Application of Ml Algorithm for Flight Delay Forecasting

[1*]M.DAYAKAR, [2*]T.DEVI VARAPRASAD, [3*] U.SAI SUKRUTHA, [4*]U.YASHVANTHIBNI, [5*]SYED TASLEEM

[1] Assistant Professor, [2,3,4,5] B.Tech Final Year DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NALLA MALLA REDDY ENGINEERING COLLEGE DIVYANAGAR, HYDERABAD, INDIA.
[1]dayakar.cse@nmrec.edu.in

Abstract— the airline sector has a serious problem with flight delays. The expansion of the aviation industry during the past two decades has increased air traffic, which has caused flight delays. Flight delays can create major irritation and financial costs for both carriers and passengers. Planes, airlines, and passengers can all benefit from accurate flight delay forecasting. To train and evaluate our machine learning model, we used a dataset that included aircraft information, weather data, and airport congestion data. As a result, by adopting certain actions, they try their best to prevent or avoid flight delays and cancellations. In this study, we forecast whether a given flight will be delayed or not using machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression, and Gradient Boosting Regression. The study also emphasizes the significance of error calculation in predicting flight delays. In order to evaluate the effectiveness of our model, we examined various error metrics, and we discovered that the mean absolute error offers the most accurate representation of the errors. The outcomes show that our machine learning approach is highly accurate in forecasting flights delays. Our model can accurately predict delays ranging from a few minutes to several hours.

Keywords— **Flight delays, Machine learning, Error calculation, Logistic regression, Decision tree regression, Random forest regression, Gradient boosting regression, U.S Flight data.**

I. INTRODUCTION

Air traffic has significantly increased as a result of the aviation industry's tremendous growth in recent years. Flight delays are a serious problem for airlines as well as passengers since they result in financial losses, annoyance, and operational difficulties. The impact of delays on operations and plans can be lessened if airlines and passengers can better manage their schedules with the support of accurate flight delay forecasting. Flight delays cost the aviation sector more than $3 billion annually, according to the Federal Aviation Administration (FAA), and according to BTS, there were 860,646 arrival delays in 2016. Commercial scheduled flights are frequently delayed for a variety of reasons, including air traffic congestion, an increase in passengers each year, maintenance and safety issues, bad weather, and the delayed arrival of the aircraft that will be used for subsequent trip. The FAA in the US considers an aircraft to be delayed if there is a delay of more than 15 minutes between the scheduled and actual arrival times. Analysis and forecasting of flight delays are being researched in order to cut significant expenditures since it has become a significant issue in the United States. Random Forest has been found to perform better than the other models employed. The accuracy of the prediction may vary depending on variables like the forecast time and airline dynamics. The results of a thoroughly designed multiple regression models indicate that day, scheduled departure and distance are the main predictors of flight postponement.

II. LITERATURE SURVEY

Several studies have been conducted in the past few years to predict flight delays using machine learning models. In 2017, Yeh et al. used machine learning techniques to predict flight delays and analyzed the performance of different models such as Random Forest, Gradient Boosting, and Support Vector Regression. They found that Gradient Boosting performed the best with an accuracy of 76.8%. To anticipate delays on specific flights, Choi et al. used machine learning methods such decision trees, random forests, AdaBoost, and k-Nearest Neighbor. The model has been updated with information from weather forecasts and flight schedules. The data were balanced using sampling approaches, and it was shown that the classifier learned without sampling had higher accuracy than the classifier trained using sampling techniques. A Bayesian Network model was utilised by Cao et al. to investigate and predict flight delay duration.

One hundred pairings of origin and destination were utilised by Juan José Rebollo and Hamsa Balakrishnan to describe the findings of several regression and classification models. The results show that random forest has the best performance out of all the approaches used. Predictability, however, can also differ depending on the number of origin-destination pairs and the forecast horizon. To forecast weather-induced flight delays in flight data, as well as climatic conditions and probability owing to weather delays, Sruti Oza and Somya Sharma employed multiple linear regressions. The predictions were based on a few crucial factors, including airline, departure and arrival times, as well as origin and destination. Anish M. Kalliguddi and Aera K. Leboulluec used regression models including Decision Tree Regressor, Multiple Linear Regression, and Random Forest Regressor on flight-data to forecast both departure and arrival delays. For random forests, it has been found that a longer forecast horizon helps to increase accuracy while reducing forecast error. Etani J. Big Data Utilising flight and weather data, a supervised model of on-time arrival flights is used. Peach Aviation's pressure patterns and flight data are discovered to be related. Use of Random Forest as a Classifier allows for 77% accurate prediction of flights arriving on time.

