Flight Delay Prediction for Indian Air Carriers with Explainable Artificial Intelligence

Jyoti Singh
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
jyoti.ba07@race.reva.edu.in

Mithun Dolthody Jayaprakash
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
mithun.dj@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence, REVA University
Bengaluru, India
rashmi.agarwal@reva.edu.in

***Abstract*—The aviation industry plays a crucial role in the world's transportation sector, and a lot of businesses rely on various airlines to connect them with other parts of the world. Flight delays are gradually increasing and bring more financial difficulties and customer dissatisfaction to airline companies. The present research takes into consideration the flight data for various Indian carriers, Airport scores, Airlines Ratings, and weather information to build a robust prediction system. This research is done in three phases. In the first phase, Classification models like Logistic Regression, Decision Tree, Decision Jungle, and Neural Network Classifier are used to predict if the flight is delayed more than 15 minutes. The most effective model is chosen, and the primary causes of the flight delay are identified using the model. These factors include both controllable and uncontrollable factors resulting in the flight delay. In the second phase, Regression models like Linear Regression, Linear Regression with Principal Component Analysis (PCA) technique, and Neural Network Regressor model are used to predict the time by which the flight is getting delayed. The third phase of the research focuses on the trips with significant delays in the chosen dataset, and actionable insights are identified and explained for each trip using the Explainable Artificial Intelligence (XAI) tool with the Local Interpretable Model-agnostic Explanations (LIME) package.**

The Neural Network Classifier is identified as the best non-linear model for the dataset under consideration, with an accuracy of 92.5% and a precision of 73.6% for predicting delayed flights. With the lowest RMSE of 3.52, MAE of 2.20, and maximum coefficient of determination(R square) of 93.44, the Linear Regression model with Principal Components stands out to be the best model for predicting delay in flight arrival time.

Keywords— Flight Delay, Indian Air Carriers, Supervised Classification Model, Supervised Regression Model, Logistic Regression, Linear Regression, Decision Jungle, PCA, Linear Regression, Neural Network, MLP Regressor, LIME, XAI.

INTRODUCTION

India's aviation sector has become one of the fastest-growing industries in the world. By offering a competitive and affordable alternative to long and tedious journeys via road or rail, the sector had established itself as a credible and affordable option. A recent study reveals that India is the world's third-largest civil aviation market. According to estimates, India will become one of the world's largest aviation markets by 2034 because of its visible growth trend. By 2037, the number of flight passengers could double to 8.2 billion, according to IATA (International Air Transport Association) [1].

Approximately 55 percent of the segment's market share was held by the passenger carrier IndiGo in the financial year 2021. IndiGo achieved the highest On-Time Performance (OTP) of 95.4 percent at four metro airports in February 2022, according to data released by the Director General of Civil Aviation (DGCA). The four airports where Indigo had the best OTP were Bengaluru (BLR), Delhi (DEL), Hyderabad (HYD), and Mumbai (BOM). Also, IndiGo had the lowest number of complaints from its domestic customers, assuring the carrier's popularity was high.

Delays in flights are not only disruptive to passengers but also costly for carriers. According to the Federal Aviation Administration (FAA), a flight is considered delayed if it is 15 minutes late than its scheduled time, whereas the flight is canceled when it is not operated at all due to some circumstances. Several factors contribute to flight delays, such as bad weather conditions, airport congestion, airspace congestion, mechanical problems, and sometimes airlines using smaller aircraft. It is often the result of these delays and cancellations that the airlines' reputation is tarnished, resulting in a drop in demand from passengers. The inefficiency of the air transportation system may have an indirect effect, increasing business costs as a result of hiring more ground staff and employees. The environment is also greatly impacted by flight delays since they result in higher than projected fuel consumption and carbon emissions. Most of the time, the fuel burn indices provided in the International Civil Aviation Organization (ICAO) engine emissions databank are used to compute the fuel usage for taxi-out. Fuel efficiency can be increased while reducing flight delays, thereby decreasing carbon emissions [2].

