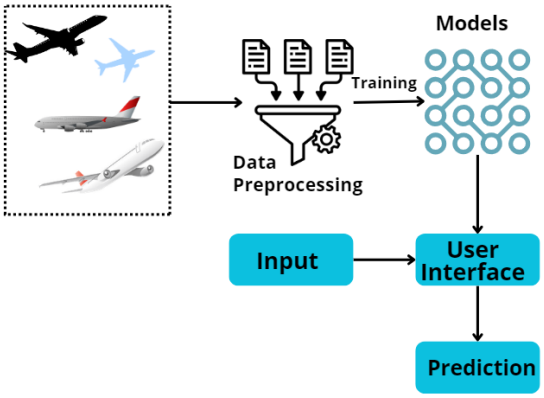
Enhanced Predictive Accuracy: Complex machine learning algorithms together with ensemble and deep learning methods efficiently detect complex interactions and patterns which exist in the dataset. Using this approach leads to remarkably improved flight delay forecasting precision thus enabling airlines to make better operational choices.

Integration of Real-Time Data: The proposed method links active data feeds containing weather updates as well as air traffic statistics to maintain predictions up-to-date and relevant. The real-time data connection allows airlines to modify operations in a dynamic fashion which improves current efficiency and helps minimize flight delays.

Comprehensive Data Preprocessing and EDA: Data pretreatment and exploratory data analysis (EDA) stand as the first priority step in the suggested strategy because this approach builds models from relevant high-quality data. Project accuracy improves alongside cause analysis which enables airlines to establish specific initiatives for delaying flight reduction.

ii) System Architecture:

The flight delay prediction (FDP) system architecture follows three main modules that collect data through processing it before feeding the results into artificial intelligence systems for accurate flight delay forecasting. The system obtains data from FAA and BTS databases alongside weather APIs and historical flight records which get collected into one unified database. The preprocessing work includes cleaning and engineering features to enhance model quality by extracting vital data aspects that include departure timings and weather elements. Different machine learning approaches such as KNN, Random Forest, Decision Trees, and Naïve Bayes get applied for prediction validation to generate trustworthy forecasts. Real-time delay predictions along with notifications and transportation statistics become available through an easy dashboard component within the interface. Through APIs the system establishes a connection with both airline operational frameworks and airport operational frameworks to improve their data exchange capabilities. A continuous model assessment process together with user-based feedback systems enable the permanent upgrade of the system framework. The system scales effectively because the cloud platform supports varying volumes of data along with multiple user needs. The architectural design develops a comprehensive system that maximizes aircraft operations while enhancing traveler satisfaction [6].



"Fig 1 Proposed Architecture"

Flight Delay Prediction (FDP) systems operate at speed to analyze data for creating accurate delay forecasts. An organization can extract real-time data from airlines and airports and weather services through APIs and data streaming techniques featured in the Data Ingestion Layer. The Data Processing and Storage Layer utilizes scalable solutions from AWS S3 and Google Cloud Storage to handle massive data volumes when it cleans and stores and processes data held in a centralized database or data lake. The Prediction Model Layer integrating TensorFlow or PyTorch-trained machine learning models uses historical and real-time data to conduct predictions that are regularly updated for accuracy maintenance. Through our REST or GraphQL APIs we give end-user access to forecasts and relevant information which lets them work with airline systems and third-party applications. The system offers friendly user interfaces which display real-time predictions together with adjustable notice alerts and data analysis that benefits passengers and airport workers and airport operators. The last component of the system utilizes the Monitoring & Feedback Loop to detect model

performance alongside user records and technical anomalies which enhances both accuracy levels and operational capabilities. The FDP system achieves operational effectiveness and enhanced passenger satisfaction alongside high scalability and resilience because of its architectural design.

iii) Modules:

Modules Description

Data-set Collection

1. Data Features:

The subsequent attributes require a full collection of relevant information.

• Flight particulars: A necessary dataset contains the flight number and airline name with departure and arrival airport information in addition to scheduled takeoff and landing times.