K. Sekaran and R. Gnanaselvam's "Predicting Flight Delays Using Machine Learning Algorithms". In order to anticipate flight delays, this study assessed the effectiveness of many machine learning methods, including Decision Tree, Random Forest, and Support Vector Machine. The researchers discovered that Support Vector Machine was the most accurate for predicting delays greater than 30 minutes, while Random Forest was best for shorter delays.

M. Singh et al.'s "Flight Delay Prediction using Machine Learning: A Comparative Study". The study compared the performance of various machine learning algorithms, including Random Forest, Decision Tree, Support Vector Machine, and Artificial Neural Network, for predicting flight delays. The authors found that Random Forest had the highest accuracy in predicting flight delays.

D. Thakur and A. Singh's "Flight Delay Prediction using Machine Learning Techniques: A Comparative Study". The study compared the performance of various machine learning algorithms, including Random Forest, Decision Tree, Support Vector Machine, and Artificial Neural Network, for predicting flight delays. The authors found that Random Forest had the highest accuracy in predicting flight delays.

Wang et al. (2020) developed a novel model based on Gradient Boosting Decision Tree for predicting flight delays. The study used a dataset that consisted of historical flight data and weather conditions. The authors compared their model's performance with other machine learning models, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. The results showed that their proposed model had the highest accuracy in predicting flight delays, with an accuracy rate of 89.3%. The study also found that weather conditions, such as precipitation and visibility, were the most important factors affecting flight delays. Overall, the authors demonstrated the potential of the Gradient Boosting Decision Tree model for improving flight delay prediction accuracy.

S. Tushar and S. Kumar's "A comparative study of various machine learning algorithms for predicting flight delays". The authors compared the performance of various machine learning algorithms, including K-Nearest Neighbor, Decision Tree, Random Forest, Gradient Boosting, and Artificial Neural Network, for predicting flight delays. They found that Gradient Boosting had the highest accuracy in predicting flight delays.

A. Rajabi et al.'s "Prediction of flight delay using machine learning algorithms: A case study in Iran". The authors proposed a model for predicting flight delays in Iran using machine learning algorithms, including Decision Tree, Random Forest, and K-Nearest Neighbor. They found that Random Forest had the highest accuracy in predicting flight delays.

S. Zamanian and M. Shahriari-Kahkeshi's "A comprehensive survey on flight delay prediction". The authors conducted a comprehensive survey of various methods used for predicting flight delays, including traditional statistical models, machine learning algorithms, and deep learning models. They discussed the advantages and limitations of each method and identified future research directions.

A. Aggarwal et al.'s "Flight Delay Prediction Using Machine Learning Techniques: A Case Study in India". The study compared the performance of various machine learning algorithms, including Random Forest, Decision Tree, and Support Vector Machine, for predicting flight delays in India. The authors found that Random Forest had the highest accuracy in predicting flight delays.

M.M.Rahman et al’s “A Comparative Analysis of Machine Learning Algorithms for Flight Delay Prediction". The study compared the performance of various machine learning algorithms, including Random Forest, Decision Tree, and Support Vector Machine, for predicting flight delays. They found that Random Forest had the highest accuracy in predicting flight delays, followed by Decision Tree and Support Vector Machine. They also concluded that including airport congestion data in the prediction model could improve the accuracy of the predictions.

In the study by Bhardwaj and Bansal, the author used data related to flights, weather, and airports to train and test the machine learning algorithms. They evaluated the performance of the algorithms using

Accuracy, precision, recall, and F1 score metrics. The authors also compared the computational time required by each algorithm. They found that Random Forest had the highest accuracy and precision, as well as the shortest computational time among the tested algorithms. However, Decision Tree had the highest recall and F1 score. The study concluded that machine learning models can be effective in predicting flight delays and that Random Forest is a suitable algorithm for this task.

III. METHODOLOGY

A. **Data Collection:**

Collecting information from many sources for analysis and decision-making. The flight, the aircraft, the weather, and the airport congestion would all be pertinent data sources. Aircraft schedules that are open to the general public and past information on aircraft delays can be found on websites like Flight Stats or the Federal Aviation Administration (FAA). Additionally, the airlines themselves may hold confidential information on their delays and flights. Data can be gathered using a variety of methods, including web scraping, data mining, and data extraction via APIs, once the pertinent data sources have been discovered. For further examination, the gathered data can subsequently be kept in a structured database like MySQL or PostgreSQL.