The primary goal of the research is to forecast flight delays for selected Indian airports and airlines. The important element of the business objective is also to identify the primary causes of delay for a specific airline and/or airport. The research goals are accomplished in three phases:

Phase I of the research aims to forecast flight delays for Indian air carriers using the Supervised Classification technique and highlight the key drivers affecting the delay. If the flight is delayed for more than 15 minutes, Machine Learning methods like Logistic Regression, Decision Tree, Decision Jungle, and Neural Network classifier are used to anticipate the delay. Additionally, this phase identifies the numerous main factors - both controllable and uncontrollable - that globally contribute to flight delays.

Phase II of the research uses the Supervised Regression technique to forecast flight arrival time delays. Regression Machine Learning models like Linear regression, PCA

technique, and Neural network Regressor model are used to predict flight arrival time delays.

Phase III of the research focuses on the journeys that have major delays and offers actionable, real-time insights into the key factors that contribute to the delays. This is achieved using the XAI tool with the LIME package. LIME enables a machine learning model to get more insights and makes each prediction understandable to pinpoint the primary causes of the delay for an individual instance.

LITERATURE REVIEW

Researchers and analysts have focused their efforts on gathering information on weather and flights to predict the causes of flight delays. Abdel-Aty and C. Lee [3] examined the Orlando International Airport's non-stop domestic flight arrival delay patterns. They mainly concentrated on the cyclical fluctuations in the demand for air travel and the weather at that specific airport. The consequences of these cyclical changes in flight delays were detected in the airport's weather and the demand for air travel.

Airlines, airports, and passengers are all affected by flight delays. All players in the commercial aviation industry rely on their predictions to make decisions. As a result of the complexity of air transportation systems, the number of ways to predict delays, and the overabundance of flight data, it became difficult to develop accurate predictions for flight delays. Based on the scope, data, and computational methods, A. Sternberg and J. Soares [4] present a taxonomy and summarize the different approaches taken to predict flight delays, with a particular focus on the use of machine learning. Furthermore, it also presents a timeline of significant works illustrating the relationship between flight delay prediction problems and research trends to address them.

The objective of the study by S. Ahmadbeygi and A. Cohn [5] is to offer computational findings based on data from a big U.S. carrier that demonstrate how appreciable operational performance gains can be made without raising anticipated expenses. In this study, the authors demonstrate how delay propagation can be minimized by allocating planning slack, and modifying the flight schedule just slightly while maintaining the original fleeting and crew scheduling choices.

By observing the weather conditions, A. A. Simmons's [6] mining technique predicts aircraft delays. He chose various classifiers, compared their performance, and then utilized WEKA and R to build their models. He has utilized many machine learning methods, such as Linear Discriminant Analysis Classifiers and Naive Bayes.

Maryam Farshchian Yazdi, Seyed Reza Kamel [7] have used Deep Learning (DL) based strategy for forecasting flight delay. DL is one of the most recent approaches used to address issues with high levels of complexity and vast amounts of data. Additionally, DL can automatically extract the key features from data. To find the right weight and bias values, the Levenberg-Marquart algorithm is used, and lastly, the output has been optimized to deliver very accurate findings.

The goal of research by Sun Choi and Young Jin Kim [8] was to mitigate the consequences of data imbalance brought on by data training. For anticipating specific flight delays, they have employed methods including Decision Trees, AdaBoost, and K-Nearest Neighbours. The model generated a binary classification to forecast the planned flight delay.

The study by Peng Hu and Jianping Zhang [9] examines the distribution of delayed, on-time, and early arrivals flights by analyzing the departure flight data from Guangzhou Baiyun International Airport in June 2020 and selecting the data from ten landing airports. It creates a random forest predictions model and analyses the selection of variables that have an impact on flight delays. It also analyses the significance of features such as the departure flight delay time, the scheduled flight time, the number of scheduled departure flights on the day, the date, and the landing airport.

