WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Prediction of Estimated Time of Arrival of an Aircraft

Cheekuri Dinesh[1], Dinesh Makkapati[2], Enumula Siddhu[3], V. Kamakshi Prasad[4]

[1]Under Graduate Student, Jawaharlal Nehru Technological University, Hyderabad - 500085. Telangana, India
[2]Under Graduate Student, Jawaharlal Nehru Technological University, Hyderabad - 500085. Telangana, India
[3]Under Graduate Student, Jawaharlal Nehru Technological University, Hyderabad - 500085. Telangana, India
[4]Professor, Dept. of CSE, Jawaharlal Nehru Technological University, Hyderabad - 500085. Telangana, India

ABSTRACT

Day by day the demand for airline transportation is gradually increasing. Due to the swift development of civil aviation, flight delays have become a significant topic and a major problem for aviation transportation systems all over the world. The air transportation industry is continuously suffering from economic losses related to flight delays. These flight delays result in a loss of time and missed business opportunities or leisure activities for the passengers and the airlines trying to make up for these passengers for the flight delays leading to extra fuel consumption and a major adverse environmental impact. To overcome these challenges and reduce the negative economic and environmental impacts caused by these unexpected flight delays and balance the increasing flight demand with the growing flight delays, aviation delays in airports must be accurately predicted. Machine learning is the term given to algorithms that allow computers to analyze data, pick up potential patterns and use them to make predictions.

Keywords: Air traffic, Air traffic flow, Air traffic management, Machine Learning, Air traffic flow management.

1. INTRODUCTION

Flight delays usually occur because the supply of airspace capacity is too low to meet the demand for air travel. Many researchers have studied that flight delays can be a result of insufficient air traffic control and irregularity of airline operations while the dominant factor that usually causes almost 75% of system delays is adverse weather conditions. These flight delays can also be caused by many different reasons involving various civil aviation agents. Disruptions in the air traffic system as a result of these factors lead to more and more subsequent delays for flights concerning numerous airports and airlines. In order to store, interpret, and forecast massive amounts of flight data, researchers can use machine learning.

Consequently, it has become quite intriguing to do a study on the analysis of aircraft delays. Various types of machine learning and data mining techniques are studied by various researchers to investigate this problem. These researchers looked at many factors, including the capacity of the airport, weather patterns, and the placement of airport facilities.

Machine learning is the term given to algorithms that allow computers to analyze data, pick up potential patterns and use them to make predictions. Algorithms for learning can shed light on the relative difficulty of learning in various contexts. The most popular machine learning algorithms are supervised and unsupervised learning, while there are many more. A supervised learning algorithm has generated a function that transforms an input into a desired output. The main forms of supervised learning algorithms are regression and classification. Unsupervised learning represents an input set devoid of labeled instances.

2. RESEARCH QUESTIONS

i. What are the reasons behind not removing outliers from the dataset, and why is it important in training the model?

ii. Why weather is not considered as a parameter in training the machine learning model?

iii. How can the accuracy be increased in the prediction of the landing time of an aircraft?

iv. Why data is obtained only from the flightradar24.com website and isn't there any source to obtain the required data?

v. Why tree-based machine learning models are performing worse when compared with linear models of regression?

3. RELATED WORKS

According to H. Khaksar and A. Sheikholeslami: Visibility, wind, and departure time are factors impacting US networks whereas aircraft type and fleet age are factors affecting Iranian airline flights. The suggested methods showed greater than 70% accuracy in estimating delay occurrence and size in both the US and Iranian whole-network. But when it comes to Indian climatic conditions the parameters such as visibility and wind cannot be predicted accurately beforehand.

Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020) studied support vector machine (SVM) model to explore the non-linear relationship between flight delay outcomes, and sensitivity analysis was performed to assess the relationship between dependent and explanatory variables. The impacts of various explanatory variables are examined in relation to delay, weather information, airport ground operation, demand capacity, and flow management characteristics. The variable impact analysis reveals factors such as pushback delay, taxi-out delay, ground delay program, and demand-capacity imbalance.

Ye, B., Liu, B., Tian, Y., & Wan, L. (2020) propose a new methodology for predicting aggregate flight departure delays in airports by exploring supervised learning methods. Individual flight data and meteorological information were processed to obtain four types of airport-related aggregate characteristics for prediction modeling.

