iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33881

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Flight Delay Prediction System using Machine Learning Approach

Durga Ambekar[1], Shreyas Jadhav[2], Aaina Jain[3], Abhay E Patil[4]
[1, 2, 3, 4]*Department of Information Technology, MCT's Rajiv Gandhi Institute of Technology, Mumbai, India*

Abstract: ***Increased growth in aviation companies has led to flight congestion inflicting flight delays. Monitoring air traffic is becoming more and more difficult. Flight delays not solely have economic effects; they also have harmful environmental properties. There are many reasons to prevent flights. Some of these are due to mechanical issues, security issues, weather conditions, airport congestion, etc. We are proposing machine learning algorithms like SVM regression, decision tree regression techniques and hybrid ensemble regression technique. The aim of this research work is to predict the flight delay, which for many countries is the most economically productive and among many transports the fastest and most convenient. By using machine learning algorithms, you can drastically save many business losses.***
Keywords: Flight Delay Prediction, Support Vector Machine, Decision Tree Regression, R^2, MSE, model, prediction

I. INTRODUCTION

Delay is one of the well-known overall performance signs of any transportation system. In particular, commercial airlines understand delay as the period of time during which a flight is delayed or postponed. Traditionally, if the departure or arrival time of a flight is usually longer than 15 minutes than planned departure and arrival time, then it is far taken into consideration that there is a arrival or departure delay with respect to corresponding airports. In this case, it is essential to have an intelligent and automated forecast system that can predict possible airline delays. This research work aims to analyse the flight information of domestic flights operated by airlines, covering the top 5 busiest airports and predicting possible delays in the arrival and departure of flights with data mining and machine learning. With tremendously growing population, timing is everything for many billionaires. The importance of flights was increased here, but due to the high cost and some continuous flight delays, a very less progress on flights was made in 1960's. Now with government aid, many airline companies have begun to make cheaper flights and provide more comfort and many airports have control of air traffic. Huge losses have been occurred in the economic status of the countries. We all know that the latest machine learning technology is one of the ways to determine flight delays. Notable reasons for commercial scheduled flights being delayed include, air traffic congestion, adverse weather conditions and arriving aircraft to be used for the flight, safety issues, and maintenance. This research work has been structured as follows: Introduction, Related Work, Problem Statement, Research Attributes, Proposed System, Results, Future Scope and Conclusion, Acknowledgment and References.

II. RELATED WORK

A good amount of research attention has been dedicated to the study of flight delays; predicting and analysing the delays and their reasons. Different researchers have studied this issue.

A. Review on Machine Learning Techniques

In one among the most effective studies performed, [1] Chakrabarty, Navoneel, et al proposed a Machine Learning Model using Gradient Boosting Classifier for predicting flight arrival delay in 2019. This model takes the flight details concerning American Airlines covering high 5 busiest airports of US as input, analyses it and gives the arrival prediction. This paper proposes a hyper-parameter approach by the use of Grid Search on Gradient Boosting Classifier Model on flight information. Another prominent research was done by [2] Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, Subhas Burman in 2017. Their study uses a measured approach to predict airline delay using gradient boosted decision tree. Manna et al explored and examined the flight data and developed a regression model using Gradient Boosting Regressor for predicting both Flight Departure and Arrival Delays respectively. Another model has been presented by [3] Juan Jose Robollo and Hamsa Balakrishnan where Rebollo applied Random Forest on an air traffic network framework for predicting flight departure delays in future. The main objective of this paper is to predict the departure delay on a particular link or at a particular airport, sometime in the future. Further a research was also done by [4] Sruti Oza, Somya Sharma, Hetal Sangoi, RutujaRaut, V.C. Kotak in 2015.
Oza at al tried investigation weather elicited flight delay prediction by implementing Weighted Multiple Linear Regression on weather-flight information having weather factors. The research found that arrival and departure delays are highly inter-related. Inter-relation between arrival and departure delays is very high (around0.9). A group of researchers, [5] Anish M. Kalliguddi and Aera K. Leboulluec constructed regression models like Decision Tree Regressor, Random Forest Regressor and Multiple Linear