Analysis of people's opinions, sentiments, and behavior on flight delays has been done by B. L. Lei Zhang [10] using opinion mining and sentiment analysis. A feature-based opinion summary also referred to as sentiment classification, is the study's output.

Researchers developed analysis algorithms that assisted them in extracting characteristics from the model using methods including Natural Language Processing, Naive Bayes, and Support Vector Machine. The majority of them concentrated on foreseeing average flight delays.

This research primarily focuses on forecasting flight delays and arrival time delays for Indian air carriers in phases on the designated routes. Additionally, the primary goal is to obtain practical insights on the major factors causing flight delays for the most troublesome or delayed flights using the XAI technique.

METHODOLOGY

Data Management

The BeautifulSoup Python package was used by online collaborators working on similar projects to collect historical flight information and weather details from *flightradar24.com* and *accuweather.com*, respectively. The flight information dataset consists of 14951 observations and 22 features during a period of two years, from January 2018 to January 2020. The weather dataset includes 14 meteorological features and 90600 observations spread throughout two years, from January 2018 to January 2020. The features are categorized based on their type as Airport scores (Departure and Arrival airport), Airline Ratings, Flight Departure and Arrival details, and Weather details (Departure and Arrival airport).

Data leakage in machine learning occurs when the data used to train an algorithm contains the information that the model is attempting to predict. As a result, the predictions made by the model after deployment are unreliable and inaccurate. Departure Time, Departure Delay, and Arrival Time leak information about Arrival Delay in the dataset. Therefore, these columns are removed from the dataset before feeding it as input to the classification and regression models.

Multicollinearity occurs when the independent variables are highly correlated. It can be difficult to fit the model and comprehend the findings if there is a high correlation between the variables. To address the multicollinearity issue, we created a correlation matrix and eliminated 5 variables that had a degree of correlation greater than 0.90.

Outliers are data points in a distribution that deviate from the overall pattern. Using the boxplot, it was deduced that the

Arrival Delay column contains outliers. Interquartile Range (IQR) is a standard method for identifying outliers. An ordered dataset is divided into 4 equal-sized groups using the quartile. Specifically, IQR refers to the middle 50% of data, which is from *Quartile 3 (Q3) - Quartile 1 (Q1)*. Outliers are defined by the interquartile range approach as values that are larger than *Q3+1.5* IQR* or smaller than *Q1-1.5* IQR*, where *IQR = Q3 – Q1*. In the current dataset, the upper whisker is 34 minutes which means flights are delayed by a maximum of 34 minutes and the lower whisker is -46 minutes which means the flights are reaching 46 minutes before Arrival time.

One Hot encoding technique is used in the research to convert the categorical columns to binary vectors. Also, the research uses MinMax scaling technique to normalize the dataset. The MinMax scaling technique sets the feature's minimal value to zero and its highest value to one. The MinMax Scaler reduces the data inside the specified range, often between 0 and 1. It scales the values to a particular value range while preserving the original distribution's shape.

Modeling

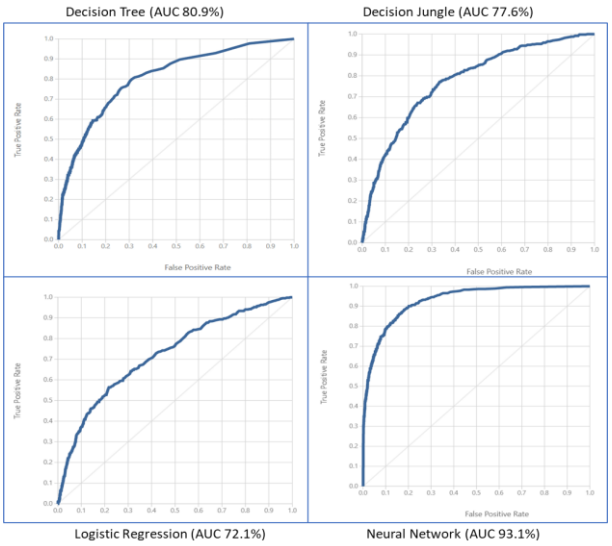
Predict if the flight is delayed by more than 15 minutes using Classification techniques

The Delay attribute is leveraged as a Label for creating Supervised Classification models. It divides the dataset into two categories: 1 for flights that are delayed by at least 15 minutes and 0 for those that are not delayed.