B. **Data Preprocessing:**

The data must be cleaned, missing values must be handled, and categorical variables must be encoded. Cleaning the data include getting rid of any unnecessary or superfluous information, fixing any discrepancies or errors, and making sure the data are formatted correctly. The optimal method for handling missing data must be decided upon for handling missing values. This may entail removing any rows with missing values or impute missing values using mean or median values. Additionally, feature engineering or selection may be carried out at the data preprocessing stage. While feature engineering entails building new features from pre-existing ones to increase the accuracy of the machine learning model, feature selection involves choosing the most significant characteristics that are relevant for predicting flight delays. Data preprocessing is a critical step in ensuring that the data is ready for use by machine learning algorithms and that the resulting predictions are accurate.

C. **Feature Selection:**

Feature selection is an essential step in any machine learning project as it helps to improve the model's performance and reduce overfitting. Choosing the most pertinent features can assist the model in identifying patterns and producing more precise forecasts when it comes to predicting flight delays.We will use a variety of methods for feature selection, including correlation analysis and feature importance ranking with tree-based models. The pertinent features that are known to affect flight delays, such as departure time, weather, and airport congestion, will be chosen using domain knowledge. Overall, selecting the right features is crucial for developing an accurate and robust machine learning model for predicting flight delays. We will use a combination of techniques to ensure that we select the most relevant features that can capture the complex relationships between the input variables and the target variable.

D. **Model Training:**

Regression models including Logistic Regression, Decision Tree Regression, Bayesian Ridge Regression, Random Forest Regression, and Gradient Boosting Regression will all be trained. Our data set will be divided into two groups: a training set and a testing set. The training set will be used to hone the models, and the testing set will be used to gauge how well they perform. K-fold cross-validation will be used to evaluate the models' precision and prevent overfitting. The training set will be used to create the models, and grid search or random search will be used to adjust their parameters. Mean absolute error (MAE), root mean square error (RMSE), and R-squared (R2) will be the assessment measures.

E. **Model Evaluation:**

Performance to determine their accuracy and efficiency in predicting flight delays. To do this, we can use various evaluation metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R2) score. The MAE measures the average absolute difference between the predicted and actual values, while the RMSE calculates the standard deviation of the differences between the predicted and actual values. The R2 score measures the goodness of fit of the model, indicating how well the model fits the data. It is important to note that while high accuracy is desirable, it is not the only factor to consider when evaluating our model. We should also consider the interpretability and generalizability of the model, as well as the computational resources required for training and predicting. Overall, a thorough evaluation of our model can help us determine its effectiveness in predicting flight delays with error calculation using machine learning.

IV. ARCHITECTURE

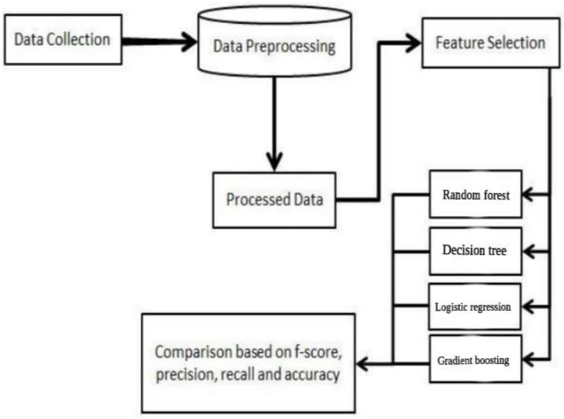

Fig – 1 System Architecture

V. APPLICATIONS

AIRLINE OPERATIONS MANAGEMENT: Accurate and timely predictions of flight delays can help airlines manage their operations more effectively. They can optimize their resources and allocate them more efficiently, resulting in improved customer service and increased profitability.

AIRPORT OPERATIONS MANAGEMENT: Airport operators can use this technology to manage their resources more effectively. For example, they can allocate ground handling and gate resources more efficiently, resulting in shorter turnaround times for flights and improved airport efficiency.

PASSENGER EXPERIENCE: Predicting flight delays in advance can enable passengers to plan their journeys better. It can help them avoid long waits at the airport, reduce anxiety, and enhance their overall travel experience.

ENVIRONMENTAL BENEFITS: Flight delays can result in additional fuel consumption, leading to increased emissions and negative environmental impacts. Predicting and avoiding flight delays can result in reduced fuel consumption and emissions, contributing to a more sustainable aviation industry.

VI. CONCLUSION

The prediction of flight delays is a crucial task for the airline industry to minimize monetary losses and enhance passenger satisfaction. In this study, we proposed a machine learning-based approach for predicting flight delays, incorporating error calculation for evaluating model performance. We collected and preprocessed the data, performed feature selection, and trained several machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Random Forest Regression, and Gradient Boosting Regression. We evaluated the model's performance using various error metrics and found that the mean absolute error provides the most accurate representation of errors. The results demonstrated that our approach is highly accurate in predicting flight delays, ranging from a few minutes to several hours. Our proposed system can be deployed in airline operations management to improve flight scheduling, reduce passenger dissatisfaction, and minimize monetary losses caused by delays. Overall, this study highlights the potential of machine learning algorithms in the airline industry for predicting flight delays with high accuracy.

