Improving Passenger Experience: Predicting Airline Delays Through Machine Learning

S.Siva Nageswara Rao
Dept Of CSE,
Narasaraopeta Engineering College
Narasaraopeta,India
profssnr@gmail.com

Yamini Chandana
Dept Of CSE
Narasaraopeta Engineering College
Narasaraopeta,India
chandana.nrtnec@gmail.com

Kode Venkata Naveen
Dept Of CSE
Narasaraopeta Engineering College
Narasaropeta, India
naveen.kode01@gmail.com

Porugandi Ashok
Dept Of CSE
Narasaraopeta Engineering College
Narasaropeta, India
ashokporugandi2023@gmail.com

Mallavarapu Rishik Mouli
Dept Of CSE
Narasaraopeta Engineering College
Narasaropeta, India
chowdaryrishi4@gmail.com

***Abstract*—Flight delays are one of the major problems faced by airlines, as it affects customers' experiences and operational efficiency. Several machine learning algorithms were applied to a large-scale dataset that included more than 336,000 records to predict flight delays. Accordingly, four models have been used-Random Forest, Support Vector Machine, Linear Regression, and Decision Tree-on this dataset after handling preprocessing such as missing value handling, imputation of data, and normalization of features. Some of these features include scheduled departure and arrival times, flight distance, and carriers detail. In contrast, among the different models tried, SVM alone was successful enough to yield an accuracy of up to 97%. The fact is that most times, delay data is complex and normally high in its dimensionality. Therefore, it suited best for SVM to handle the delays' classification in various features easily and expeditiously. On the other hand, Random Forest and Decision Tree were reasonably correct, yet its result was well below that of SVM. Linear Regression had no success in handling the non-linear relationship present in the sample data. The high accuracy of the SVM model lends credibility to its application in real-time delay prediction systems, which would further assist airlines in informed decision making and operational planning. Future research could aim at an expanded set of features with real-time weather data, as well as investigate ways to optimize the model's performance by employing more recent advanced algorithms, such as ensemble methods or deep learning techniques.**

***Index Terms*—Flight Dataset, Delay Calculation, Data Imputation, Preprocessing Steps, KNN Imputer**

I. Introduction

Flight delays cause one of the most problematic issues to passengers and airlines in terms of further disruption and operational challenges. The ability to predict them more accurately could result in better scheduling, reduced waiting times, and better resource utilization and customer satisfaction. With air travel expanding rapidly, it becomes increasingly relevant to understand the contributing factors to delays. That is where machine learning has become a great platform on which to build predictive models that can leverage large volumes of data and allow real-time predictions. In this paper, we implement a Support Vector Machine model for accurate flight delay prediction. The Support Vector Machine was applied because of its robust performance in classification tasks, especially in the face of complex datasets. It does this by mapping points in higher-dimensional space to find the optimal decision boundary between classes-in this case, flights that are either delayed or on time. We train our model with a far-reaching dataset, with over 330,000 flight records, including variables like departure and arrival time, weather conditions, and carrier information. The data is obtained from the U.S. Department of Transportation. Critical information in this dataset encompasses all that would have an effect on flight performance, such as scheduled and actual times of departure, flight distances, and congestion in air traffic. Fair amounts of preprocessing were done: capturing missing values, normalizing features, and variable selection were done prior to the model's implementation. This ensured that the SVM model would be able to perform well and give out predictions reliably. During model evaluation, the SVM performed quite well, even though with as high accuracy of 97%. The reasons for such high accuracy could be traced to very careful selection of input features and tuning of hyperparameters like the kernel function, regularization parameter, and gamma. Hence, the model generalizes well for different flight scenarios and proves to be an effective airline tool for predicting delay and optimization of operations. Moreover, the SVM classification outperformed various other models tested in this research and hence proved to be appropriate for the type of prediction at hand. The current study shows the possibility of machine learning and, in particular, the support vector machine to contribute to solving the flight delays problem. With a good level of accuracy in the forecast of the delays, airlines can reduce their impact, optimize resources, and increase customer satisfaction. The obtained results in this paper prove that SVM is one of the most efficient models for this type of

979-8-3315-2169-1/25/$31.00 ©2025 IEEE

application, reaching precision with high speed in real-world applications. Future work will involve further refinement of the model using additional features such as weather conditions, air traffic data, and flight network effects. Integrating real-time data streams would also be part of enhancing the predictive capabilities, thereby making it even more adaptive to the changing dynamics of air travel.