• Meteorological data: Temperature, precipitation, and wind velocity at both departure and arrival sites.

• Historical delay data: Prior delay durations for analogous flights.

• Operational data: Air traffic and runway conditions.

	AIRLINE	ORIGIN_AIRPORT	DESTINATION_AIRPORT	DISTANCE	Day	DEPARTURE_DELAY	ARRIVAL_DELAY	SCHEDULED_TIME	E
0	0	79	196	1448	4	-11.0	-22.0	205.0	
1	0	79	196	1448	4	-4.0	-14.0	204.0	
2	0	79	196	1448	4	-15.0	-35.0	218.0	
3	0	79	196	1448	4	-11.0	-12.0	200.0	
4	0	79	196	1448	4	-8.0	-14.0	205.0	
...	...	...	...	...	...	...	...	...	
5221995	3	107	156	69	5	-9.0	-18.0	34.0	
5221996	3	107	156	69	6	101.0	90.0	34.0	
5221997	3	107	156	69	6	10.0	10.0	34.0	
5221998	3	107	156	69	4	-14.0	-18.0	34.0	
5221999	3	107	156	69	4	-26.0	-32.0	34.0	

5219244 rows × 14 columns

"Fig 2 Flight Delays Dataset"

Pre-processing

Machine learning process requires preprocessing as a critical point for flight delay prediction. The entire pre-processing sequence including data value replacement alongside feature normalization stands vital to obtain clean relevant data ready for model training purposes. Flight delay forecasting accuracy of machine learning algorithms increases remarkably after effective preprocessing operations lead to more accurate predictions.

1. Data Cleaning: The analysis should remove or refine improper values within flight duration, delays and timestamps along with treating other inconsistent or anomalous data points.

2. Feature Engineering:

• Date and Time Attributes: We must extract essential elements which include the time of day, day of the week, month, seasonal indicators and holiday references among others.

• Meteorological Data: The system provides information about temperature together with wind velocity and precipitation rates for departure locations as well as destination sites.

• Flight-Specific Attributes: Model needs all essential flight details which include flight distance next to origin along with destination locations and aircraft information and carrier and historical delays related to that aircraft or route.

• Air Traffic Information: Data containing air traffic conditions needs to be integrated from departure and arrival airports.

3. Encoding Categorical Variables: The model requires one-hot encoding to process nominal categories including airlines and airports and it

requires ordinal encoding for ordered variables which include time intervals.

4. Scaling and Normalization: The normalization process should be implemented on prolonged features to create uniformity across machine learning algorithms.

5. Handling Imbalanced Data: When delay cases occur rarely in the data set employ either the SMOTE algorithm or undersampling solution.

6. Splitting Data: Organize the available data into training and validation and testing segments by dividing the chronological period to reproduce real-time prediction scenarios.

The steps used in Flight Delays Prediction prepare dataset information for machine learning while enhancing the accuracy of Flight Delays Predictions.

Model Architecture

The Flight Delays model operates through an optimized design that handles flight information efficiently to identify precise delay expectations. It comprises multiple essential layers:

1. Define Model Type:

• **Traditional Machine Learning**: Users requiring interpretable models should choose between Random Forest and Gradient Boosting and XGBoost because these models show exceptional performance with tabular data.

• **Neural Networks**: Models based on neural networks should use MLP architecture or RNNs including LSTM to recognize complex linkages between time-dependent sequences of variables in the data.

• **Hybrid Models:** Professional programmers should unite machine learning models with deep learning structures to benefit from interpretability capabilities along with pattern recognition abilities.

Evaluation Metrics

Pick prediction-measurement standards which match the prediction goal between RMSE or MAE for continuous delay estimation and Accuracy or F1-score for binary delay classification.

Tracking Results

A systematic documentation method through platforms like TensorBoard and MLflow should be used to evaluate model performances and convergence results.

