www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC
Gestão, Inovação e Tecnologias

An Approach of Applying Machine Learning Model in Flight Delay Prediction- A Comparative Analysis

Saksham Somani[1]; Priyanshu Pandey [2]; Meghna Sharma[3]; M. Safa[4*]

[1]Final Year UG Students, Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

[1]sakshamsomani@gmail.com

[2]Final Year UG Students, Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

[2]priyanshupandey1309@gmail.com

[3]Final Year UG Students, Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

[3]meghas@srmist.edu.in

[4*]Assistant Professor, Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India.

[4*]safam@srmist.edu.in

Abstract

Flight delay is a major issue in the aviation industry. In commercial aviation, if a flight reaches its destination 15 minutes later than the scheduled arrival, it is said to be delayed. Flight delays cause a great deal of bother to travelers. It could make them late to their booked occasions or miss a corresponding flight, accordingly prompting outrage and dissatisfaction. Likewise, travelers may not generally be entitled for a refund when a postponement happens. Carriers report that couple of the numerous reasons prompting most flight delays are carrier glitches, climate conditions, support issues with the airplane and congestion in air traffic. Rapid development in airline industry has led to an increased number of aircrafts in the skies, this has brought about air-gridlock causing flight delays. Flight delays are not only extremely undesirable financially but also have adverse environmental effects. Air-traffic management is becoming increasingly challenging. The aim of our research work is to predict the delay of flights due to various factors using machine learning and deep learning so as to minimize losses and increase customer satisfaction. This Machine Learning model could be integrated with Airlines systems for the use of staff and customers and it could also rank Airlines and flights based on delays.

Key-words: Flight, Delay, Machine Learning, Algorithms, Naïve Bayes, SVM, Decision Trees.

Received: 26.04.2021 – Accepted: 15.05.2021

1. Introduction

The aviation industry around the globe incurs huge losses due to various factors, one of these factors is Airline Delay. Airline delay tends to be onerous for each entity involved i.e. airports, airlines and passengers. Precise and meticulous prediction of Airline delay using the factors which play prodigious role are going to be the key to attenuate the losses and increase customer satisfaction. In the paper, several machine learning and deep learning algorithms have been employed to produce a comparative study with respect to the accuracy of each algorithm.

The Flight Delay Prediction System consists of the following steps:-

• Data Preprocessing • Data Visualization

• Feature Selection• Data Balancing

The objective is to analyze and predict flight departure delays for a sample of flights and optimize different models to get best possible predictions for the same.

Some of the main goals pertaining to this project can be:-

- Predict the delay of flights due to various reasons.
- Increase customer satisfaction while minimizing loses due to flight delays.
- To create a highly accurate model for flight delay prediction using ML and deep learning algorithms.
- Reduction of the magnitude of Air Traffic Congestion that may add to the delay of the flights.
- To determine how different combinations of feature selections affect the accuracy of the model.

2. Problem with the Existing System

There are various factors causing flight delays including weather conditions, ATC restrictions, mechanical issues etc. These cause companies lots of losses and unwanted expenses. People waste a lot of essential time too. Flight delays are initiated by a variety of aspects including technical issues, poor weather conditions, ATC problems, labor strikes, medical emergencies, congested airports, etc. The Indira Gandhi International Airport, Delhi discovers that the significant reason for postponed flights are by and large traditionalist reasons like boarding/check-in issues, logistical/clerical errors, deferral of airplane from another carrier, awaiting onboard staff, rerouting, diversion, airplane change for non-technical reasons.