Regressor on flight data for predicting both departure and arrival delays in 2017. Predictive model developed in this investigation can lead to better management decisions allowing for effective flight scheduling. In addition, the highlighted important factors can give an overview into the root cause of aircraft delays. Also, [6] Brett Naul applied Logistic Regression, Naive Bayes Classifier and Support Vector Machine on flight information for prediction of flight departure delay. The aim of his analysis work was to use massive historical datasets to create predictions concerning the timing of future flights way prior to.

B. Review on Deep Learning Techniques:

Another simulation by [7] Young Jin Kim, Sun Choi, Simon Briceno, Dimitri Mavris centered on a Deep Learning Approach using Recurrent Neural Networks (RNNs) for predicting flight delay. In their detailed research work, they used long short-term memory RNN architecture for predicting airline delays. Further, [8] Sina Khanmohammadi, Salih Tutun, Yunus Kucuk projected a Deep Learning Approach using Artificial Neural Network (ANN) and additionally introduced a new type of multilevel input layer ANN. The results counsel that the projected methodology may be effective for specific issues that embrace many nominal variables, such as the transportation problem. One amongst the restrictions of this study is that it needs to be self-addressed in our future work is the complexity of the proposed method (as the quantity of variables will increase, the quantity of connections will also considerably increase).

C. Review in Big Data Approach:

Finally, one of the most important research analyzed by [9] Loris Belcastro, Fabrizio Marozzo, Domenico Talia and Paolo Trunfion proposed a Big Data to predict Flight Delays. The main goal of their research work was to implement a predictor of the arrival delay of a scheduled flight due to weather conditions by analyzing and mining flight data and the respective weather conditions using parallel algorithms implemented as MapReduce programs executed on Cloud Platform for weather affected flight delay prediction.

All the works done in recent years have limitations with respect to accuracy in prediction, large data, weather characteristics to name a few are still related to the topic in a way that contributes to the progress of this article, so here we have included studies that employed a support vector machine (SVM) model to explore the non-linear relationship between flight delay outcomes and another model that explored a broader spectrum of factors. This model could possibly affect the flight delay and proposed a decision tree regressor based model for generalized flight delay prediction. The implemented methods are faced to limitations, because these methods cannot combat against the abundant data volume and complicated computations.

III. PROBLEM STATEMENT

The main objective here is to utilize the available flight operational data and data mining techniques to construct an analytical model. The analytical model constructed here is used to predict the flight delay based on some of the flight attributes which will be discussed in the latter section of this paper. Additional models will be created to determine the most likely cause of a flight delay and to predict the approximate duration of the delay.

IV. RESEARCH ATTRIBUTES

In this paper, we tend to extract some attributes that have an effect on the flight delay, associate formulating them as an input vector x within the projected model as shown below in Table 1.

Airports Reviewed	UA, AA, US, F9, B6, OO, AS, NK, WN, DL, EV, HA, MQ, VX
Flight On-time Performance Information (Input)	Scheduled Departure, Departure Delay, Scheduled Time, Elapsed Time, Air Time, Distance, Scheduled Arrival, Arrival Delay, Previous Arrival Delay, Previous Departure Delay
Selected Features	Scheduled Departure, Departure Delay, Scheduled Time, Elapsed Time, Air Time, Distance, Scheduled Arrival, Arrival Delay, Previous Arrival Delay, Previous Departure Delay, Month, Day of Month, Airline Name, Origin Latitude, Origin Longitude, Destination Latitude, Destination Longitude
Classification (Output)	1 - indicates occurrence of delay 0 - indicates absence of delay
Regression (Output)	Numerical value (Score) of the flight delay prediction

Table 1. Feature Study

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

V. PROPOSED SYSTEM

An outline of the model developed to predict delays of individual flights is shown in Figure 1. The model consists of two main parts, the training process and the prediction process. The training process starts with data collection. Historical flight data and data corresponding to airlines, airports are collected and they are joined together using the scheduled departure time and airport as the join keys. In the pre-processing step, estimating missing data and normalization are performed. Then the training set is finally ready and it is used to train the predictive model with sampling techniques. Data for the prediction process is collected and pre-processed in the same way as the training set. After that it is fed into the model trained with the training data. In the end, the model assigns each data point a label.