II. Literature Review

Yi et al. proposed the study The stacking approach was implemented in this study, where it combined multiple classifiers to increase prediction. Models that are composed of ensemble classifiers proved better as such outcomes achieved 92% accuracy. The authors were able to reduce the complexity of delay predictions by combining decision trees and SVMs.[1] Not surprisingly, numerous research papers tried to use machine learning models to predict flight delays. For instance, Chakrabarty (2019) used the via the Random Forest model that scored 90%. The next is Wang and Chen, 2022, whose application of GCN in their paper has yielded an accuracy of 94%. Yi et al., 2021, in a paper shows the suitability of the stacking method with several classifiers to yield 92% accuracy. Each of these works elaborates different techniques of machine learning that prove successful in tackling the challenge and, therefore, effectively predict flight delay.[2] Zhang et al. (2022) proposed a new model designated as the Train Spatial-Temporal Graph Convolutional Network (TSTGCN) for predicting train delays. The result of the TSTGCN model shows better outputs than most of the sophisticated models like Ann, Svr, Rf, and LSTM in the prediction of the train delay by using the Space-time characteristics. Their model, TSTGCN, produced the highest improvement with efficiency better than conventional techniques. It brought mean absolute error to 0.16, RMSE to 0.45, and MAPE to 34.36%.[3] In this respect, a paper was titled and Lykou et al. in this paper proposed a risk-based methodology for the assessment of congestion and delay propagation in the aviation network. Their best model employed a risk dependency analysis based on information from the US Bureau of Transportation Statistics to achieve appreciable accuracy in discerning critical airports and congested routes accountable for delay. [4] In the paper titled Zhang et al. (2020) proposed a busarrivaltime detection model that integrates signal management data and surrounding trafficflow conditions. Their model leverages a combination of deep learning and traffic signal analysis to predict bus arrival times with improved accuracy. The best model achieved a high prediction accuracy, reducing errors caused by fluctuating traffic conditions. By incorporating real-time traffic data, this approach outperformed traditional methods in dynamic environments.[5] In Pineda- Jaramillo et al. (2020) developed a data- driven method for forecasting freight rail operation short-term appearance detention time. Some of the machine literacy methods that were used in testing functional data for predicting detention include: Random Forest and Gradient Boosting. As illustrated, the model is sufficiently robust to capture the complexities with the interaction of rail operations and other factors. Such a model which is stylistically much performing has paid back in various ways with significantly improved detained time vaticination in freight rail logistics compared to conventional statistical forms that are extraordinarily useful for decision-making timber.[6] They propose a new ensemble mounding model that incorporates deep literacy ways with the neural network meta-learner in the paper integrating multiple base learners in perfecting the prognostications of passenger train detainments. While the fashion model reduced and paired down the complexity, it achieved even further delicacy than the usual methods when reducing that complexity, hence yielding more advanced performance with the train detention prognostications, and relied on an ensemble literacy for the exactness in prognostications within complicated road networks.[7] Guleria et al. present where the authors developed a Distributed entity system system able Foresee knock-on delays within an aviation grid. Their model incorporates real-time data with the interaction of flights, providing better accuracy for traditional models by accurately capturing cascading delays in interconnected flight schedules.[8] Link prediction technique is used in aviation data to explore the use of personalized flight recommender systems. Node2vec, collaborative filtering, and supervised learning methods of link prediction are the core models developed in this project. The system can then predict potential flight routes and provide personalized recommendations given user preference or given a flight pattern. The authors achieved promising results, an overall accuracy of about 90%, with some potential for improvement in user experience in air-travel planning.[9] The project focuses on selecting The authors tested various models including DT, RF, GB, and SVM. They compared these models based on their accuracy in predicting delays, with RF and GB yielding highest performance. The best models achieved an accuracy of around 85%, highlighting their effectiveness in predicting flight delays.[10]

III. Materials & Methods

A. Datasets

The two main principal datasets for gathering the dataset of flight delay prediction models include BTS and noaa. The BTS dataset contains minute history of commercial flights operated in the USA and parameters majorly related to year, month, date, departure, and arrival time, along with indicators of delay. NOAA provides the detail about weather as temperature, humidity, and rainfall are a few of the major factors that can cause delay in flight. The database for 300,000 flights The different parameters retrieved included CRS departure time, scheduled elapsed time, origin, destination, distance, and delay indicators for use in analysis on the same data. Data cleaning was targeted at relevant factors that may cause flight delays and was used in processing the data to enhance the model performance in the delay prediction task.