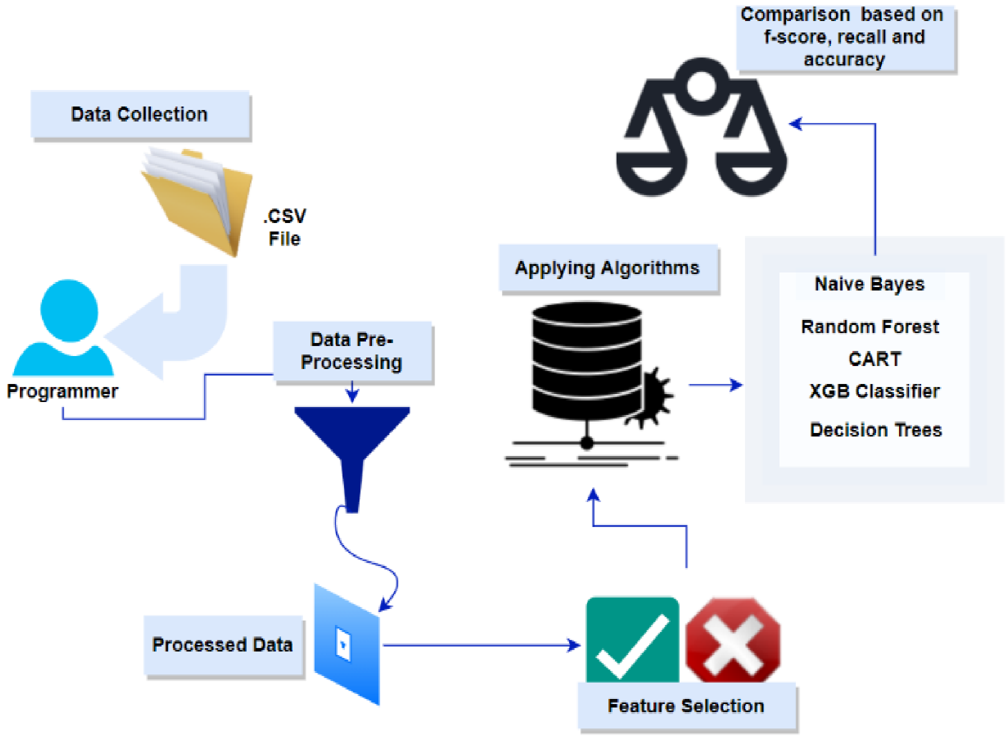
3. Literature Review

- [1] This paper is predicated on a flight delay system that emphasizes on the effects of weather on the delay of a flight.
- [2] Considering the recording machine characteristic of the SVM, a sensitivity analysis was performed to assess the link between dependent and explanatory variables.
- [3] The paper focuses on developing a world optimization version of the Expectation Maximization algorithm, borrowing ideas from Genetic Algorithms.
- [4] A model built on the information sets obtained from each flight delays and weather information sets and a sampling technique is applied to balance the data.
- [5] In this paper, the author checked out completely different cubic centimeter techniques/algorithms to undertake to predict whether an aircraft will be late or not before it's even proclaimed on the departure boards. Thus, this info was checked out as a vicinity of explorative Data Analysis (EDA).
- [6] The scientists of this paper utilized different AI and profound learning strategies like, Random Forest, Support vector Machine, K-means and so on, and considered different components that may cause a flight delay, may it be because of climate or NAS delay. The yield of the investigation was that, random forest gives the best outcomes for the taken dataset.
- [7] The research in this paper was made utilizing the Information Mining and AI methods to foresee delay for a neighborhood carrier by the name American Airlines, for the best 5 busiest air terminals. Procedures like Gradient Boosting Classifier and Hyper-parameter Tuning were utilized in this examination.
- [8] Diverse programming resembles AWS and sparkle were utilized for the application. The issue was characterizing e4as a Multiclass Characterization issue for this investigation.
- [9] Flight postpone forecast can be separated into two fundamental classifications, delay propagation and root deferral and cancellation. Five fundamental techniques were proposed for displaying: statistical analysis, probabilistic model, network representation, operations research, machine learning and AI.
- [10] This study was finished utilizing the Light GBM procedure of AI. Flight postpone expectations are chiefly centered around the air terminal-based information as it were. It can likewise be considered as future extent of study.

- [11] The model uses Memento and Resilient Back Optimized Propagation that the Resilient Back Optimized Propagation spreads speedier than back propagation and therefore the model is prepared and thusly has been expanded.
- [12] A multifaceted methodology, a novel profound conviction network technique is utilized to mine the inward examples of flight delays. Support Vector regression is inserted in the created model to play out an administered calibrating inside the introduced prescient design.