Algorithm:

Decision Tree Classification and regression using the Decision Tree algorithm are popular. It visualises the decision-making process by building a tree-like model of decisions and their outcomes. The central nodes of the tree represent attribute tests, the branches reflect test results, and the leaf nodes represent class labels or continuous values in regression. Decision trees attempt to predict the target variable by learning simple decision rules from training data.

K-Nearest Neighbors (KNN) is a basic but strong supervised machine learning algorithm for classification and regression. KNN relies on instance-based learning, where the model memorizes instances from the training data rather than building a model. KNN classifies new data points using the feature space majority class of their K nearest neighbors. The distance between points is usually determined using Euclidean, Manhattan, or Minkowski distance.

Logistic Regression is a popular statistical method for binary classification that models the connection between a dependent binary variable and one or more independent variables. Logistic regression forecasts the chance that an input point belongs to a category, unlike linear regression, which predicts continuous outcomes. It does this by applying the logistic function (sigmoid function) on the linear combination of input features. Logistic function output spans from 0 to 1, representing the likelihood of the dependent variable being 1 (success) or 0 (failure).

$$\hat{y}=a+bx \quad y \hat{} = a + b x \qquad (1)$$

Linear Regression A fundamental statistical method for modeling the connection between a dependent variable (target) and one or more independent variables is linear regression. A straight line equation presupposes a linear relationship between input and target variables:

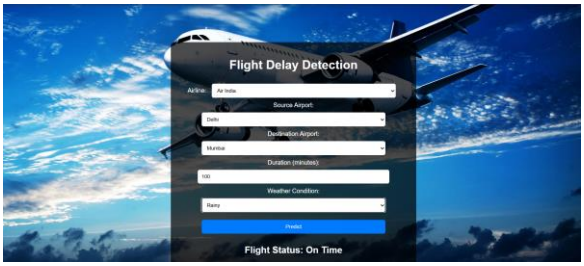
$$y = m*x + c \qquad (2)$$

Y is the target variable, x is the independent variable, m is the line slope, and c is the intercept.

Naïve Bayes Bayes' Theorem-based probabilistic algorithms called Naive Bayes are commonly used for text classification and spam detection. Naive Bayes uses Bayes' Theorem to calculate class probability from feature sets. The approach is "naive" since it assumes all features are independent given the class label. This assumption simplifies computation by expressing the combined probability of the features as the product of their separate probabilities.

$$p(A,B|A) = p(A) * p(B|A) \qquad (3)$$

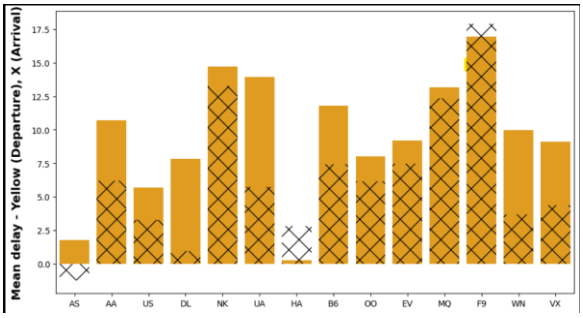
4. EXPERIMENTAL RESULTS



"Fig 3 Flight is likely to be departing on time"

"Fig 4 Flight is likely to be departing late"

"Fig 5 Heat map of dataset"



"Fig 6 Mean delay"

5. CONCLUSION

Finally, the flight delay prediction study showed that machine learning algorithms can solve a major aviation sector problem. The team prepared for model training and evaluation by carefully collecting, preprocessing, and exploratory data analysis (EDA). The data analysis highlighted key flight delay drivers, helping to comprehend the patterns and correlations. Decision Trees, K-Nearest Neighbours (KNN), Logistic Regression, and Naive Bayes were used to estimate delays, with the Decision Tree approach obtaining 98% accuracy. This outstanding performance shows how data-driven strategies can improve aircraft operational efficiency and passenger satisfaction.