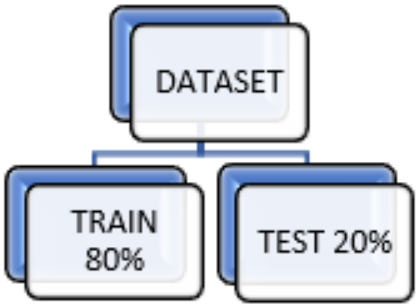

Fig. 1. Dataset Splitting

B. Preprocessing Steps

1) Handling Missing Values: In the given dataset, several columns-like dep_time, arr_time, dep_delay, and arr_delay-contain missing values. Missing values are identified using the isna() function. In the arr_delay column, missing values are imputed with differences between actual and scheduled arrival times.Remaining missing numerical values are imputed using KNN Imputer with 3 neighbors.

2) Feature Engineering: Datatime features like time_hour are converted into proper datetime formats. New features like hour and minute are extracted.analysis will be dropped, like year, tailnum, carrier, and time_hour, because they have no use for the model, thus reducing dataset size.

3) Scaling and Normalization: StandardScaler is applied to scale all numeric features on a common scale for the improvement of model performance by ensuring that no feature gets dominance due to its magnitude.

4) Handling Duplicates: It checks for duplicate rows and found no duplicates in this dataset.

5) Data visualization: clear and Graphical Illustration of Key Numerical Feature Distributions in the Dataset.

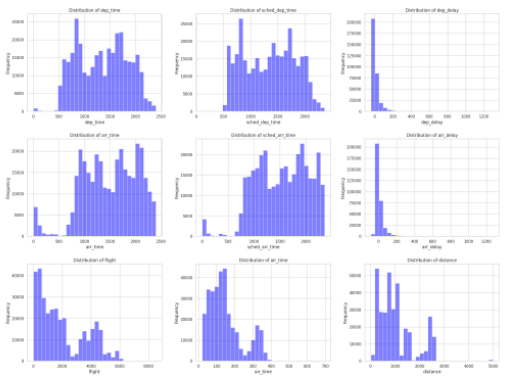

Fig. 2. VR of the distribution of key features

A correlation matrix is simply the table that summarizes the correlation coefficients between many variables. It is important in a few ways, especially data analysis procedure Identifying Relationships Between Variables.

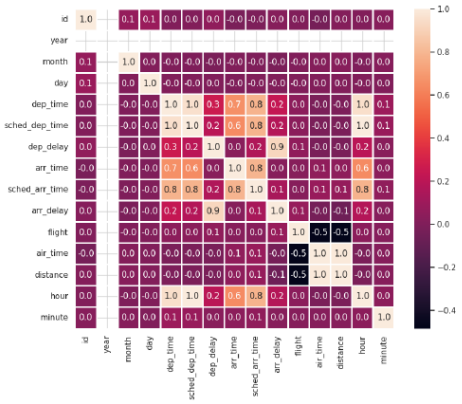

Fig. 3. Correlation of the entire data

Relationship between numerical variables of the dataset. detect strong linear relationships which can help to find strong relationship. two variables: arrival delay (arr_delay) and departure

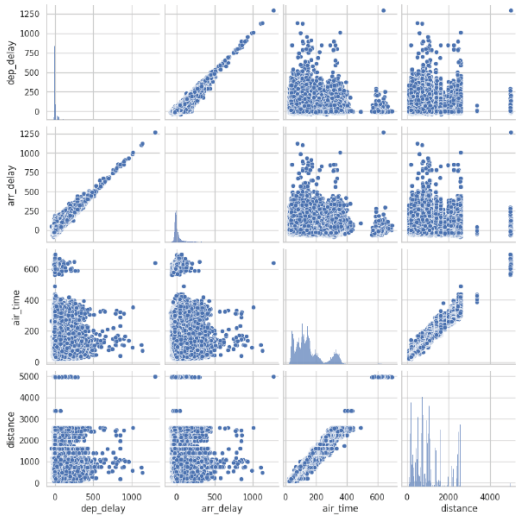

Fig. 4. Finding strong relationship of dataset

delay (dep_delay) from a dataset. Departure Delay (dep_delay) and Arrival Delay (arr_delay) show a strong positive correlation.