The prepared dataset is divided into two subsets using the Test-Train split methodology in the 30:70 ratio respectively. The Train dataset is used to fit the machine learning model, and the test dataset is used for evaluation purposes.

We have used Decision Tree, Two-Class Decision Jungle, Logistic Regression, and the Neural Network Classifier to predict and classify the data in two classes i.e. delay and no delay.

Fig. 1. depicts the ROC (Receiver Operating Characteristic) curve of all the Classification models with Area under the ROC curve (AUC). Neural Network Classifier tops the chart with an AUC of 93.1%.



ROC curve of Supervised Classification models

The Key Metrics values for the classification models used to predict flight delay are shown in Table I. The Neural Network classifier is the best non-linear model for the dataset under consideration, with an accuracy of 92.5% and a precision of 73.6%.

CLASSIFICATION MODELS KEY METRICS

Model	**Key Metrics**			
	Accuracy	***Precision***	***Recall***	***F1 Score***
Decision Tree	68.50%	26.30%	43.40%	32.75%
Two-class Decision Jungle	88.9%	40.7%	6.8%	11.7%
Logistic Regression	88%	0	0	0
Neural Network classifier	92.5%	73.6%	54.6%	62.7%

The top 5 independent variables influencing flight delays are listed in Table II, with flight Duration topping the list with 13.2% followed by Arrival Airport On-Time Rating with 5.9%.

TOP 5 IMPORTANT VARIABLES

Sr. No.	***Independent variable***	***Importance***
1	Duration	13.2%
2	Arrival_AirportOnTimeRating	5.9%
3	Deprture_AirportOnTimeRating	5.6%
4	Arrival_AirportServiceRating	5.4%
5	Departure_windspeedKmph	4.9%

Predict the delay in flight Arrival time using Regression techniques

Supervised Regression machine learning algorithms are used to predict continuous valued output based on the independent input variables. For the prepared dataset under consideration, Arrival Delay is taken into consideration as the output variable to predict the delay in the flight's arrival time.

We have used the Linear Regression model, Linear Regression with Principal Components, and Regression Neural Network algorithms to predict the delay in the flight arrival time.

PCA: This is a dimensionality-reduction technique, used to reduce the dimensionality of big data sets by condensing a large collection of variables into a smaller set that retains the majority of the large dataset's information.

Covariance Matrix computation is done by PCA to understand how the variables in the input data set differ from the mean to one another or to determine whether there is a link between them.

The covariance matrix, which has entries for all potential pairs of the initial variables, is a $p * p$ symmetric matrix

(where p is the number of dimensions). For instance, the covariance matrix for a collection of three-dimensional data containing the variables x, y, and z is a three-by-three matrix in the form of (1).

$$\begin{bmatrix} Con(x,x) & Con(x,y) & Con(x,z) \\ Con(y,x) & Con(y,y) & Con(y,z) \\ Con(z,x) & Con(z,y) & Con(z,z) \end{bmatrix} \quad (1)$$

To identify the Principal components of the data, eigenvectors and eigenvalues are computed from the covariance matrix. The new variables created as a result of the basic variables' linear combinations or mixtures are known as principal components. These combinations are made in a way that most of the information included in the original variables is condensed or squeezed into the first components, which are the new variables (i.e., principal components), and these are uncorrelated. In the current project, 14 principal components are created using the PCA technique.

These 14 Principal component factors are fed to a Linear Regression model with Arrival Delay as the label to be predicted. Principal components have been found to significantly reduce the error rate of the Linear Regression models when used as input features.

The significance of the top 3 primary component factors is displayed in Table III, with factor 8(Carrier) ranking the highest.