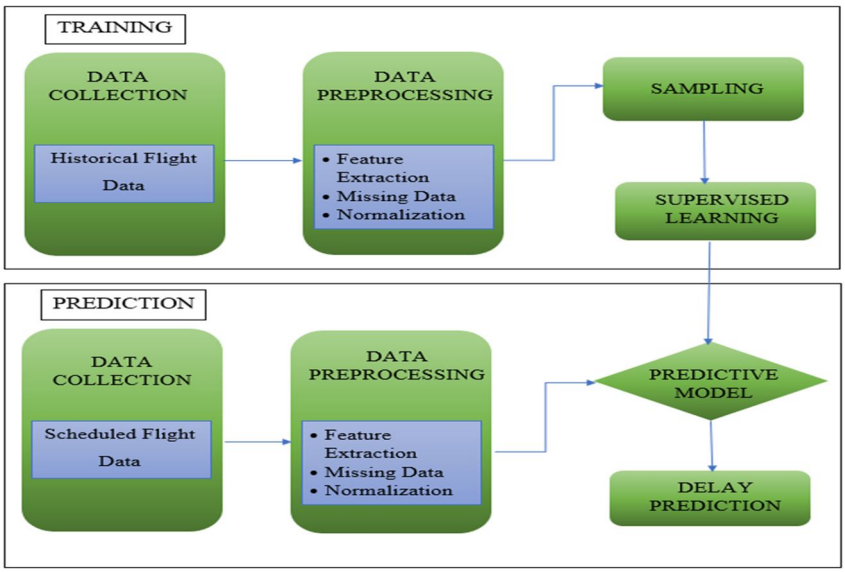


Fig 1: Summary of the model developed

The above system proposed consists of 2 phases:

A. Data and Pre-Processing

1) *Data Collection:* To train and test models, we tend to use a publicly available dataset for domestic air traffic from Kaggle. The native supply of our dataset is the on-line Bureau and Transportation Statistics database. Datasets of varied airports, airlines, flights were incorporated together with the help of joining keys and therefore the resultant dataset was the ultimate dataset on which the whole models was deployed. The info set is from the year 2004-2019 and consists of much above 3 Million examples with following features categorized as follows:

a) Data about flight (day, month, year, airline, flight number, tail number)
b) Data about origin and destination (origin airport, destination airport)
c) Data about the departure (scheduled departure, departure time, departure delay, taxi)
d) Data about the flight-journey (air time, distance, hour, minute, time-hour)
e) Data about the arrival (scheduled arrival, arrival time, arrival delay)

2) *Data Pre-Processing:* Air traffic information for major airports and corresponding weather information is extracted. Following the foundations of BTS, flights that attain the gate within 15 minutes of the scheduled time are possibly regarded as on-time. All of the canceled and diverted flights in the training set are considered to be delayed. To deal with missing data usually encoded as blanks, NaNs or other place holders data pre-processing is done. Such datasets however don't appear to be compatible with scikit-learn estimators that suppose that each value in an array are numerical, and that all have and hold meaning. A basic strategy to use incomplete datasets is to discard the whole rows and/or columns containing missing values. However, this comes at the cost of losing data which can be valuable (even though incomplete A more potent method is to impute the lacking values, i.e., to deduce them from the recognized a part of the data. Here we can use Imputer class of sklearn module and the methodologies used are transform and fit-transform.