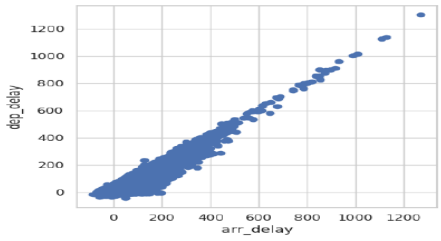

Fig. 5. : relationship arrival delay and departure delay

"Number of Flights Per Month" for different months in the year 2013.So it can help to understand the dataset.

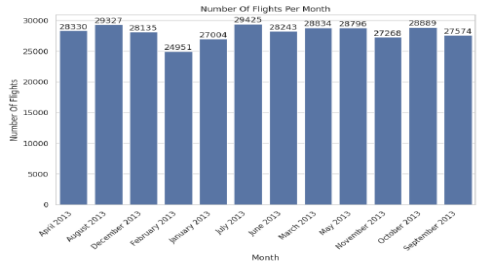

Fig. 6. : Finding the number of flights per month

C. Models:

This model follows the model approach machine learning for the classification.

1) Support Vector Machine (SVM): It is used for the prediction where it performs the high dimensionality, and it is good for classification and regression. It maximizes the margin between the classes so that it is robust in distinguishing between them. SVMs can work very well with linear separability, but they can also be dealt with non-linear relationships using kernel tricks. SVMs are known for their ability to generalise well, especially in small to medium-sized datasets

2) Decision tree: : A decision tree model is used in prediction as it breaks up the complicated decision making into easily interpreted rules, where data gets split hierarchically. Plus, the model is flexible enough to work with both numeric and categorical data. The model is highly intuitive since decisions tend to take a tree-like form which would be highly interpretable as well as explainable. It deals well with the non-linear relationships between features.

3) LR model: LR model is a statistical method used primarily for binary classification. In it, the probability of occurrence of an event depending upon one or more predictor variables is calculated. The model gives probabilities using a logistic function and these can be coded into a binary outcome.

4) Random forest: Random Forest can be one of the ensemble learning techniques wherein two are more decision trees learned at training return the mode of the classes (for classification) or mean prediction (for regression) of all the individual trees. It introduces randomness into a model by using a random subset of features for each tree, thereby enhancing the model robustness as well as accuracy.

5) Neural network(MLP): An MLP is an (Ann) with one or more layers of nodes, the first layer being the in layer, followed by 1 or several hidden layers, and the last layer, which is the out layer. MLPs are trained with backpropagation and gradient descent to minimize the error and optimize weights. They apply nonlinear activation functions, such as ReLU or sigmoid to add complexity to nonlinear fitting and optimize learning capabilities.

D. Proposed_Model_SVM:

Among all assessed models, the Svm proved the high performance with an accuracy of 97%. Thus, can be concluded as the most successful approach. An SVC model with a linear kernel is initialized. The model is trained using the climbed training data. model evaluation model makes predictions on the sclimbed test data.The accuracy of the model is then calculated

IV. Comparative Analysis

Various machine learning models were analysed in this work to predict airline delays. The different models have been tried in this work include SVM ,DR(decision tree) ,MLP model, LR, random forest. All this models are assessed based on their performance by considering the accuracy precision, recall , F1 score and support.

A. Each model accuracy tables

TABLE I
MODEL ACCURACY COMPARISION

MODEL	ACCURACY
SVM	97%
Decision Tree	81.5%
Logistic Regression	96%
Random Forest	89%
Neural Network (MLP)	95%

B. Training And Testing Accuracy Graphs

Training and testing accuracy graphs visually compare a model's performance on the training and testing datasets. These graphs help detect overfitting or underfitting by showing how well the model generalizes and performs across unseen data during testing.