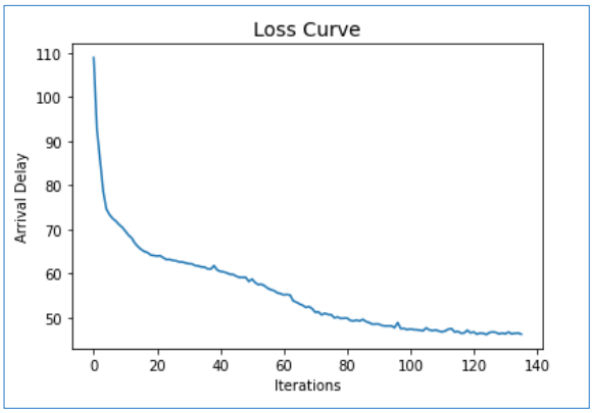
SIGNIFICANCE RATINGS OF THE TOP 3 PRINCIPAL COMPONENT FACTORS

Features	**Factors Significance**		
	Factor 8 (Carrier) 0.663	***Factor 10 (Wind) 0.643***	***Factor 6 (Mistiness) 0.543***
Carrier Rating	0.975		
Vistara	0.954		
Departure Wind Gust		0.88	
Departure Wind speed		0.866	
Departure Cloud cover			0.701
Arrival Dew point			0.631
Departure Humidity			0.614

Multi-layer Perceptron Regressor (MLPRegressor): This research uses Neural Network - MLPRegressor of Scikit Learn library to predict the delay in flight arrival time. MLP Regressor is a type of Artificial Neural Network (ANN). A minimum of three layers of nodes make up the simplest MLP: an input layer, a hidden layer, and an output layer. In the current MLPRegressor, 8 hidden layers are used to build up the model.

The square error serves as the loss function in regression scenarios. "Adam" is the default optimizer and it can handle quite large datasets if any optimizer is not specified. The "sigmoid" and "hyperbolic tan" functions are supported by MLPRegressor in addition to "RELU" activation.

A Loss curve during training is one of the most used graphs for neural network debugging. It provides us with an overview of the training procedure and the way the network learns. Fig. 2. depicts the Loss curve of the MLP Regressor model. The below curve indicates that the model is training reasonably well because the loss is decreasing with each iteration.



Loss curve for MLP Regressor

The Key Metrics values for the regression models used to predict a delay in the flight arrival time are shown in Table IV. With the lowest RMSE and MAE values and maximum coefficient of determination, we conclude that the Linear Regression model with Principal Components is the best model for predicting delay in flight arrival time.

REGRESSION MODELS KEY METRICS

Model	**Key Metrics**		
	Root Mean Squared Error (RMSE)	***Mean Absolute Error (MAE)***	***Co-efficient of Determination (R^2)***
Linear Regression	121.03	40.64	3.43
Regression Neural Network	97.81	7.51	48.91
Linear Regression with PCA	3.52	2.20	93.44

Local justifications for a single incidence of flight delay using the XAI

With huge and diverse data sets, one frequently needs to rely on complicated models to produce the best results. However, their user interpretability is typically poor, necessitating in some circumstances the acceptance of a loss in model performance. Toolkits have been created that show the causes of model predictions to improve the interpretability of even complicated deep learning models. XAI is a set of tools and frameworks that help to understand and interpret predictions made by complex machine learning models.

There are two ways to deliver explanations utilizing the XAI technique: Shapley Additive explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME).

LIME is used in the current research to produce local explanations. LIME, an XAI technique, aids in illuminating a machine learning model and making each prediction's particular implications understandable. The technique is appropriate for local explanations because it describes the classifier for a particular single instance. LIME transforms the input data to produce a succession of artificial data that only partially retains the original features. The LIME explanations are created in just two easy steps: Importing the module and then fitting the explainer with the training values, features, and the target.