The following data fields were extracted from the ultimate dataset for every scheduled flight because those are factors having impacts on flight delays.

a) Year
b) Month
c) Day of Month
d) Day of Week
e) Arrival and Departure Schedule in Local Time
f) Arrival Delay Indicator: 0(zero) if the scheduled arrival time deducted from the actual arrival time is less than 15 minutes, whereas 1 if the scheduled arrival time deducted from the actual arrival time is greater than or equal to 15 minutes.

B. Predictive Model

The model consists of 2 stages:

1) Departure Delay Prediction
2) Arrival Delay Prediction
3) Departure Delay Prediction: The dataset was split into training data and test data in the ratio 90:10 for the purpose of evaluating the model. The results of the model are illustrated in Table 2 later. The Support Vector Machine Regression model performed better than the other models with a R-Squared score of 0.17.
4) Arrival Delay Prediction: Similar to the departure delay prediction phase, the dataset was split into training data and test data in the ratio of 90:10. The results of the classification stage in the model are shown in Table 3 later. As seen from the table, the Support Vector Machine Regression model performs the best with R^2 score of 0.33, followed by Decision Tree Regressor with R^2 score of 1.0. According to the standards set by the Airports Authority of India (AAI), a flight is said to have a class value of 0 if it departs or arrives no later than 15 minutes from the scheduled time, otherwise it is said to have a class value of 1 and is considered as a delay. If the classification stage outputs 0, then there is an absence of delay. If the classification stage outputs 1, then the regression stage predicts the value of delay in terms of a predictive score.

The following algorithms showed the highest accuracy from those implemented:

a) Support Vector Machine
b) Decision Tree - Regression
c) Stacking Algorithm (Hybrid Algorithm consisting of Random Forest Regressor, Decision Tree Regressor, Logistic Regression and SVM Algorithm)

- *Support Vector Machine (SVM):* The concept of SVM classification algorithm is: by the use of a nonlinear transform $\phi(x)$ to map the entered information to a higher-dimensional space, after which doing linear classification of the entered information in dimensional feature space, in order that the optimal hyperplane is constructed.
- *Decision Tree – Regression:* Decision tree builds regression or classification models among the range of a tree structure. In this algorithm the dataset is broken down into smaller subsets and at the same time, an associated decision tree is also incrementally developed. The final result is a tree with decision nodes and leaf nodes-a choice node has a pair of a lot of branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a selection on the numerical target. The highest decision node in every tree that corresponds to the best predictor known as root node. Decision trees can handle every categorical and numerical data.
- *Stacking Algorithm:* The construct of stacking rule is to tackle a retardant by breaking it into multiple sub-problems with every sub-problem obtaining handled by totally different algorithms. Every rule has its own set of strengths and weaknesses. The intuition concerned in stacking rule is to mix the strengths of various algorithms in-order to get a model that would cut back the error rate and boost the general accuracy of the system. Stacking is taken into account as a hybrid rule because it consists of multiple algorithms. that consists of base learners and a final learner. The results obtained by the bottom learners are treated as input for the ultimate level regressor. The work of this final level regressor is to beat the errors made by the bottom learners and to spice up the general accuracy. Here, the stacking algorithm used is a hybrid of Random Forest Regressor, Decision Tree Regressor, Logistic Regression and SVM.

In this paper we have used a mix of three generic algorithms augmenting each other to resolve problems they are not designed to resolve. Since most machine learning algorithms are designed for a selected dataset or task, combining multiple Machine Learning algorithms will greatly improve the overall result by either serving to tune each other, generalize, or adapt to unknown tasks.