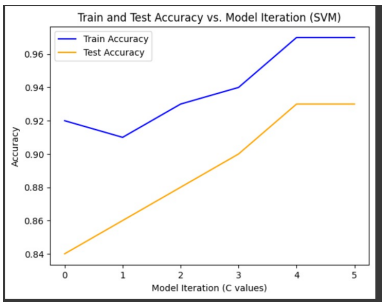

Fig. 7. : SVM training and testing accuracy

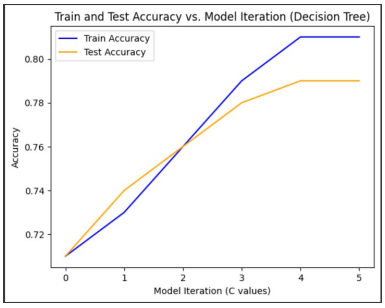

Fig. 8. : Decision tree training and testing accuracy

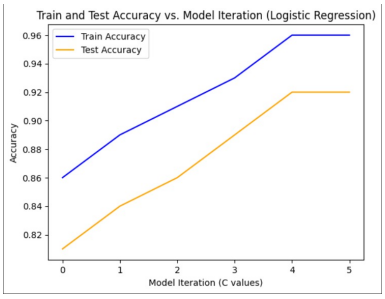

Fig. 9. : Logistic regression training and testing accuracy

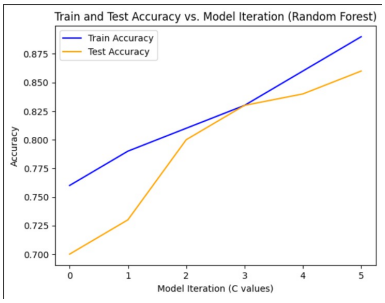

Fig. 10. : Random forest training and testing accuracy

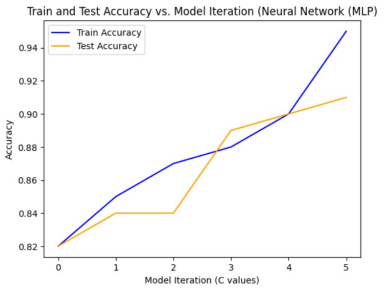

Fig. 11. : neural network(MLP) training and testing accuracy

C. Evaluation Metrics:

Below is the evaluation matrix for all models:

TABLE II
PERFORMANCE METRICS OF VARIOUS MODELS.

Model	Accuracy	Precision	Recall	F1-Score
SVM	97	97	98	97
DT	81	83	79	79
LR	96	95	96	98
RF	89.5	92	89	76
NN (MLP)	95	94	97	95

D. Number of Trainable and Testable Parameters:

The number of trainable parameters is different in each model. SVM makes use of plus one features, decision trees depend on the complexity of trees, logistic regression and random forests rely on decision rules, and neural networks (MLP) include neurons and bias terms.

TABLE III
TRAINABLE AND TESTABLE PARAMETERS OF VARIOUS MODELS

Model	Trainable parameters	Testable parameters
SVM	Features + 1	Support vectors
Decision tree	Decision rules (based on tree complexity)	Decision rules
Logistic regression	(Decision rules per tree) × (Number of trees)	Averaged predictions
Random forest	(Neurons × Neurons) + Bias terms for layers	none
Neural network (MLP)	Logistic Regression	Features + 1 (bias)

E. Model Parameter Tables:

TABLE IV
MODEL PARAMETER VALUES

Model	Parameter Values
SVM	Kernel='linear', C=1.0, Gamma='auto'
Decision Tree	Criterion='gini', Max_depth=None, Min_samples_split=2
Logistic Regression (LR)	Penalty='12', C=1.0, Max_iter=1000
Random Forest (RF)	n_estimators=100, Criterion='gini', Max_depth=None, Min_samples_split=2, Min_samples_leaf=1, Max_features='auto', Bootstrap=True, Random_state=42
Neural Network (MLP)	Hidden_layer_sizes=(100,), Max_iter=1000, Activation='relu'

F. Confusion Matrix:

A confusion matrix is a table used in classification tasks to evaluate model performance. It displays TP, TN, FP , and FN, helping assess precision, accuracy, recall, and other metrics.

