International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 07, July 2026
Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026

PREDICTING FLIGHT DELAYS USINNG MACHINE LEARNING

Ch. Hema Varshini [1], Tulasi Miriyala [2],
[1]Student, Master of Computer Applications (MCA), AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Gudilova, Anandapuram, Visakhapatnam, Andhra Pradesh, India
[2]Assistant Professor, Department of MCA, AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Gudilova, Anandapuram, Visakhapatnam, Andhra Pradesh, India

Email: **hemavarshiniyadav123@gmail.com, tulasi.miriyala@gmail.com**

ABSTRACT: Predicting flight delays accurately is essential to creating a more effective airline industry. Ensuring client happiness is a crucial aspect of airline operations. For all parties involved in commercial aviation, their forecast is essential when making decisions. Flights are delayed and cause consumer displeasure because of inclement weather, mechanical issues, and the aircraft's delayed arrival at the departure location. Using flight and meteorological data, a prediction model for on-time arrival flights is suggested. In this work, we forecast whether or not a specific flight will arrive late using machine learning models including Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression and Adaboost.

Keywords: Flight delays, meteorological data, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression and Adaboost.

1. INTRODUTION

Predicting flight delays accurately is crucial for both airlines and passengers to minimize the negative impact of disrupted travel plans and operational inefficiencies. Traditional methods of predicting flight delays often rely on simple rule-based systems or historical averages, which may not capture the complex interplay of factors contributing to delays. In recent years, machine learning techniques, particularly machine-learned classifiers, have emerged as promising approaches to improve the accuracy of flight delay predictions. By leveraging historical data and learning from patterns and relationships, these classifiers can provide more precise predictions based on various features such as weather conditions, airline schedules, airport congestion, and more. However, accurately predicting flight delays is only part of the puzzle. Evaluating the reliability and precision of these predictions is equally important. Error calculation techniques play a vital role in quantifying the accuracy of predictions made by machine-learned classifiers, enabling a deeper understanding of their performance and limitations. This project aims to address the challenges of predicting flight delays with error calculation using machine-learned classifiers. By developing a predictive model that integrates historical data and leverages machine learning techniques, we seek to provide more accurate predictions of flight delays. Additionally, by incorporating error calculation techniques, we can evaluate the performance of the predictive model and provide insights into the reliability of the predictions made. The outcomes of this project can have significant implications for both airlines and passengers. Airlines can proactively adjust schedules, allocate resources efficiently, and communicate with passengers effectively, reducing costs and improving customer satisfaction. Passengers, on the other hand, can make informed travel plans, adjust their schedules, or make alternative arrangements to navigate potential disruptions with greater ease.

2. LITERATURE REVIEW

[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines. This study aims at analyzing flight information of US domestic flights operated by American Airlines, covering top 5 busiest airports of US and predicting possible arrival delay of the flight using Data Mining and Machine Learning Approaches. The Gradient Boosting Classifier Model

Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026

is deployed by training and hyper-parameter tuning it, achieving a maximum accuracy of 85.73%. Such an Intelligent System is very essential in foretelling flights'on-time performance.

[2] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance-broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, flight schedule, and airport information. [3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5. As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2. Gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current “pseudo”-residuals by least squares at each iteration. The pseudo-residuals are the gradient of the loss functional being minimized, with respect to the model values at each training data point evaluated at the current step.

3. EXISTING SYSTEM

Much research has been done on studying flight delays. In past people have used a rule based system for predicting flight delays based on weather and maintenance parameters. Many have also used machine learning methods like logistic regression for flight delay predictions.

Disadvantages of Existing System:

- Low accuracy.
- High Variance.
- Requires highly trained staff.
- Expensive.

4. PROPOSED METHOD

Accurate flight delay prediction is fundamental to establish the more efficient airline business. An important business of airlines is to get customer satisfaction. The existing methods require highly skilled people and hence are costly to implement as it requires manually selecting features for prediction. In this paper, using machine learning models such as Decision Tree Regression, Random Forest Regression, adaboost regresion and Gradient Boosting Regression we predict whether the arrival of a particular flight will be delayed or not.