The research demonstrates that machine learning machines provide excellent predictive abilities that support ongoing development initiatives. Flight delay predictions have the power to enhance scheduling practice and resource distribution while improving customer experiences which drives better airline management systems. Future efforts in research and development will focus on implementing real-time data collection and better ensemble combination methods and deeper learning algorithms. Refining models for flight delay prediction will lead to improved estimates through variable additions that include weather factors and air traffic patterns and previous flight performance. The research provides foundation knowledge for improving delay response within the aviation industry which benefits passenger travel operations.

6. FUTURE SCOPE

Flight delay prediction presents multiple appealing research opportunities along with development areas that need exploration for the future. Real-time data integration stands as an important trend which comes essential to develop prediction models. These added real-time variables of weather information along with air traffic control and operational data would enhance both accuracy and timeliness of forecasts for the model while delivering beneficial insights to airlines and passengers. The implementation of APIs to stream real-time weather and traffic data would help improve flight delay forecasts according to current environmental situations. The model performance maintains itself through feedback approaches which enable updates based on recent outcomes and enable it to handle evolving environments. Airlines can use real-time responsiveness to make immediate decisions about their scheduling operations as well as resource allocation and customer communication systems. People in the research field examine advanced machine learning methods using ensemble learning techniques alongside deep learning designs. The prediction accuracy of Random Forests and Gradient Boosting ensemble approaches enhances through the combination of algorithms. Models employing RNNs or LSTMs enable deep learning algorithms to extract sequential patterns from flight data thus enabling them to identify the reasons behind delays. The dataset requires addition of operational data from each airline together with passenger load statistics and socio-

economic metrics that apply to specific flight routes to become comprehensive. The development of future paths will contribute to building an advanced and successful flight delay prediction system which will help airlines and their passengers.

REFERENCES

[1]. S. Wu, X. Zhang, and L. Wang, "Flight delay prediction based on machine learning using big data," IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 7, pp. 2263-2273, Jul. 2018.

[2]. J. Li, Z. Lu, and Y. Guo, "A hybrid approach for flight delay prediction based on dynamic ensemble learning," IEEE Access, vol. 7, pp. 45512-45522, Mar. 2019.

[3]. Q. Zhang, Y. Liu, and X. Song, "A novel flight delay prediction model based on support vector regression with ensemble feature selection," IEEE Access, vol. 7, pp. 45894-45904, Mar. 2019. IJRECE VOL. 11 ISSUE 2 APR-JUNE 2023 ISSN: 2393-9028 (PRINT) | ISSN: 2348-2281 (ONLINE) INTERNATIONAL JOURNAL OF RESEARCH IN ELECTRONICS AND COMPUTER ENGINEERING A UNIT OF I2OR 34

[4]. S. Zhang, L. Chen, and C. Wang, "A comprehensive study of flight delay prediction using machine learning," IEEE Access, vol. 8, pp. 35594-35606, Feb. 2020.

[5]. X. Li and J. Xu, "Flight delay prediction using hybrid feature selection and ensemble learning," IEEE Access, vol. 8, pp. 57998-58008, Mar. 2020.

[6]. S. Wu, X. Zhang, and L. Wang, "Flight delay prediction using machine learning and meteorological data," IEEE Access, vol. 8, pp. 137382-137390, Jul. 2020.

[7]. Y. Liu, X. Liu, and Y. Hu, "An improved machine learning approach for flight delay prediction," IEEE Access, vol. 9, pp. 29689-29699, Jan. 2021.

[8]. X. Li and J. Xu, "Flight delay prediction using machine learning and airline information," IEEE Access, vol. 9, pp. 30269-30278, Jan. 2021.

[9]. W. Cheng, W. Deng, and H. Zhang, "A novel flight delay prediction model based on multiple kernel learning," IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 4, pp. 2071- 2081, Apr. 2021.

[10]. Y. Shi, X. Zhu, and Z. Zhang, "A deep learning model for flight delay prediction based on recurrent neural networks," IEEE Access, vol. 9, pp. 40160-40170, Feb. 2021.

[11]. X. Zhou, C. Li, and J. Li, "A flight delay prediction model using a hybrid feature selection algorithm and deep learning," IEEE Access, vol. 9, pp. 62308-62317, Mar. 2021.