VI. RESULTS

A. *Data Visualization*

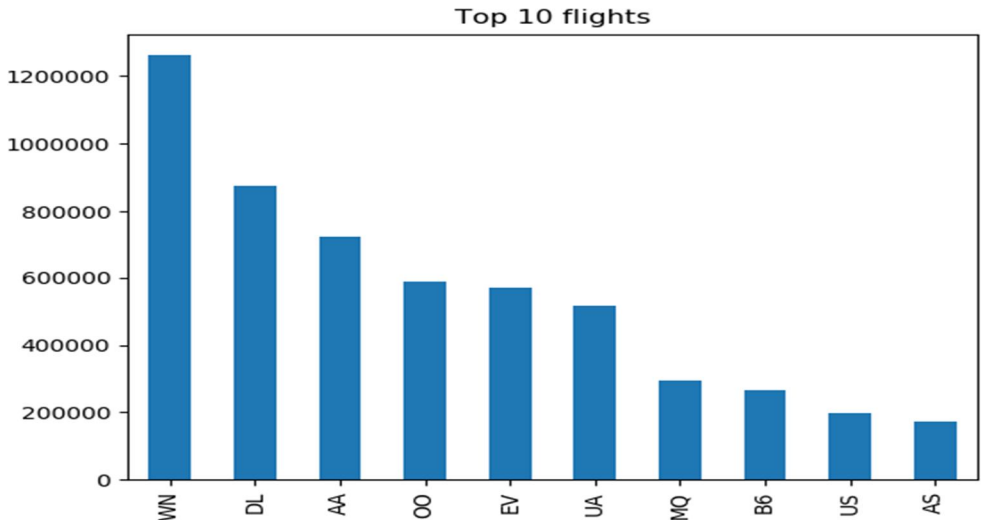


Fig 2: Top 10 flights of US domestic air traffic

In figure 2, we represent the top 10 flights of United States domestic air traffic.
The dataset consist of 3 million examples.
By data visualization we found the number of flights that arrive on time or are delayed for United States.

B. *Comparative Analysis*

SVM, Decision Tree Regression, Stacking models are applied on a dataset using Python programming language, to classify if a flight is delayed based on characteristics such as Origin, Destination, Scheduled Departure, Departure Delay, Scheduled Time, Arrival Delay, Previous Arrival Delay, Previous Departure Delay, Month, Day of Month, Airline Name.

1) *R^2 Score:* R-Squared is a statistical measure of work that indicates the quantity of variation of a dependent variable is explained by the independent variable(s) in a regression model. The scale of R^2 score is instinctive that is it ranges from zero to one. Zero indicating that the projected model doesn't improve prediction over the mean model, and One indicating accurate prediction. Better accuracy of the model, higher the R-squared.
2) *Mean Squared Error Score:* MSE is the average of the square of the error that is used because the loss function for least squares regression. It is the total, over all the information points, of the square of the subtraction between the expected and actual target variables, divided by the total number of information points.