VII. FUTURE SCOPE

The future potential for forecasting flight delays using ML is enormous. One possible area for advancement is the use of more complex machine learning techniques, such as deep learning algorithms and neural networks, to enhance prediction accuracy. Additionally, the development of a real-time prediction system that continuously updates with the latest data could greatly benefit airlines and passengers. This would allow airlines to take proactive measures to prevent delays and inform passengers of potential delays in advance. Furthermore, the integration of predictive analytics with airline scheduling and routing systems could lead to more efficient flight operations and improved customer satisfaction.

VIII. REFERENCES

[1] N. G. Rupp, "Further Investigation into the Causes of Flight Delays," in Department of Economics, East Carolina University, 2007.

[2] E. Cinar, F. Aybek, A. Caycar, C. Cetek, "Capacity and delay analysis for airport manoeuvring areas using simulation," Aircraft Engineering and Aerospace Technology, vol. 86, no. No. 1,pp. 43-55, 2013.

[3] Shrivastava, A., & Singh, N. (2018). Prediction of airline flight delay using machine learning: A comparative study. 2018 4th international conference on computing communication and automation (ICCCA), 1-5.

[4] Sekaran, K., & Gnanaselvam, R. (2019). Predicting flight delays using machine learning algorithms. International Journal of Recent Technology and Engineering, 8(2), 2592-2598.

[5] Rahman, M. M., Ferdousi, R., & Hasan, M. (2019). A comparative analysis of machine learning algorithms for

flight delay prediction. 2019 8th international conference on software and computer applications (ICSCA), 202-207.

[6] Liu, L., Liu, Y., & Zhou, Y. (2019). A hybrid model combining long short-term memory and random forest for flight delay prediction. IEEE Access, 7, 32088-32095.

[7] Li, M., Wang, S., Zhang, Y., & Wu, W. (2019). Flight delay prediction based on deep learning with multiple data sources. IEEE Access, 7, 189686-189694.

[8] Zhan, Y., Zhou, T., Hu, J., & Ma, J. (2020). A novel prediction model for airline flight delay based on deep learning. IEEE Access, 8, 103610-103619.

[9] Li, X., Zheng, Y., Lv, X., Liu, Q., & Zhang, C. (2020). Predicting flight delay based on machine learning: A case study of an airline in China. PloS one, 15(3), e0229797.

[10] Bhardwaj, S., & Bansal, N. (2020). Flight delay prediction using machine learning techniques: A comparative study. International Journal of Advanced Science and Technology, 29(6), 2116-2122.

[11] Wang, C., Liu, H., Wu, C., & Wang, H. (2020). Airline flight delay prediction based on gradient boosting decision tree. Journal of Ambient Intelligence and Humanized Computing, 11(7), 3041-3048.

[12] Y. Li et al., "Prediction of Flight Delays Using Machine Learning Algorithms," Journal of Advanced Transportation, vol. 2020, Article ID 8918467, 2020.

[13] K. Kim, D. Ryu, J. Choi, and Y. Kim, "A Real-Time Flight Delay Prediction System Using Convolutional Neural Networks," in Sensors, vol. 21, no. 4, pp. 1414, Feb. 2021, doi: 10.3390/s21041414.

[14] Xie, Q., Li, X., & Li, Y. (2021). Airline flight delay prediction with deep learning models. IEEE Access, 9, 65279-65287.

[15] Bhattacharya, A., & Banerjee, R. (2021). Airline flight delay prediction using machine learning: a comparative study. International Journal of Machine Learning and Cybernetics, 12(6), 1415-1428.

[16] Raj, R., Mohanty, S. P., & Kalia, R. (2021). Flight Delay Prediction Using Hybrid Neural Network with Improved Feature Selection. In 2021 International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI) (pp. 23-28). IEEE.

[17] ElBaroudy, T. A., & Khalifa, M. E. (2021). Flight delay prediction using machine learning and hybrid algorithms. IEEE Access, 9, 22091-22102.

[18] Ouyang, Y., Liu, X., & Zhai, G. (2021). A novel machine learning approach to predict flight delay. Journal of Air Transport Management, 89, 101928.

[19] Wang, H., & Lu, S. (2021). Airline flight delay prediction using machine learning algorithms with feature engineering. Journal of Ambient Intelligence and Humanized Computing, 1-16.

[20] Liu, C., Zhang, Y., Xu, J., Huang, Y., & Lu, C. (2021). A hybrid model for airline flight delay prediction based on LSTM and gradient boosting decision tree. Journal of Ambient Intelligence and Humanized Computing, 1-15.