Stefanovič, P., Štrimaitis, R., & Kurasova, O. (2020) did an analysis to predict the interval of time delay deviation using seven algorithms: probabilistic neural network, multilayer perceptron, decision trees, random forest, tree ensemble, gradient boosted trees, and support vector machines and observed that the highest accuracy is obtained using the tree model classifiers and the best algorithm of this type to predict is gradient boosted trees.

Yuemin Tang records information on flights departing from JFK airport for one year and was used for the prediction. Seven algorithms (Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Gradient Boosted Tree) were trained and tested to complete the binary classification of flight delays.

Khanmohammadi, S.; Tutun, S.; Kucuk used an artificial neural network (ANN), the proposed method was applied to predict the delay of incoming flights at JFK airport, where the neurons of each sublayer of the input layer symbolize the delay sources at different levels of the system, and the activation of each neuron represents the possibility of being the source of overall delay.

3. IMPLEMENTATION

Actual Time of Arrival (ATA) of an aircraft prediction through proposed approaches that are based on machine learning algorithms. Parameters that enable the effective estimation of the Actual Time of Arrival are identified, after which **Multi Linear Regression**, **Bayesian linear regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression** are used to estimate the Actual Time of Arrival of a particular aircraft. These methods were tested on **Dubai International Airport (DXB)** and **Rajiv Gandhi International Airport (HYD)** flight datasets.

3.1 DATA COLLECTION

The data used for analysis is obtained from https://www.flightradar24.com/ which is a global flight tracking service that provides us with real-time information about thousands of aircraft around the world.

It consists of four datasets containing data about the past one-year travel details of unique flights that have arrived to or departed from Rajiv Gandhi International Airport (VOHS) and Dubai International Airport (DXB) in the three days from 27 October 2022 to 29 October 2022.

The first dataset contains past one-year travel details of **258 flights arriving at Rajiv Gandhi International Airport (VOHS)** containing 77853 lines of individual flight information, the second dataset contains details of **261 flights departing from Rajiv Gandhi International Airport (VOHS)** containing 76973 lines of individual flight information, the third dataset contains details of **624 flights arriving to Dubai International airport(DXB)** containing 147249 lines of individual flight information and the fourth dataset contains details of **633 flights departing from Dubai International airport(DXB)** containing 152789 lines of individual flight information.

All the datasets contain 8 columns. All the data collected is from the past year and only unique flights are considered. A detailed description of data set attributes is presented in Table 1.

3.2 DATA DESCRIPTION

Table 1. Attribute description for the data set.

Attribute Name	**Description**	**Type**
FLIGHTS	Flight ID	Object
DATE	Date of flight	datetime64[ns]
FROM	Origin	Object
TO	Destination	Object
AIRCRAFT	Aircraft Type and Registration	Object
FLIGHT TIME	Time taken for the airplane to travel from origin to destination	Object
STD	Scheduled Time of Departure	Object
ATD	Actual Time of Departure	Object
STA	Scheduled Time of Arrival	Object
STATUS	Status of the flight	Object
ATA	Actual Time of Arrival	Object

3.3 DATA PREPROCESSING

Data cleaning is a very important step in data preprocessing, Here is where each dataset is cleaned. This means the model is prepared to fit the needed format, by eliminating unwanted values such as outliers, extreme values, or values, which aren't wanted to be processed, transformed, or adapted to the desired format.

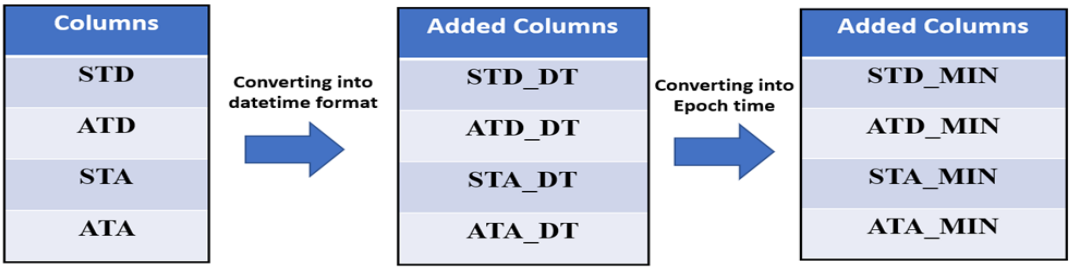
Adjustments are made to the dataset such as rows containing null values being deleted. The first dataset contains 1037 rows with null values, the second contains 991 null-valued rows, the third contains 2580 null-valued rows and the fourth dataset contains 4420 rows with null values. All these rows are deleted from the dataset.