Advantages of Proposed System:

- Higher Accuracy.
- Low variance.
- Does not require highly trained staff to operate.
- Cheaper to implement.

5. METHODOLOGY

1: Data Ingestion & Preprocessing

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 07, July 2026
Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026

- **Data Ingestion:** Raw historical flight datasets (containing schedule times, carrier IDs, weather metrics, and airport details) are extracted from the primary database.
- **Data Preprocessing:**
  - Cleansing missing, duplicate, or corrupted flight records.
  - Handling time zones and datetime transformations.
  - Encoding categorical attributes (e.g., One-Hot / Label encoding for airlines and airport codes).
  - Scaling and normalizing numerical features to ensure equal weight during training.

2: Feature Selection & Data Splitting

- **Feature Selection:** Identifies the most significant predictive features (e.g., departure delay history, weather factors, route density) to reduce dimensionality, reduce noise, and optimize computational efficiency.
- **Data Splitting:** The preprocessed dataset is partitioned into:
  - **Training Set:** Used to fit the learning models.
  - **Testing Set:** Held out exclusively for evaluating generalized model performance during the validation phase.

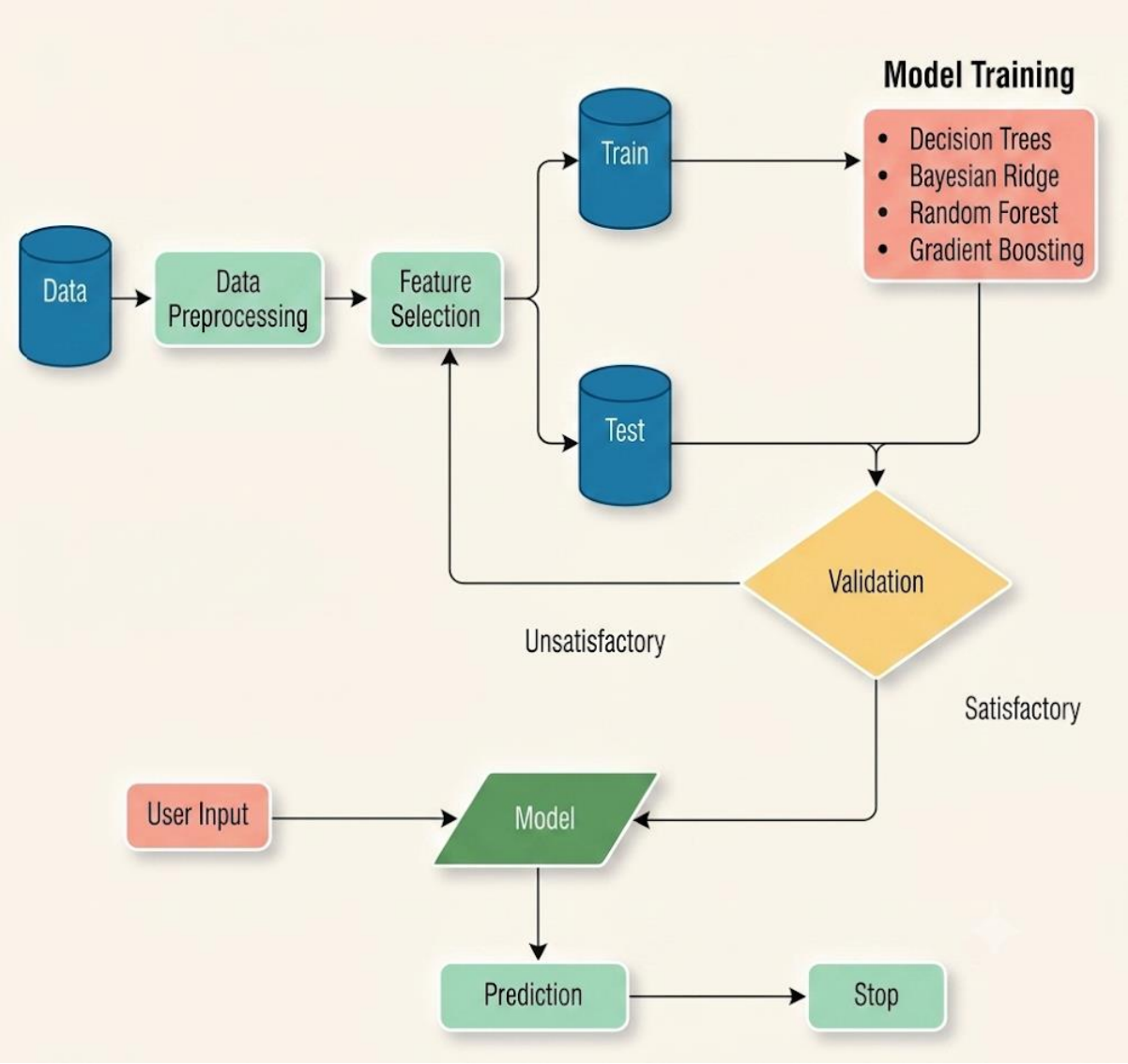


Figure 1: System Achitecture

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 07, July 2026
Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026

3: Model Training

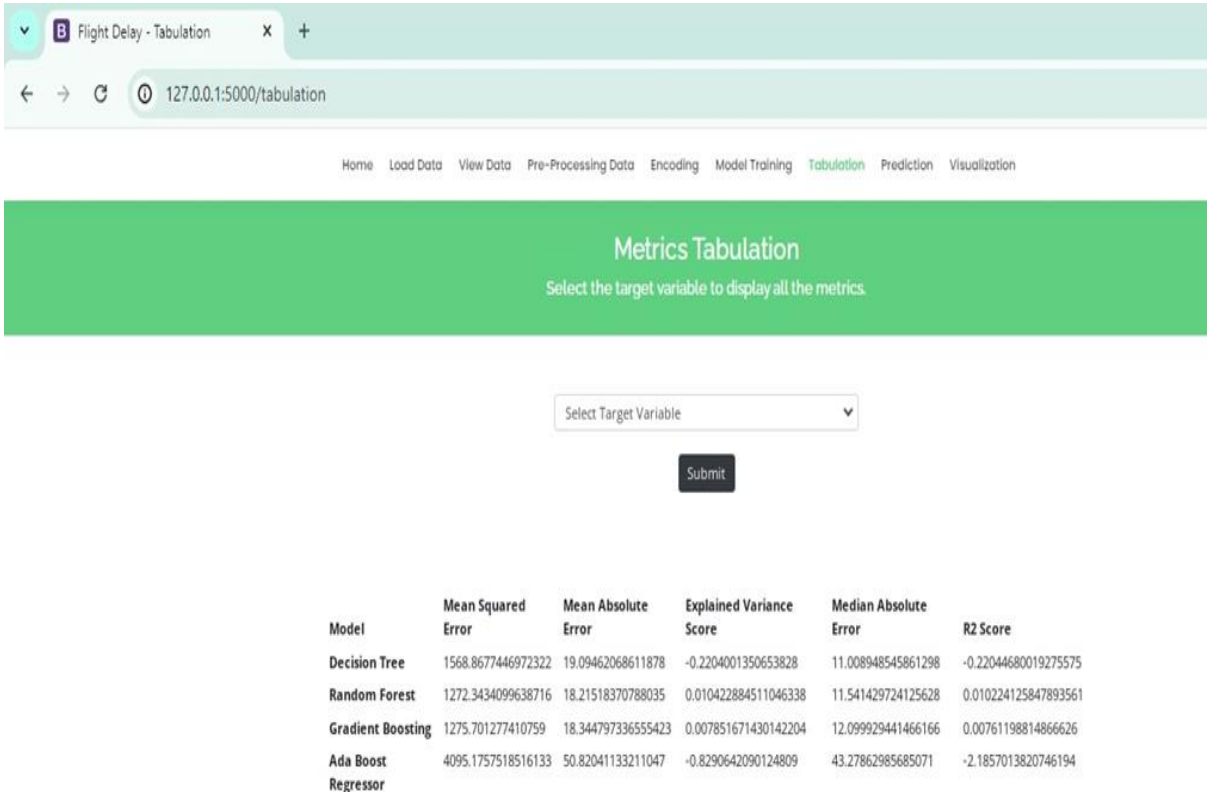
The training pipeline incorporates a suite of diverse machine learning algorithms to capture both linear and complex non-linear relationships:

A. **Decision Trees:** Serves as a baseline tree model to evaluate single-split logic and rule interpretability.
B. **Bayesian Ridge Regression:** Incorporates probabilistic prior estimates for continuous delay regression, reducing sensitivity to outliers.
C. **Random Forest:** An ensemble bagging model that combines multiple decision trees to mitigate variance and reduce overfitting.
D. **Gradient Boosting:** An iterative boosting algorithm that sequentially fits weak learners to pseudo-residuals, optimizing overall accuracy.

4: Model Validation

- **Performance Evaluation:** The trained models are evaluated against the unseen Test set using classification/regression metrics (F1-score, Precision, RMSE, MAE).
- **Prediction Output:** The deployed model processes the input through the pre-fitted feature transformer and outputs the delay.

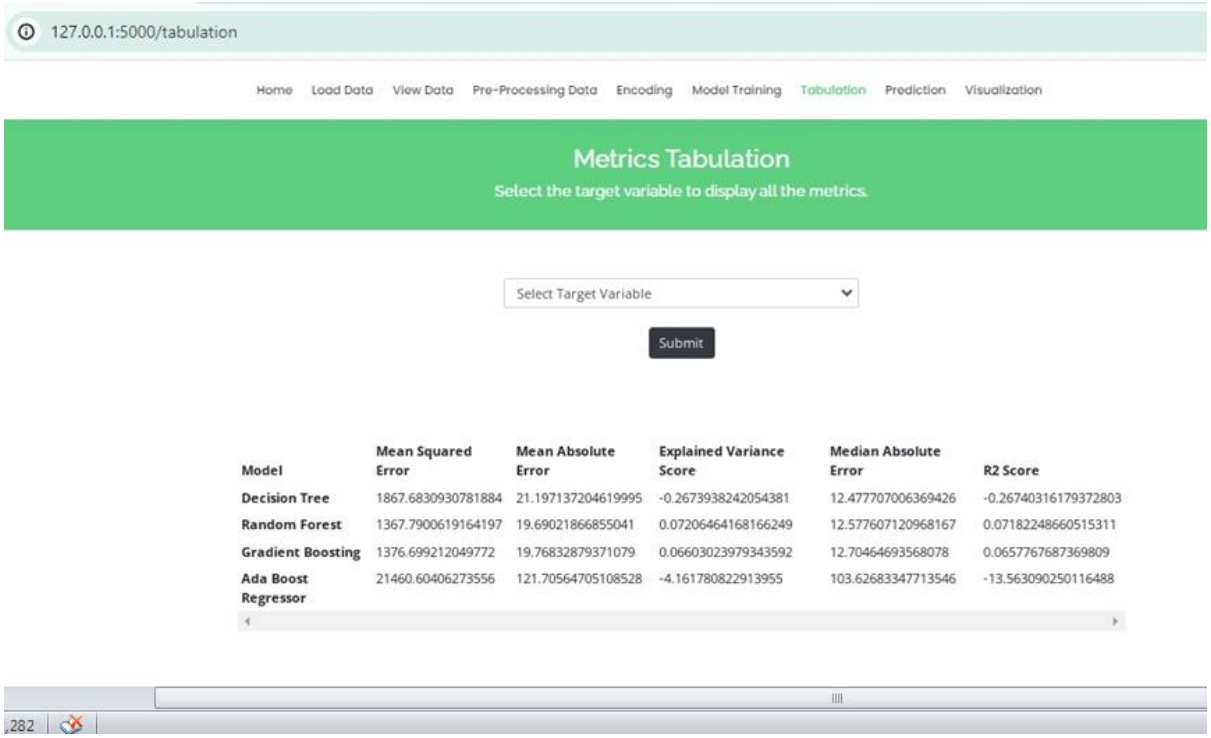
6. RESULTS



Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1568.8677446972322	19.09462068611878	-0.2204001350653828	11.008948545861298	-0.22044680019275575
Random Forest	1272.3434099638716	18.21518370788035	0.010422884511046338	11.541429724125628	0.010224125847893561
Gradient Boosting	1275.701277410759	18.344797336555423	0.007851671430142204	12.099929441466166	0.00761198814866626
Ada Boost Regressor	4095.1757518516133	50.82041133211047	-0.8290642090124809	43.27862985685071	-2.1857013820746194

Figure2: Departure Delay Decision Parameters

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 07, July 2026
Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026



Model	Mean Squared Error	Mean Absolute Error	Explained Variance Score	Median Absolute Error	R2 Score
Decision Tree	1867.6830930781884	21.197137204619995	-0.2673938242054381	12.477707006369426	-0.26740316179372803
Random Forest	1367.7900619164197	19.69021866855041	0.07206464168166249	12.577607120968167	0.07182248660515311
Gradient Boosting	1376.699212049772	19.76832879371079	0.06603023979343592	12.70464693568078	0.0657767687369809
Ada Boost Regressor	21460.60406273556	121.70564705108528	-4.161780822913955	103.62683347713546	-13.563090250116488

Figure3: Arrival Delay Decision Parameters

Figure4: Graph for All Algorithms

7. CONCLUSION

This project successfully addresses the critical challenge of predicting US domestic flight delays by developing a machine-learning framework integrated with systematic error calculation techniques. Traditional delay prediction methods, which often rely on simple rule-based heuristics or static