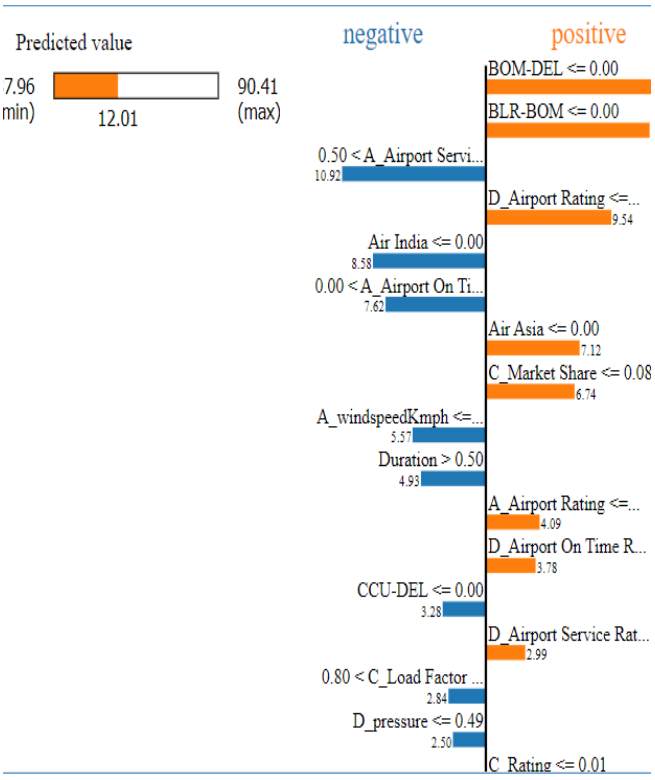
Phase III of the research focuses on the journeys that have major delays and offers actionable, real-time insights into the key factors that contribute to the delays. This is achieved using the XAI tool with the LIME package.

Fig. 3. depicts the LIME explainability for SpiceJet flight from BLR to DEL. The model predicts the delay by 12 minutes and explains the prediction as below:

a) The negative side attributes (the ones shown in Blue) on the left side indicate the attributes contributing to the arrival of the flight before the scheduled arrival time. For example, Arrival Airport (DEL) Service Rating is greater than 0.50, hence flight should reach 10 minutes before time.

b) However, the positive characteristics (the ones shown in Orange) are more substantial and are more responsible for the flight delay. For example, Departure Airport (BLR) rating level is 0, which contributes 9.54 minutes to the delay.

c) On the extreme right, we can observe the real features' value for the instance under consideration.



Local explanations for SpiceJet flight from BLR-DEL

Table V lists the top 5 key controllable and uncontrollable drivers of delay identified by LIME explanations for the SpiceJet flight from BLR-DEL. The features that can be controlled and tackled easily are the Airport Rating, On-Time Performance rating of the Airport, and Airline. Although weather parameters cannot be controlled, Airline can plan to schedule the flight time based on the weather conditions.

TOP 5 IMPORTANT VARIABLES

Sr. No.	***Controllable Features***	***Uncontrollable features***
1	Departure Airport Rating	Route
2	Carrier Market share	Departure precipitation
3	Arrival Airport Rating	Arrival visibility
4	Departure Airport On-Time Performance rating	Departure Cloud cover
5	Carrier On-Time Performance Rating	Arrival precipitation

With this knowledge, airlines and the airport authorities must collaborate to provide better services; otherwise, both customers and airlines risk continuing to lose money and, more importantly, time.

CONCLUSION

Based on the comparative study of the various types of Classification Models' Key performance metrics, the Neural Network Classifier model emerged as the best model with an accuracy of 92.5% and a precision of 73.6% for predicting the delayed flight. While Linear Regression Model with Principal Components is identified as the best Regression model with the lowest RMSE of 3.52, MAE of 2.20, and maximum coefficient of determination(R square) of 93.44 for predicting delay in flight arrival time. The XAI technique can be used by the airport and airline authorities to locally interpret the main causes of the arrival delays for the trip with substantial delays. By controlling these factors, one can ensure that delays are kept to a minimum, lessening the inconvenience for passengers and the expense of taxing and fuel used when aircraft are delayed. Market share, carrier OTP rating, load factor of the airline, OTP ratings of the departure and arrival airports are identified as the most significant controllable features affecting flight delays. The destination Airport's wind speed, precipitation, humidity, and visibility are identified as the most significant uncontrollable features affecting flight delays. Although the uncontrollable feature cannot be controlled, Airline can plan to schedule the flight time based on the weather conditions. The controllable features that can be controlled and tackled easily are the OTP rating of the Airport and Airlines. According to the analysis, customers are more likely to choose the carriers with a higher OTP rating. Even though consumers cannot select the departure and arrival airport based on OTP rating, airport authorities must attempt to improve their services and management to prevent flight delays.