Further preprocessing is done such as the columns **FROM, TO,** and **AIRCRAFT** are split into desired columns i.e. (**FROM=>(FROM, ORIGIN)**), (**TO => (TO, DESTINATION)**), **(AIRCRAFT=>(AIRCRAFT, REGISTRATION)).**

New columns **STD_DT, ATD_DT, STA_DT,** and **ATA_DT** are added which are the columns STD, ATD, STA, and ATA in datetime format. The values of these columns in minutes in epoch time are also calculated and these columns are named as **STD_MIN, ATD_MIN, STA_MIN,** and **ATA_MIN.**

New columns such as the scheduled flight time and the actual flight time, the scheduled flight time in minutes and the actual flight time in minutes, the departure delay in minutes, and the arrival delay in minutes are added.

A column determining the exponential moving average of the arrival delay(in minutes) grouped by the attributes **TO** and **AIRCRAFT** considering a span of 3 days is added and is named **EMA.**



A detailed description of the data set attributes after data preprocessing is presented in Table 2.

Table 2. Attribute description for the data set after data preprocessing.

Attribute Name	Description	Type
FLIGHTS	Flight ID	Object
DATE	Date of flight	datetime64[ns]
FROM	Origin airport	Object
TO	Destination airport	Object
AIRCRAFT	Aircraft Type	Object
FLIGHT_TIME	Time taken for airplane to travel from origin to destination	Object
STD	Scheduled Time of Departure	Object
ATD	Actual Time of Departure	Object
STA	Scheduled Time of Arrival	Object
ATA	Actual Time of Arrival	Object
STATUS	Status of the Aircraft	Object
REGISTRATION	Registration of the aircraft	Object
ORIGIN	Origin airport in IATA code	Object
DESTINATION	Departure airport in IATA code	Object
STD_DT	Scheduled Time of Departure in datetime format	datetime64[ns]
ATD_DT	Actual Time of Departure in datetime format	datetime64[ns]
STA_DT	Scheduled Time of Arrival in datetime format	datetime64[ns]
ATA_DT	Actual Time of Arrival in datetime format	datetime64[ns]
STD_MIN	Scheduled Time of Departure in minutes in epoch time	float64
STA_MIN	Scheduled Time of Arrival in minutes in epoch time	float64
ATD_MIN	Actual Time of Departure in minutes in epoch time	float64
ATA_MIN	Actual Time of Arrival in minutes in epoch time	float64
ST_MIN_DIFF	Scheduled flight time in minutes	float64
AT_MIN_DIFF	Actual flight time in minutes	float64
STA-STD	Scheduled flight time	timedelta64[ns]

ATA-ATD	Actual flight time	timedelta64[ns]
ATD-STD_MIN	Departure Delay in Minutes	float64
ATA-STA_MIN	Arrival Delay in Minutes	float64
EMA	Exponential Moving Average	float64

4. Feature selection

The value which is to be predicted here is the **ATA_MIN** column which contains the actual time of arrival of the aircraft. The features which are selected so as to predict ATA_MIN is the origin airport for the arrival datasets and the destination airport for the departure datasets, the type of aircraft and the scheduled time of departure in minutes i.e. **STD_MIN**, the scheduled time of arrival in minutes i.e. **STA_MIN**, the actual time of departure in minutes i.e. **ATD_MIN.**

Since STD_MIN, STA_MIN, and ATA_MIN are in an integral format unlike STD_DT, STA_DT, and ATA_DT which are columns with similar data but in datetime format, we select STD_MIN, STA_MIN, and ATA_MIN for prediction since some of the selected algorithms can only deal with numerical data. As we have seen in the earlier section, the data set contains both numerical variables and categorical variables. These categorical variables are needed to be converted to numerical variables to avoid the algorithms failing to work when faced with categorical data. Therefore, the method of one-hot encoding is used to convert category labels to unique integer numbers to be used for training and testing.