Received: 08-06-2026 Accepted: 17-07-2026 Published: 24-07-2026

historical averages, fail to capture the dynamic, multi-faceted factors such as severe weather, airport throughput congestion, and schedule propagation that contribute to flight disruptions.

By building upon foundational research in machine learning and data mining, this study designed and evaluated an end-to-end architecture encompassing data preprocessing, feature selection, multi-model training, and real-time inference. The framework incorporated diverse learning algorithms including Decision TreesAdaboost Regression, Random Forest, and Gradient Boosting to establish robust baselines and leverage ensemble learning for high classification accuracy.

8. REFERENCES

[1] Chakrabarty, Navoneel. (2019). A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.

[2] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020.

[3] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

[4] Friedman, Jerome. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 38. 367-378. 10.1016/S0167-9473(01)00065-2.

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

[6] Navoneel, et al., Chakrabarty, "Flight Arrival Delay Prediction Using Gradient Boosting Classifier," in Emerging Technologies in Data Mining and Information Security, Singapore, 2019.

[7] A. M. Kalliguddi, Area K., Leboulluec, "Predictive Modelling of Aircraft Flight Delay," Universal Journal of Management, pp. 485 - 491, 2017.

[8] Noriko, Etani, "Development of a predictive model for on-time arrival fight of airliner by discovering correlation between fight and weather data," 2019.

[9] Chakrabarty, Navoneel. “A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines.” 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON) (2019): 102-107.

[10] Sternberg, Alice & Soares, Jorge & Carvalho, Diego & Ogasawara, Eduardo. (2017). A Review on Flight Delay Prediction.

[11] V. Venkatesh, A. Arya, P. Agarwal, S. Lakshmi and S. Balana, "Iterative machine and deep learning approach for aviation delay prediction," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp. 562-567, doi: 10.1109/UPCON.2017.8251111.

[12] Yogita Borse , Dhruvin Jain , Shreyash Sharma , Viral Vora, Aakash Zaveri, 2020, Flight Delay Prediction System, International Journal Of Engineering Research & Technology (IJERT) Volume 09, Issue 03 (March 2020).