4. Proposed Methodology

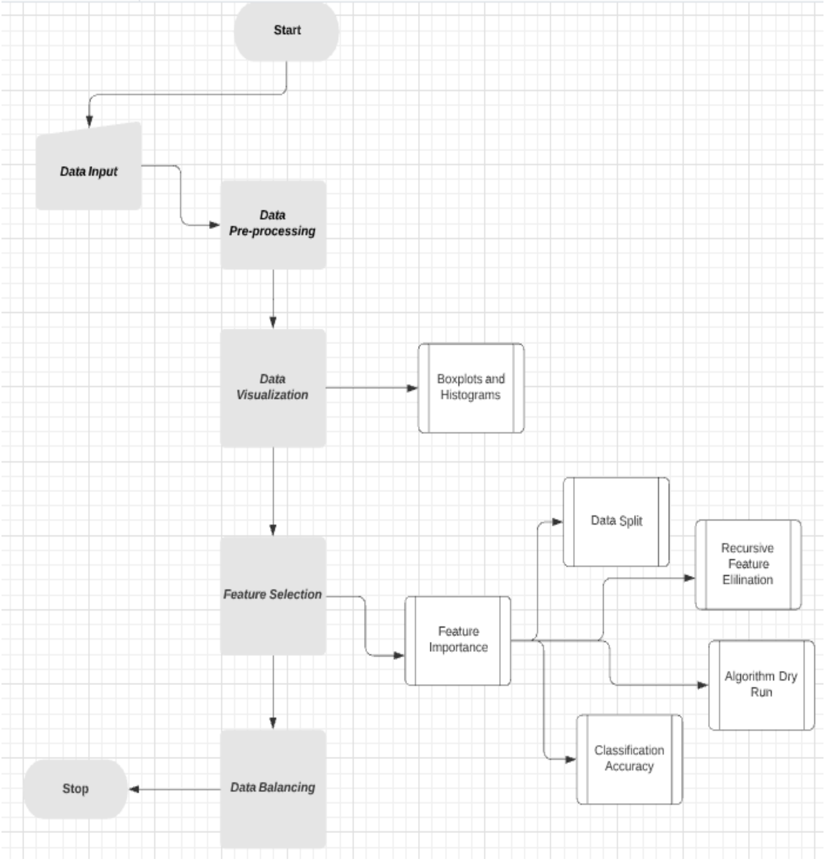
Figure 1 - Proposed Methodology Flow



The first step is Data Collection, the primary and the most crucial step towards building a model. The dataset is collected from the airline website, after which it is pre-processed before being used for the project. Once we have the processed data, a feature selection algorithm (Recursive Feature Elimination) is applied to get the features from the original dataset that would give the best results for our project. After that we split the data into testing and training datasets before applying the various machine learning algorithms to the dataset. Finally, we compare the different algorithms, based on various factors like Accuracy, f1-score etc., to check which algorithm gave the best results for the taken dataset.

5. System Architecture

Figure 2 - System Architecture Diagram



A. Data Preprocessing

It's the very first step that needs to be performed while doing a project. A numerous amount of data is available on the internet nowadays on various topics. People have worked to collect large amounts of data and tried to make it available for the use of others in one single place. One such site is the Kaggle data warehouse. We have also taken our dataset from the Kaggle site itself. The data we are talking about here is a flight delay dataset, from the year 2015. The dataset contains various columns like Year, Date, and Flight number, Delay time and many other columns which we call features. Once the dataset has been chosen that fulfils our requirement, we download the dataset in the form of .CSV file. We import the dataset into our Jupiter notebook. For our project we then selected only 20,000 tuples randomly from the dataset. The next step is to clean the dataset and remove all the null values from it. After this step, we have clean data with which we can proceed with our project.

Received: 26.04.2021 – Accepted: 15.05.2021

B. Data Visualization

Here and there information doesn't bode well until you can take a look at in a visual structure, for example, with diagrams and plots. Having the option to rapidly visualize your datasets for yourself as well as other people is a significant ability both in applied insights and in applied AI.

Figure 3 - Box Plot Data Visualization

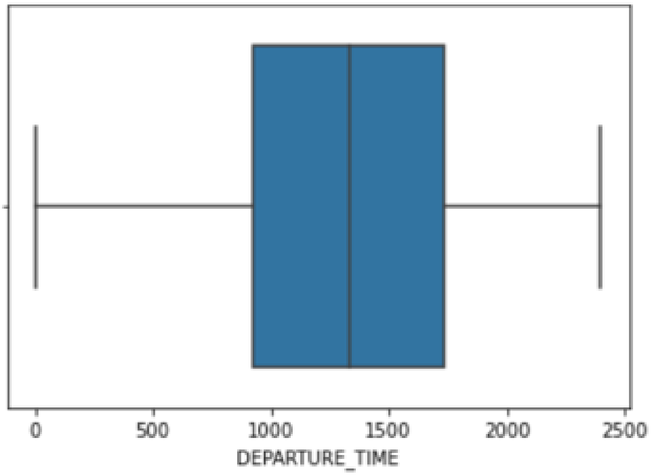
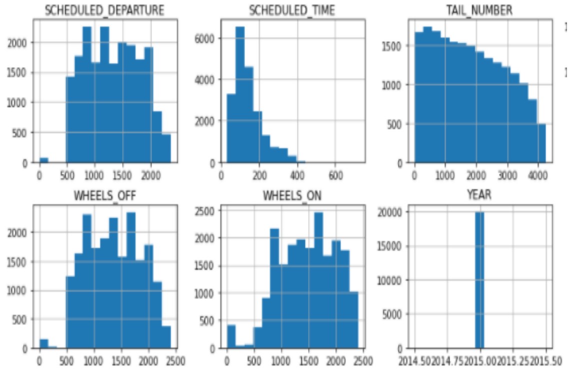