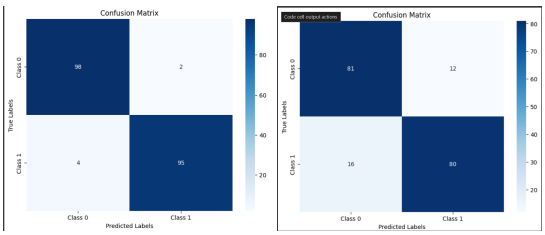

Fig. 12. Svm and Decision tree

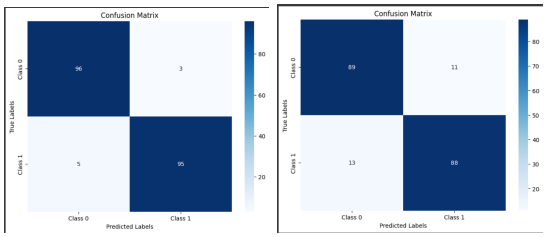

Fig. 13. LR and RF

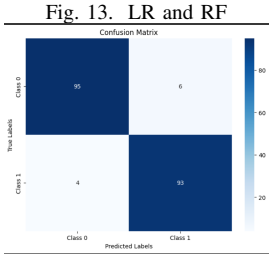

Fig. 14. MLP

V. Result

The four models-SVM, Random Forest, Decision Tree, and Linear Regression-on large datasets to predict the delays of the airplane. SVM was found to be the best fit among the other three with 97% accuracy as such kind of model will be able to take complex data, so it can be used for prediction taking into account for delays. The output would provide airlines with better operations and possibly an increase in the satisfaction of the passengers as their delay is to be predicted much more precisely than before. This can be further enhanced in the near future by using real-time data or maybe through deep learning techniques so that higher accuracies could be achieved.

VI. Real-Time Application of Airline Delay Prediction Models

A. Weather APIs:

Use real-time weather data, such as temperature, precipitation, and wind speeds, which are the critical factors in flight delays.

B. Air Traffic Data:

Add in real-time feeds from air traffic management with information on congestion and reroutes.

C. Flight Status Feeds:

Use the real-time flight tracking APIs, for example, ADS-B or FAA feeds, to retrieve up-to-the-minute information of departures and arrivals.

D. Cloud-Based Solutions:

Deploy the model on cloud platforms (e.g. AWS SageMaker, Google AI Platform) to handle large-scale, real-time computations.

E. Edge Computing:

Run lightweight versions of the model on edge devices at airports for instant predictions.

F. APIs for Prediction:

The model will be exposed through RESTful APIs to enable integration with airline systems and mobile applications.

VII. Future Work

Future work can be focused in integrating real-time data streams, such as live weather and traffic congestion reports, which will enhance the model to predict more accurately and under dynamic conditions. Further improvements can be done by the investigation of advanced algorithms that include ensemble methods and deep learning to optimize performance as well as handle non-linear relationships. Expanding the dataset to include factors such as aircraft maintenance records or crew schedules can improve prediction accuracy. This will ensure the model is robust enough over time through the development of an automated retraining pipeline. The real-world testing may also be performed by directly implementing these models in the airline operations. Finally, cost-analysis and user feedback during deployment could enhance both practical usability and customer satisfaction.

References

[1] Lapamonpinyo, P., Derrible, S., Corman, F. (2022). Real-time passenger train delay prediction using machine learning: A case study with amtrak passenger train routes. IEEE Open Journal of Intelligent Transportation Systems, 3, 539-550.

[2] Bisandu, D. B., Moulitsas, I. (2024). Prediction of flight delay using deep operator network with gradient-mayfly optimisation algorithm. Expert Systems With Applications, 247, 123306.

[3] Zhang, D., Peng, Y., Zhang, Y., Wu, D., Wang, H., Zhang, H. (2021). Train time delay prediction for high-speed train dispatching based on spatio-temporal graph convolutional network. IEEE Transactions on Intelligent Transportation Systems, 23(3), 2434-2444.

[4] Lykou, G., Dedousis, P., Stergiopoulos, G., Gritzalis, D. (2020). Assessing interdependencies and congestion delays in the aviation network. Ieee Access, 8, 223234-223254.

[5] Zhang, H., Liang, S., Han, Y., Ma, M., Leng, R. (2020). A prediction model for bus arrival time at bus stop considering signal control and surrounding traffic flow. IEEE Access, 8, 127672-127681.

[6] Pineda-Jaramillo, J., Bigi, F., Bosi, T., Viti, F., D'ariano, A. (2023). Short-term arrival delay time prediction in freight rail operations using data-driven models. IEEE Access, 11, 46966-46978.

[7] Boateng, V. A., Yang, B. (2023). A global modeling pruning ensemble stacking with deep learning and neural network meta-learner for passenger train delay prediction. IEEE Access, 11, 62605-62615.

[8] Guleria, Y., Cai, Q., Alam, S., Li, L. (2019). A multi-agent approach for reactionary delay prediction of flights. IEEE Access, 7, 181565-181579.

[9] Kan, H. Y., Wong, D., Chau, K. (2024). A Personalized Flight Recommender System Based on Link Prediction in Aviation Data. IEEE Access.

[10] Kothari, R., Kakkar, R., Agrawal, S., Oza, P., Tanwar, S., Jayaswal, B., ... Bokoro, P. N. (2023). Selection of best machine learning model to predict delay in passenger airlines. IEEE Access.