All other variables have very little influence on predicting flight delays and so are dropped. Some new features can also be added which can have a considerable influence on the prediction of arrival time of an aircraft such as the distance between the origin and the destination airports can be considered as an important feature. This distance is needed to be added to our data, for this we can gather information on various different airports from the internet. The downloaded dataset contains the latitudes, longitudes, the name of the airport, and the iata_code of each airport, this code is used to map this data to our preprocessed datasets. Latitudes and longitudes of both origin and destination airports are added to the dataset.

The distance between the airports is calculated based on these coordinates. The haversine distance formula is used for this purpose and the distance in km is added as a new column to our preprocessed data as **Distance_km** which is included as a new feature for prediction.

5. Prediction

A supervised machine learning approach was used in this study. A dataset has a target variable, the goal is often to train a computer to create a regression system [8] to predict that target variable. The primary goal of this project is to use label data to anticipate the real arrival time of aircraft. Therefore, the supervised learning regression algorithm was chosen as more suitable. Predicting aircraft arrival times can be viewed as a regression problem where given data is used to predict a target variable (in this case, actual aircraft arrival times). Five supervised machine learning algorithms, **Multiple Linear Regression, Bayesian Regression, Decision Tree Regressor, Random Forest,** and **Support Vector Regression** are considered.

Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of the dependent variable. This technique allows analysts to determine model variation and the relative contribution of each independent variable to the overall variance.

Bayesian linear regression characterizes the mean of one parameter by the weighted sums of other variables. This type of conditional modeling aims to determine the prior distribution of regressors and other variables that explain regressand assignments, and ultimately, depending on observations of regression coefficients, out-of-sample predictions of regressands are made possible.

Decision tree regression observes the features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

Random forest is a classifier that takes a set of decision trees for different subsets of a given data set and averages them to improve the prediction accuracy of that data set. Instead of relying on a single decision tree, random forest takes predictions from each tree and predicts the final output based on the majority vote of the predictions.

Support vector regression is a supervised learning algorithm used for predicting discrete values. Support vector regression uses the same principles as SVM. The basic idea behind SVR is finding the best-fit line. In SVR, the best-fit straight line is the hyperplane with the maximum number of points.

All the datasets are split into training data and testing data, 67% of the entries are considered as training data while the remaining 33% of the entries are used as testing data. The same split is done for all four datasets. Un-normalized data is used for the first four machine learning models while normalized data is considered for support vector regression.

6. RESULTS

Table-A: Performance matrix comparison for **HYD-arrivals**

Type of data: Un-normalized

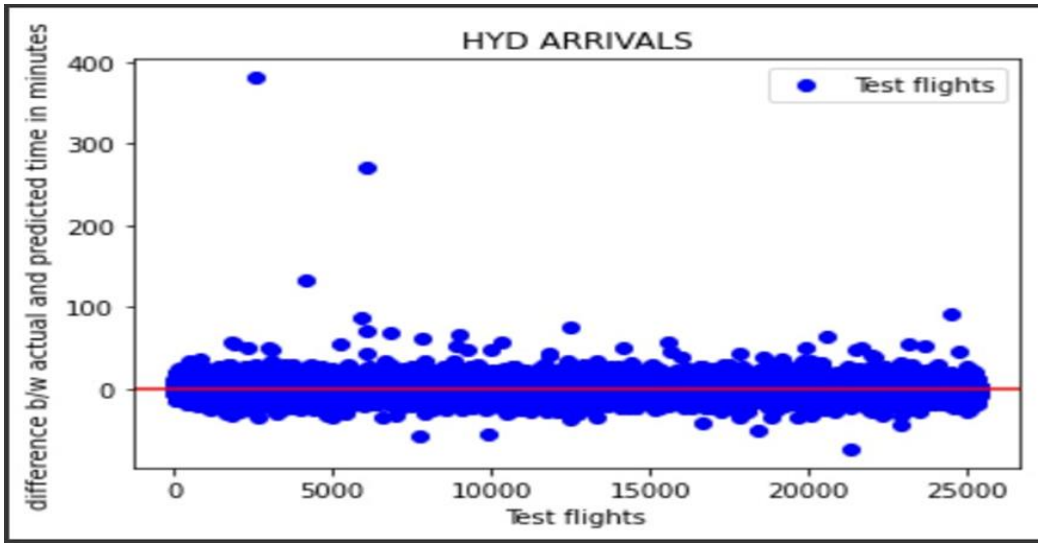
Algorithm	**r2 score**	**Root mean square error**
Multiple Linear Regression	**0.9999999973627461**	**7.705998706845894**
Bayesian Regression	0.9999999973621461	7.706875368527697
Decision Tree Regressor	0.9999998522285787	57.68304599205984
Random Forest	0.9999998669463381	54.73515800170144