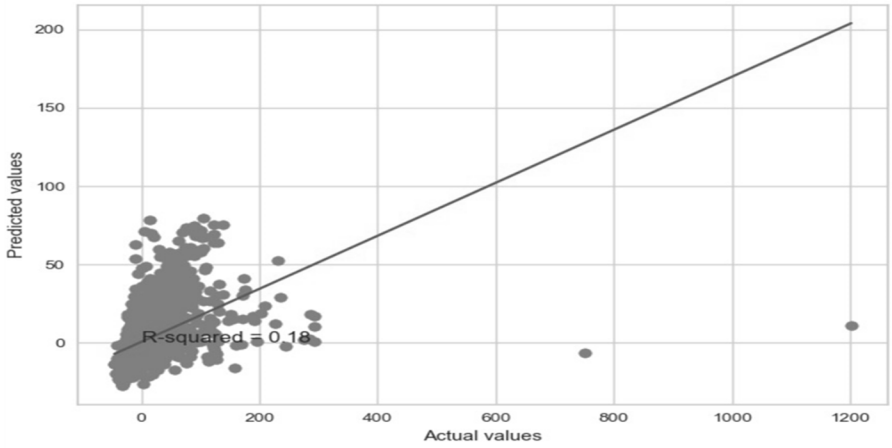


Fig 3: R^2 score for SVM model – Departure Delay

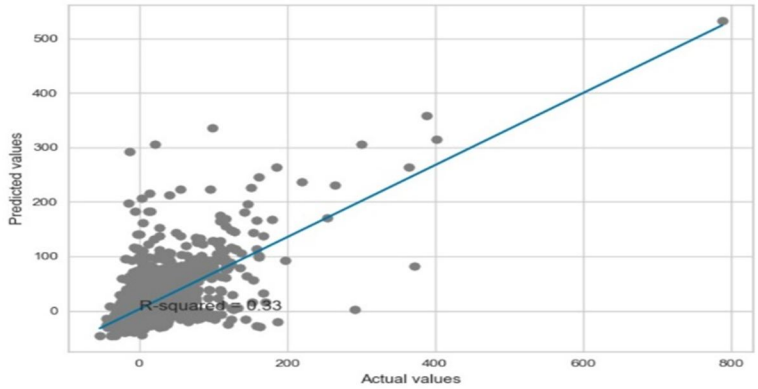


Fig 4: R^2 score for SVM model – Arrival Delay

Figure 3 and figure 4 represent implementation of SVM model and the R^2 score achieved for departure delay and arrival delay is 0.18 and 0.33 respectively.

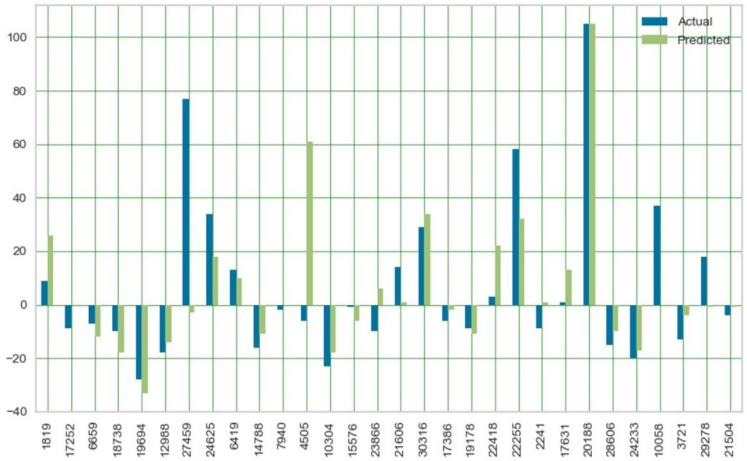


Fig 5: Delayed flights representation

Figure 5 represents the analysis of Actual vs Predicted number of flights that are delayed for United States.
The blue bar shows the actual delay of the flights and green bar shows the predicted flight delay on basis of R^2 score for SVM model.

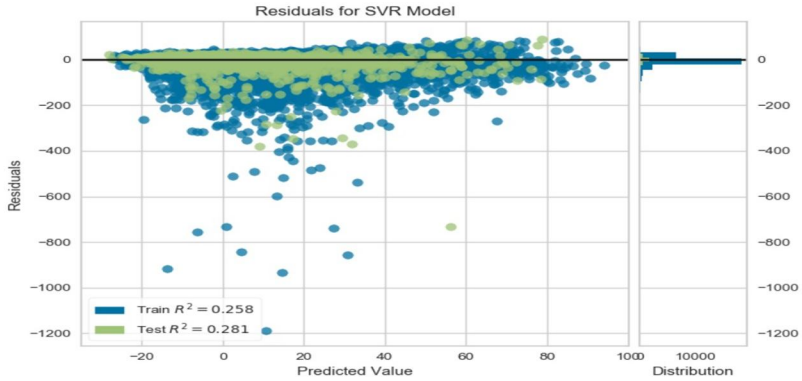


Fig 6: Comparative analysis of training and testing data of R^2 score

In figure 6, we have represented the R^2 score for both training and testing dataset using SVM model.
The Blue dots represent Training results and Green dots represent Testing results.