Figure 4 - Data Visualization with Histogram



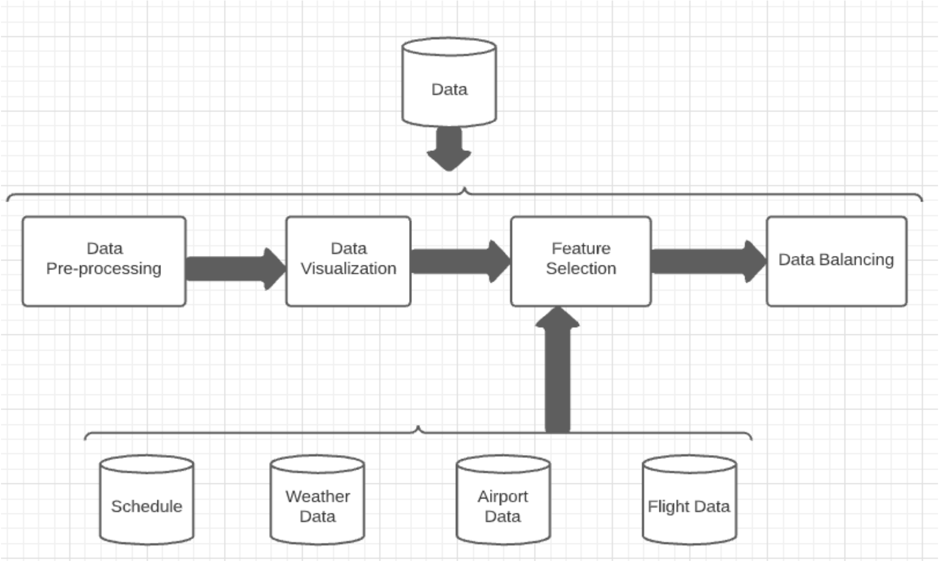
C. Feature Selection

Feature/Attribute selection is the process of selecting only those features which are relevant to the study. In our project, out of the originally available 23 features only 16 were deemed as important for the project, thus the remaining were removed from the dataset. In our project, we have used the Recursive Feature Elimination technique to get the important features for our project. The goal of a recursive feature elimination (RFE) is to select features by recursively considering the smaller sets of features.

D. Data Balancing

Imbalanced data distribution is a common machine learning aspect. It actually means that the distribution of the data is not uniform and that one of the classes is higher than the others. Basically, the total number of observations are not equal for all the classes in a classification dataset. Thus, in order to ensure accuracy, several techniques such as oversampling/under sampling are performed to balance the given set of data.

Figure 5 - Steps towards Data Balancing



6. Methods

Naive Bayes: Running on the basic concept of Bayes theorem, Naive Bayes is a classification technique and it developed by assuming that if a feature is present in a dataset it is independent of the other features present in the same dataset. This algorithm is generally known to outperform even the highly complex machine learning models. Naive Bayes makes it easier to predict the class of test datasets. In cases where the assumption of independence is true, the Naive Bayes algorithm performs much better than other algorithms. This project, the Naive Bayes algorithm calculates the probability of a delayed flight. Once the model has been trained, it is applied to the testing dataset to get the accuracy. Though the Naive Bayes algorithm has been classified as a bad estimator in some cases, it worked just fine for us.

Mathematical Equation:

$$P(c|X) = \frac{P(X|c)P(c)}{P(X)} \quad (1)$$

Received: 26.04.2021 – Accepted: 15.05.2021

Classification and Regression Trees: In CART, the decision trees lay the input vectors to identify values in the leaves. The branches of the trees then split the input values based on the observation values. This process is repeated till the leaves are reached.

Decision trees: Decision trees are a supervised type of learning algorithms in which the data is continuously split according to certain parameters. A decision tree can be best explained with the help of two terms, decision nodes and leaf nodes, the leaf nodes act as the outcomes, whereas the decision nodes are where the data is split. Decision trees are of two types, namely, Classification and Regression. While making a decision tree, different types of questions are asked, at each node of the tree. Based on the questions asked, the information gain can be calculated corresponding to it. The entropy of the dataset after a transformation is compared with that of the same dataset before the transformation, this allows one to calculate the information gain. It is then used to choose a feature which would then be split to further the tree.