Type of data: Normalized

Support Vector Regression	0.9963509453895333	0.060692013356475734

It can be observed from Table-1 that the RMSE values are very low for linear models of regression the model i.e., Multi Linear Regression and Bayesian Regression gives better prediction on the Hyderabad arrivals dataset.

The minimum RMSE is 7.7 is observed in **Multi Linear Regression** and **Bayesian Regression.**



The linear models of regression performed much better when compared with tree-based regression models.

Table-B: Performance matrix comparison for **HYD-departures**

Type of data: Un-normalized

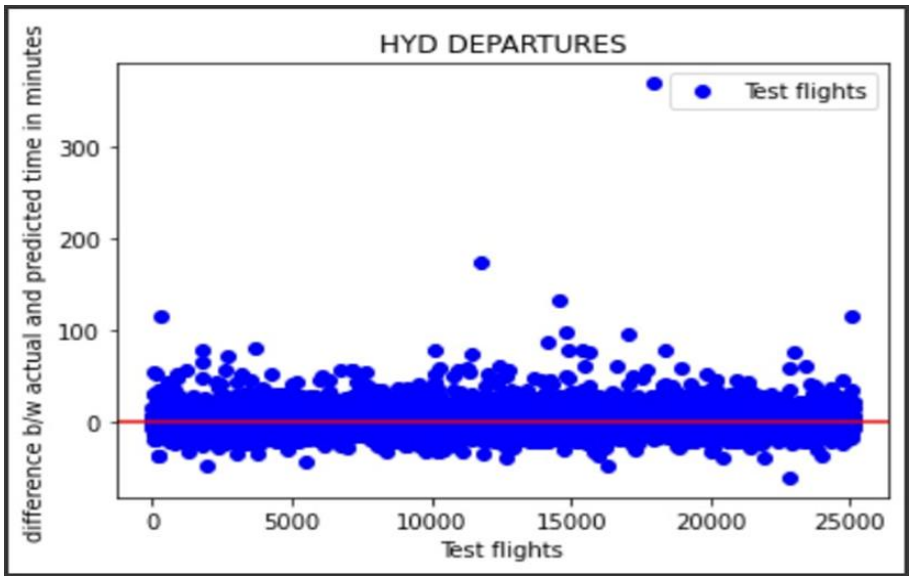
Algorithm	**r2 score**	**Root mean square error**
Multiple Linear Regression	0.9999999969364027	7.705998706845894
Bayesian Regression	**0.9999999969372853**	**8.1836561228295238**
Decision Tree Regressor	0.9999998736481813	52.56356899205456
Random Forest	0.9999998943162011	48.072676749164096

Type of data: Normalized

Support Vector Regression	0.9965459030963801	0.05874255044903601

It can be observed from Table-2 that the RMSE are very low for linear models of regression i.e., Multi Linear Regression and Bayesian Regression the model gives better prediction on Hyderabad departures dataset.

The minimum RMSE is 8.1 is observed in **Multi Linear Regression** and **Bayesian Regression.**



The linear models of regression performed much better when compared with tree-based regression models.