FUTURE SCOPE

Further study and scope extension are possible in the below areas:

a) In-depth analysis of the variables influencing airports' and airlines' OTP ratings

b) The research's forecasts and key driver analyses can be used to develop more real-time applications

c) A broader dataset can be taken into consideration for accurately capturing the impact of controllable and uncontrollable factors

d) The focus of this research is mostly on aircraft and meteorological data for India. The project scope can be extended to other countries like China, the United States, the United Kingdom, Russia, and more. One can broaden the research's reach by including flight information from international flights rather than simply domestic ones

e) The project can be extended to create a prediction model for the departure time delay, as passengers tend to complain more about uncertainty and departure delays

ACKNOWLEDGMENT

We would like to convey a heartfelt thanks to all the mentors at RACE, Dr. J. B. Simha, Mr. Ravi Shukla, and Mr. Ratnakar Pandey for their continuous support throughout the learning journey. We would like to express a special thanks to Dr. Shinu Abhi, Director REVA Academy of Corporate Excellence for her cordial support, valuable guidance, and information at various stages, that helped in completing the research.

REFERENCES

[1] S. S. Statista, "Market share of airlines across India in financial year 2021, by passengers carried," 2021. https://www.statista.com/statistics/575207/air-carrier-india-domestic-market-share/

[2] D. M. M. . Dissanayaka, V. Adikariwattage, and H. R. Pasindu, "Evaluation of Emissions from Delayed Departure Flights at Bandaranaike International Airport (BIA)," vol. 186, no. Apte 2018, pp. 143–146, 2019, doi: 10.2991/apte-18.2019.26.

[3] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak, "Detecting periodic patterns of arrival delay," *J. Air Transp. Manag.*, vol. 13, no. 6, pp. 355–361, Nov. 2007, doi: 10.1016/J.JAIRTRAMAN.2007.06.002.

[4] A. Sternberg, J. Soares, D. Carvalho, and E. Ogasawara, "A Review on Flight Delay Prediction," no. March, 2017, doi: 10.1080/01441647.2020.1861123.

[5] S. Ahmadbeygi, A. Cohn, and M. Lapp, "Decreasing airline delay propagation by re-allocating scheduled slack," *IIE Trans. (Institute Ind. Eng.*, vol. 42, no. 7, pp. 478–489, 2010, doi: 10.1080/07408170903468605.

[6] A. A. Simmons, "Flight Delay Forecast due to Weather Using Data Mining," *Semant. Sholar*, 2015.

[7] M. F. Yazdi, S. R. Kamel, S. J. M. Chabok, and M. Kheirabadi, "Flight delay prediction based on deep learning and Levenberg-Marquart algorithm," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00380-z.

[8] S. B. and D. M. Sun Choi, Young Jin Kim, "Prediction of weather-induced airline delays based on machine learning algorithms," *IEEE*, 2016.

[9] P. Hu, J. Zhang and N. Li, "Research on Flight Delay Prediction Based on Random Forest," 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2021, pp. 506-509, doi: 10.1109/ICCASIT53235.2021.9633476.

[10] B. L. Lei Zhang, "Sentiment Analysis and Opinion Mining," *Encycl. Mach. Learn. Data Min.*, vol. 30, no. May, pp. 503–523, 2012, doi: 10.1007/978-3-319-60435-0_20.