XGB Classifier: It's said that if you are having a hard time with predictive modelling, use XGBoost. A highly complex algorithm, powerful enough to overcome various types of irregularities in the data. It is easy to build a model using XGB, whereas it is comparatively difficult to make an already existing model better using the same as this algorithm uses multiple parameters. Parameter tuning is a must to improve the model.

Random Forest: Random Forest is a commonly used ML algorithm that is a part of the other supervised learning algorithms. It is useful for both classification and regression problems. The concept behind this algorithm is ensemble learning, which is a process of solving complex problems and improving the performance of the model. It is a classifier that has multiple decision trees based up different subsets of the given dataset and takes the average to improve the predictive accuracy of the model. In our case, the algorithm mixes different features and forms numerous trees. The average of all the accuracies from the trees is taken to display the final accuracy, and that's how the algorithm works. The greater the number of decision trees, the better the result is.

Mathematical Representation and Equations

In order to develop a tree, the Gini Index needs to be determined and is calculated by the following formula.

$$Gini\ (y, S) = 1 - \Sigma_{cj\ \epsilon\ dom\ y} \left(\frac{|\sigma y = cj\ S|}{|S|}\right) \qquad (2)$$

ISSN: 2237-0722
Vol. 11 No. 3 (2021)
Received: 26.04.2021 – Accepted: 15.05.2021

7. Results

Figure 6 - Accuracy Comparison of different Models

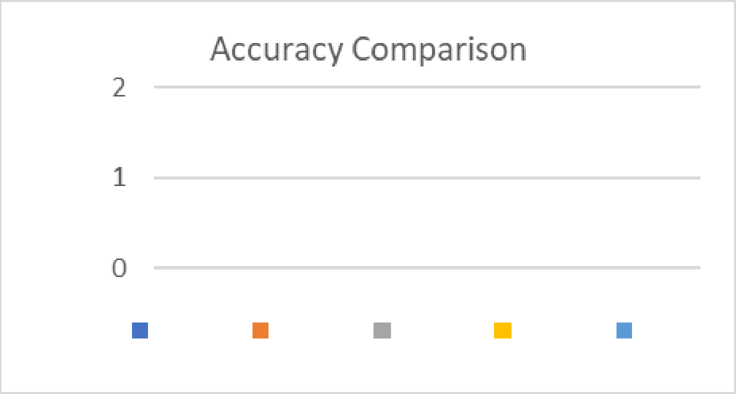


Figure 7 - An Overall Comparison

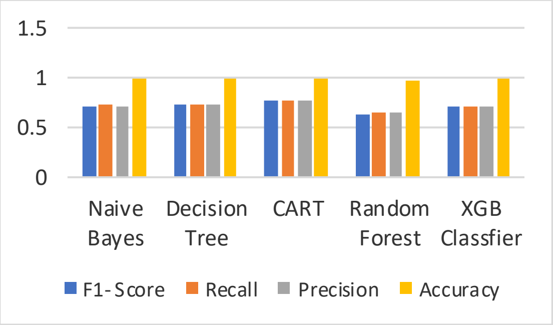


Table 1 - Overview of Results

	F1-score	Recall	Precision	Accuracy
Naive Bayes	0.709	0.722	0.705	98.65%
Decision Tree	0.721	0.723	0.720	98.90%
CART	0.767	0.771	0.765	99.15%
Random Forest	0.636	0.647	0.645	97.24%
XGB Classifier	0.705	0.700	0.717	98.50%

Naïve Bayes, Random Forest, CART, Decision Trees, XGB Classifier algorithms are applied on a dataset using Python programming language, and their accuracies were determined based on characteristics such as Origin, Destination, Departure_Delay, Arrival_Delay, Late_Aircraft_Delay, Month, etc.

Delay Classification: Parameters like Origin, Destination, Arrival_Delay, Departure_Delay our used to train the model against Arrival_Delay attribute of the dataset. Thus, performing different algorithms on the dataset, we obtain – No. of Training Samples = 15962; No. of Testing Samples = 3991. The model performance is evaluated on the following metrics as follows:

- The extent of correctness with which the model predicts the samples is given by validation accuracy.
- The number of relevant instances retrieved w.r.t the total amount of relevant instances is known as Recall = TP/(TP+FN)
- The number of correctly predicted positive observations in comparison to the total predicted positive observations is known as Precision = TP/(TP +FP).
- F1-Score is the harmonic mean of recall and precision is defined as the statistical measure of accuracy.