Table-C: Performance matrix comparison for **DXB-arrivals**

Type of data: Un-normalized

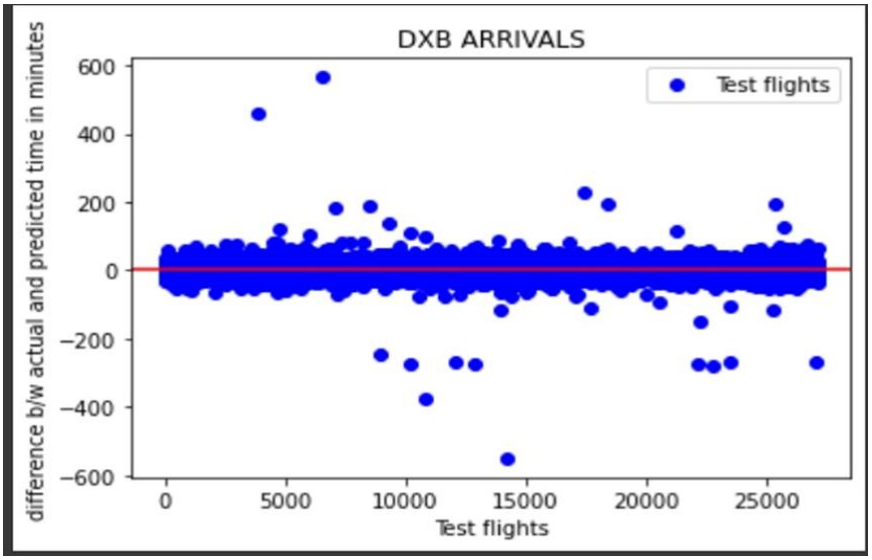
Algorithm	**r2 score**	**Root mean square error**
Multiple Linear Regression	0.9999999900523118	15.032348985341725
Bayesian Regression	**0.9999999901984422**	**14.921528775090138**
Decision Tree Regressor	0.9999998736481813	77.47908431196237
Random Forest	0.9999998943162011	48.072676749164096

Type of data: Normalized

Support Vector Regression	0.9965459030963801	0.05874255044903601

It can be observed from Table-3 that the RMSE values are low for linear models of regression i.e., Multi Linear Regression and Bayesian Regression the model gives better prediction on the Dubai Arrivals dataset.

The minimum RMSE is approximately 15 is observed in **Multi Linear Regression** and **Bayesian Regression.**



The linear models of regression performed much better when compared with tree-based regression models.

Table-D: Performance matrix comparison for **DXB_departures**

Type of data: Un-normalized

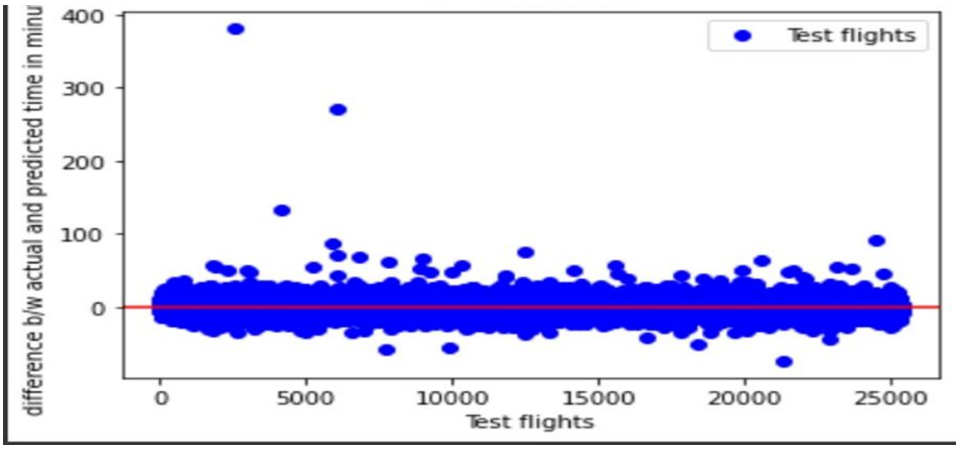
Algorithm	**r2 score**	**Root mean square error**
Multiple Linear Regression	0.9999999834141766	19.404557492646532
Bayesian Regression	**0.9999999834274867**	**19.396769848392996**
Decision Tree Regressor	0.9999997411921722	76.65207724704528
Random Forest	0.9999996975597403	82.86192644668358

Type of data: Normalized

Support Vector Regression	0.988459030862435	0.06892729675086151

It can be observed from Table-4 that the RMSE values are for linear models of regression i.e., Multi Linear Regression and Bayesian Regression the model gives better prediction on the Dubai departures dataset.