DEPARTURE DELAY				
	TRAINING		TESTING	
REGRESSOR	R^2 score	MSE	R^2 score	MSE
SVM	*0.18*	*31.63*	*0.18*	*32.00*
DECISION TREE	*1.0*	*0.0*	*0.32*	*28.98*
STACKING	*0.86*	*12.64*	*0.05*	*34.19*

Table 2: Departure Delay scores

ARRIVAL DELAY				
	TRAINING		TESTING	
REGRESSOR	R^2 score	MSE	R^2 score	MSE
SVM	0.33	32.17	0.34	13.99
DECISION TREE	1.0	0.0	0.09	36.27
STACKING	0.88	12.7	0.20	33.87

Table 3: Arrival Delay scores

Table 2 and Table 3 summarizes the accuracy and precision of R^2 score and MSE after implementation of the three algorithms ,i.e. SVM, Decision Tree and Stacking depicting departure and arrival delays.

VII. FUTURE SCOPE AND CONCLUSION

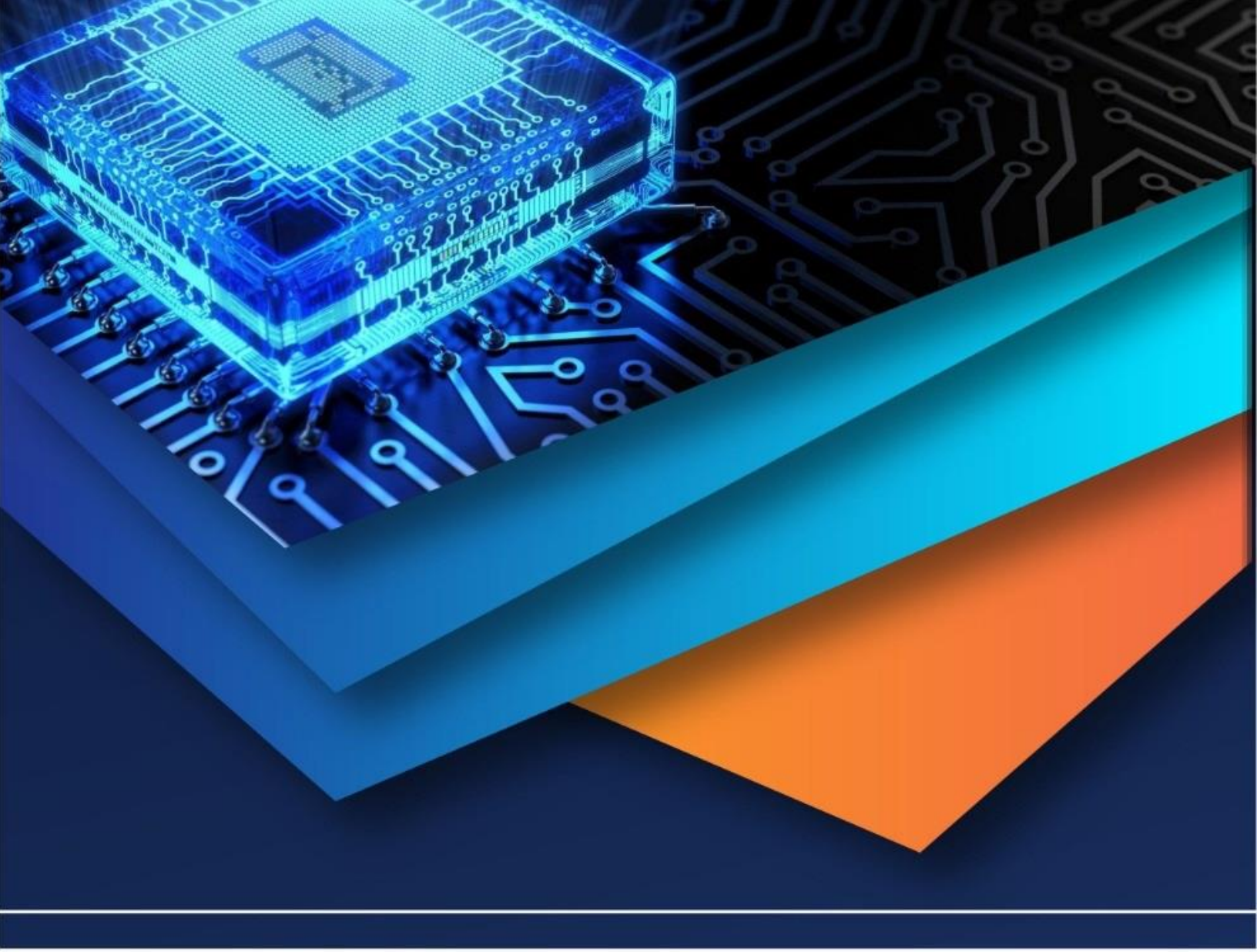
Implementation of the SVM model given the set of attributes (feature values), is able to accurately predict the Arrival Delay and Departure Delay if an aircraft travelling from a specific origin to a destination with a specified set of parameters will arrive on time or get delayed, with an accuracy of 0.18 (Departure Delay) and 0.33 (Arrival Delay) An accuracy near to 1 succinctly proves the efficiency of this model, for our purpose. Thus, this fulfills our requirement of determining the delay for any given aircraft, given merely the parameters of it. The Future Scope of this work involves the appliance of additional and advanced, novel pre-processing techniques, Machine Learning-Deep Learning Hybrid Models tuned with Grid rummage around for achieving higher model performance. Also additional research on using live dataset of airlines and flights can be used to give better and more accurate predictions that are valuable to the Airlines as well as the commuters.