8. Conclusion

Predicting flight delays has been a topic of interest for a lot of researchers for some time now. It not only holds economic but environmental as well ethical value. Since the issue of flights delays is very important from multiple perspectives, the prediction models must bear extremely high accuracy. The experimental investigation has showed that the best results are yielded by using the tree based models. CART (99.15%), Decision Trees (98.90). Naïve Bayes and XGB Classifier have also shown significantly high accuracy. All of the algorithms give an accuracy score of more than 98%. The implemented supervised ML model indicates that the best way to predict flight delays are by using tree based ML algorithms. The possibility of a brilliant flight delay prediction system by incorporating machine learning is a novel commitment in the field of data science and it will lessen delay related problems for both the aviation industry and their customers.

Future Enhancements

Due to the lack of enhanced computing hardware, we had to restrict our project with 20000 tuples. Well with the increase in computational power of computers, future researchers can increase the number of tuples, and work on a much larger dataset than we have. This may certainly increase the delay prediction accuracy to a greater extent. Researchers can also try to implement deep learning in much better ways to predict flight delays. Also, with the significant development in this field of

ML based flight delay prediction, it is about time that the aviation industry implements these models in real time so that people and airlines can save money as well as time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Kim, Y.J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC).*

Chakrabarty, N. (2019). A data mining approach to flight arrival delay prediction for American Airlines. *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON).*

Nigam, R., & Govinda, K. (2017). Cloud based flight delay prediction using logistic regression. *2017 International Conference on Intelligent Sustainable Systems (ICISS).*

Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., & Vijayaraghavan, V. (2017). A machine learning approach for prediction of on-time performance of flights. *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC).*

Yogita Borse, Dhruvin Jain, Shreyash Sharma, & Viral Vora, Aakash Zaveri. (2020). Flight delay prediction system. *International Journal of Engineering Research and, V9*(03).

Esmaeilzadeh E., & Mokhtarimousavi S. (2020). Machine learning approach for flight departure delay prediction and analysis. *Transportation Research Record: Journal of the Transportation Research Board, 2674*(8), 145-159.

Tu, Yufeng and Ball, Michael O. and Jank, Wolfgang. (2005). Estimating Flight Departure Delay Distributions - A Statistical Approach with Long-Term Trend and Short-Term Pattern. Robert H. *Smith School Research* (*RHS 06-034).*

K., & Rao, K.B. (2018). Machine learning approach to predict flight delays. *International Journal of Computer Sciences and Engineering, 6*(10), 231-234.

Natarajan, V., Meenakshisundaram, S., Balasubramanian, G., & Sinha, S. (2018). A novel approach: Airline delay prediction using machine learning. *2018 International Conference on Computational Science and Computational Intelligence (CSCI).*

Carvalho, L., Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J.A., Brandão, D., Carvalho, D., & Ogasawara, E. (2020). On the relevance of data science for flight delay research: A systematic review. *Transport Reviews*, 1-30.

Yazdi, M.F., Kamel, S.R., Chabok, S.J., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data, 7*(1).

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2017). A statistical approach to predict flight delay using gradient boosted decision tree. *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*.

Safa M, and Pandian A "A review on big IoT data analytics for improving QoS-based performance in system: Design, opportunities and challenges", (2021) *Lecture Notes in Networks and Systems*, 130, 433-444.

Piyush Raj, M. Safa, Nitish Kumar, P J Sahrudh, Shivaditya Singh. (2020). Enhanced Smart Music Controller by Applying CNN in IoT. *International Journal of Advanced Science and Technology, 29*(06), 2739 - 2749.

Abhirup Bose, M. Safa, Sanjay Bhargav Siddi, Manas Raj Anand, Vivek Kumar. (2020). Implementation of Dynamic Lighting & Augmented Reality (DLAR) Smart Home System for Deaf and Hard-of-Hearing (DHH) Residents. *International Journal of Advanced Science and Technology, 29(*06), 2724 - 2738.

Aadhineni Ganesh, M. Safa, M.Roopanjali, P.V.K Sai Sri Harsha, D. Anjana Tulasi. (2020). An Approach for Monitoring and Analyzing Parking Occupancies using IoT. *International Journal of Advanced Science and Technology, 29*(06), 3536 - 3545.