The minimum RMSE is 19.3 is observed in **Multi Linear Regression** and **Bayesian Regression.**



The linear models of regression performed much better when compared with tree-based regression models.

7. Conclusion

In this paper, we applied a machine learning problem to predict the actual time of arrival of an aircraft. A supervised machine learning approach in the form of regression was used for prediction. For this prediction, we used 5 machine learning algorithms and for the algorithm performance evaluation, we used 3 metrics. After applying the regressor to the arrival time predictions, we compared those three measures to assess the performance of each model.

The result shows that the highest values of r2 score and the least mean square error and root mean square errors are generated by the Bayesian Regression algorithm. Every model has an r2 score of more than 0.99 which indicates that all the models perfectly fit the datasets and the Bayesian Regression algorithm gives a root mean squared error of less than 10 for the two Hyderabad airport datasets and less than 20 for the two Dubai airport datasets i.e., 7.7 for VOHS arrivals, 8.18 for VOHS departures, 14.9 for DXB arrivals and 19.39 for DXB departures which indicates that the bayesian regression algorithm works well in predicting the arrival time of an aircraft.

Other regressors also show good performance. Multilinear and Bayesian regressors give significantly lower error rates when compared to two other regressors which are tree-based i.e., Decision Tree regressor and Random Forest regressor and SVR, these are not linear models unlike the Multilinear and Bayesian regression models.

Algorithm comparison results show that the linear model regression algorithms tend to better predict the arrival time of aircraft for these data sets.

8 REFERENCES

1. Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. Scientia Iranica. https://doi.org/10.24200/sci.2017.20020
2. Xu, N.; Sherry, L.; Laskey, K.B. Multi-factor model for predicting delays at US airports. Transp. Res. Rec. 2008, 2052, 62–71. https://journals.sagepub.com/doi/abs/10.3141/2052-08.
3. 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. https://doi.org/10.1177/0361198120930014
4. Beatty, R. Preliminary Evaluation of Flight Delay Propagation Through an Airline Schedule. In Proceedings of the 2nd USA/Europe Air Traffic Management R&D Seminar, Orlando, FL, USA, 1–4 December 1998 https://arc.aiaa.org/doi/abs/10.2514/atcq.7.4.259.
5. Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. Sustainability, 12(7), 2749. https://doi.org/10.3390/su12072749
6. Allan, S.S.; Beesley, J.A.; Evans, J.E.; Gaddy, S.G. Analysis of delay causality at Newark international airport. In Proceedings of the 4nd USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 3–7 December 2001. https://archive.ll.mit.edu/mission/aviation/publications/publication-files/ms-papers/Allan_2001_ATM_MS-14812_WW-10283.pdf
7. ATLIOĞLU, M. C., BOLAT, M., ŞAHİN, M., TUNALI, V., & KILINÇ, D. (2020). Supervised learning approaches to flight delay prediction. Sakarya University Journal of Science. https://doi.org/10.16984/saufenbilder.710107
8. Pyrgiotis, N.; Malone, K.M.; Odoni, A. Modeling delay propagation within an airport network. Transp. Res. Part C 2013, 27, 60–75. https://www.sciencedirect.com/science/article/pii/S0968090X11000878
9. Stefanovič, P., Štrimaitis, R., & Kurasova, O. (2020). Prediction of flight TIME deviation for Lithuanian airports using supervised machine learning model. Computational Intelligence and Neuroscience, 2020, 1–10. https://doi.org/10.1155/2020/8878681
10. Oladipupo, T. (2010). Types of machine learning algorithms. New Advances in Machine Learning. https://doi.org/10.5772/9385
11. Nibareke, T., & Laassiri, J. (2020). Using big Data-machine learning models for DIABETES prediction and flight DELAYS ANALYTICS. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00355-0
12. Khanmohammadi, S.; Tutun, S.; Kucuk, Y. A new multilevel input layer artificial neural network for predicting flight delays at JFK airport. Procedia Comput. Sci. 2016, 95, 237–244. https://www.sciencedirect.com/science/article/pii/S1877050916324942
13. https://www.flightradar24.com