VIII. ACKNOWLEDGEMENT

This paper and the research behind it would not have been possible without the exceptional support of our project guide and supervisor, Mr. Abhay Patil. His enthusiasm, knowledge and exact attention to detail have been an inspiration and kept the work on track from the beginning. We would also like to thank our friends and family who supported us and also offered deep insight into the study.

REFERENCES

[1] Chakrabarty, Navoneel, et al. ”Flight Arrival Delay Prediction Using Gradient Boosting Classifier.” Emerging Technologies in Data Mining and Information Security. Springer, Singapore, 2019. 651-659.
[2] Suvojit Manna, Sanket Biswas, Riyanka Kundu, Somnath Rakshit, Priti Gupta, Subhas Burman ”A statistical approach to predict flight delay using gradient boosted decision tree”, International Conference on Computational Intelligence in Data Science(ICCIDS), 2017
[3] Juan Jose Robollo and Hamsa Balakrishnan ”Characterization and Prediction of Air Traffic Delays”,
[4] Sruti Oza, Somya Sharma, Hetal Sangoi, Rutuja Raut, V.C. Kotak ”Flight Delay Prediction System Using Weighted Multiple Linear Regression”, International Journal Of Engineering And Computer Science ISSN:2319-7242, Volume 4 Issue 4 April 2015, Page No. 11668-11677
[5] Anish M. Kalliguddi and Aera K. Leboulluec ”Predictive Modeling of Aircraft Flight Delay”, Universal Journal of Management 5(10): 485- 491, 2017, DOI: 10.13189/ujm.2017.051003
[6] Brett Naul ”Airline Departure Delay Prediction”,
[7] Young Jin Kim, Sun Choi, Simon Briceno, Dimitri Mavris ”A deep learning approach to flight delay prediction”, 35th Digital Avionics Systems Conference (DASC), 2016
[8] Sina Khanmohammadi, Salih Tutun, Yunus Kucuk ”A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport”, doi.org/10.1016/j.procs.2016.09.321
[9] Loris Belcastro, Fabrizio Marozzo, Domenico Talia and Paolo Trunfion ”Using Scalable Data Mining for Predicting Flight Delays”
[10] https://en.wikipedia.org/wiki/Flight cancellation and delay



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY