iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** II **Month of publication:** February 2022

DOI: https://doi.org/10.22214/ijraset.2022.40253

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: 7.538
Volume 10 Issue II Feb 2022- Available at www.ijraset.com

Aircraft Ticket Price prediction using Machine Learning

Janhvi Mukane[1], Siddharth Pawar[2], Siddhi Pawar[3], Gaurav Muley[4]
[1, 2, 3, 4]*Department of Information Technology, Vidyalankar Institute of technology, Mumbai, India*

Abstract: ***With ever increasing air route connectivity throughout the world, air travel has become a common, integral and faster way to travel. Predicting fares for airlines is an important as well as challenging task since a constant fluctuation in fares is observed and it is known to be dependent on varied set factors. With tremendous study in area, it is observed that using Machine Learning, Artificial Intelligence and Deep Learning techniques an estimation of flight fares at a given time can be obtained within seconds. In this paper, we use a Machine Learning Regression approach to predict flight fare by providing basic details of departure date and time, arrival time, source, destination, number of stops and name of the airline. The results show that Random Forest Regression Model provides highly optimal results.***
Keywords: Machine Learning, Random Forest, Prediction models, Airfare Prices, data analytics

I. INTRODUCTION

Everybody knows that holidays always call for a much-needed vacation and finalizing the travel itinerary becomes a tedious task. With the worldwide growth of internet and E-commerce, commercial aviation industry has witnessed a tremendous growth and has become a regulated marketplace. [1] Hence, for Airline revenue management, different strategies like customer profiling, financial marketing, social factors are used for setting ticket fairs. It is often seen that airfares are low when tickets are booked months in advanced and then they rise when booked in urgency. [2]. But, number of days/hours until departure isn't the only factor which decides flight fare, there are numerous other factors as well.

Because of this complex pricing model of aviation industry, customers find it very difficult to find a perfect and cheapest ticket deal. To solve this problem, Machine Learning and Deep Learning based several technologies and modals are developed and extensive research is also underway. This paper throws light on Machine Learning based Flight fare Prediction System which uses Random Forest Regression to estimate prices of airline tickets. Various features that influence prices are also studied along with system's experimental analysis. In Section II, literature survey was carried out wherein, technical papers and some existing models and systems were studied. Differences in the features considered are also mapped down, In Section III, the proposed system is described in detail along with the workflow and its features. In Section IV, implementation part of the model is discussed. In Section V, results are presented along with various comparisons between findings. In Section VI, conclusions are stated and possible advances for future research are mentioned.

II. LITERATURE REVIEW

K. Tziridis, Th. Kalampokas, et.al in [3] have developed an airfare price prediction system. The paper begins with a piece of general information about Machine learning and then the authors further proceed to the methodology comprising of four distinct phases of Feature Selection that influence airfare prices, collection of data from Greek Aegean Airlines, Selection of accurate ML Regression model, and its evaluation. The airline dataset had the following eight features- departure and arrival time, number of free luggage, days before departure, number of intermediate stops, holiday, time of day, any day of the week. The authors performed prediction using eight state-of-art regression Machine Learning models including, MLP, GRNN, ELM, Random Forest Regression Tree, Regression Tree, Bagging Tree, Bagging Regression Tree, Regression SVM, and Linear Regression. Performances of these ML models were also compared and evaluated. The Bagging Regression Tree model outperforms other models with its accuracy of 87.42%.

Tianyi Wang, Samira Pouyanfar, et. al in [4] states the problem of market segment level airfare price prediction and propose a novel application for the same using a Machine learning approach. For training and evaluation of the proposed model, two public datasets, DBIB and T-100 were collected with minimal features. The methodology includes data cleaning, data transformation, data pre-processing, selection of extracted features, and applying ML model. The extracted features include distance, seat class, passenger volume, load factor, competition factor, LCC presence, Crude oil price, CPI, and Quarter. Random Forest Model is used for development because of its best performance on the data in comparison to other models including LR< SVM and Neural Networks. This prediction framework achieves high accuracy with an R squared score of 0.869.

Tao Liu, Jian Cao, et. al in [5] address the problem of airfare forecasting and introduce an ACER framework for airfare price prediction which predicts the lowest ticket price available before departure day. The model is deployed using three steps, namely Feature Selection and Extraction, Selecting a Forecast algorithm, and Multistep Forecasting. The dataset is collected from leading OTAs in China. For feature extraction, a matrix-like a schema is used with matrix rows comprising consecutive departure dates and columns with the number of days before departure. Model's input features include prices of the same itinerary, prices of itineraries departing in the last few days, statistical values, route features, and airfare searching times. Bayesian Regression is used as the base model and result analysis is based on the metrics of RMSE. Results from the experimental analysis showed that ACER performed better with an error of just between 3.7% and 6%.

Supriya Rajankar, Neha Sakharkar, et, al. in [6] put forward Machine Learning Regression methods to predict the price of a flight ticket at a given time. The paper describes its methodology which starts with the data collection process and the dataset is procured from makemytrip.com. This dataset has seven components namely, Date of journey, time of departure, place of departure, time of arrival, place of destination/arrival, airway company, and total fare. Next, the data is cleaned, pre-processed and analysis is performed using different AI models. Authors perform a comparative study of results based on the performance of various Machine Learning models like LR, Decision tree, SVM, KNN, Random Forest, and Bagging Regression Tree. It was observed that KNN gives R-squared value nearing 1 indicating high accuracy.

Juhar Ahmed Abdella, Nazar Zaki, et, al. in [7] present a review of deep learning and social media data-based Airline ticket price prediction model. The authors introduce the current airline ticket pricing situation with the factors that affect ticket prices. They also touch upon the strategies which airlines induce to increase their revenue and maximize profits. This model helps its users by advising them whether to buy tickets or wait for a suitable time to get the optimal deal. It uses data mining techniques like Rule Learning, Reinforcement Learning, time-series methods, and their combinations to achieve greater accuracy in predicting the fare of flights. Features considered for the study include flight number, hours till departure, the current price of a ticket, airline, and its route. The model attained maximum accuracy of 61.9% when a combination of the above-mentioned techniques was used.

III. PROPOSED SYSTEM

We are proposing a system that helps the user to predict the price of an airline ticket with optimum accuracy. Firstly, the user needs to fill the required input fields provided on the webpage. The input fields include the information about the date of the journey i.e., the date of departure and the departure time suitable for the user to start his flight. Up next, the user needs to select the arrival time. Source and destination are to be chosen by the user from the dropdown menu linked to the input field. Later, he/she has to select the number of halts in the journey which will impact the cost of the ticket. Lastly, the most important factor is the choice of the airline company that the user chooses to travel with. A dropdown menu is attached to for the same. Upon providing all the input fields, and clicking the 'Submit" button, the system enables the user to predict the price of the airline ticket.

IV. METHODOLOGY

Following steps were performed while building the system.

A. Data Collection

Both the training and testing datasets have been extracted from Kaggle data repository. They contain categorical as well as nominal data related to the Indian Airlines from the year 2019. The dataset provides vital information about some impacting features to predict the fare of a flight - such as the places of departures and arrivals, time of departure and arrivals, the route of the flight, the number of halts during the journey and the price of the ticket depending on those features. It's an enormous dataset of 10683 rows and 11 columns (each representing one attribute).

B. Data Pre-processing

While pre-processing the data, we converted the date of journey, departure time and the arrival time from string datatype to date-time object and extracted the numeric values from them; the month-date numeric value from the date of journey attribute and hour-minute numeric value from the departure time and arrival time attributes respectively. Later, we have implemented the 'One hot encoding' method for the nominal categorical data and the label encoding method for ordinal categorical data present in both the training as well as the testing dataset. 'One hot encoding' is a process of converting the categorical data variables into numerical values thus making it suitable to use while implementing machine learning algorithms. One hot encoding method was applied to nominal categorical data attributes such as the 'source', the 'destination' and the 'airline company' chosen by the user.

'Label encoding' helps us convert the labels into numeric values in order to make the dataset suitable for use. Label encoding method was applied to the nominal categorical data attributes such as the 'total number of halts in the journey'. The columns were re - arranged at the last step.

C. Data Cleaning

The null values present in the training dataset where removed. A few columns which were of no use for the feature selection process were deleted from the dataset. The columns of attributes having the categorical data were dropped from the dataset after the new columns containing the numerical values extracted from the pre-processed data were stored for the prediction. Thus, the training dataset suitable for use was obtained and it had the following attribute columns.

Table I
Description of the Attributes

Data Attribute	Description
Total Stops	The number of halts in the journey
Journey Day	The numerical value of 'day' selected from the calendar
Journey Month	The numerical value of 'month' selected from calendar
Dep_hour	The numerical value of 'hour' in departure time
Dep_Min	The numerical value of 'minutes' in departure time
Arrival_hour	The numerical value of 'hour' in arrival time
Arrival_Min	The numerical value of 'minutes' in arrival time
Duration_Hour	The numerical value of 'hours' in duration time
Duration_Min	The numerical value of the minutes in duration time
Airline Company (One hot encoding applied)	Display '1' for the chosen Airline company and display '0' for the rest
Source (One hot encoding applied)	Display '1' for the chosen Source and display '0' for the rest
Destination (one hot encoding applied)	Display '1' for the chosen Destination and display '0' for the rest

D. Generating the Model

The model has been generated using the Random Forest Regression.

E. Presenting the Final Prediction

The user input fields will be provided on a webpage developed using the flask framework. The webpage body was built using HTML5 and the same was styled using CSS3. After the user fills all the required input fields and submits the form, the data will be sent to the generate random forest regression model and the predicted value of the ticket price will be displayed.

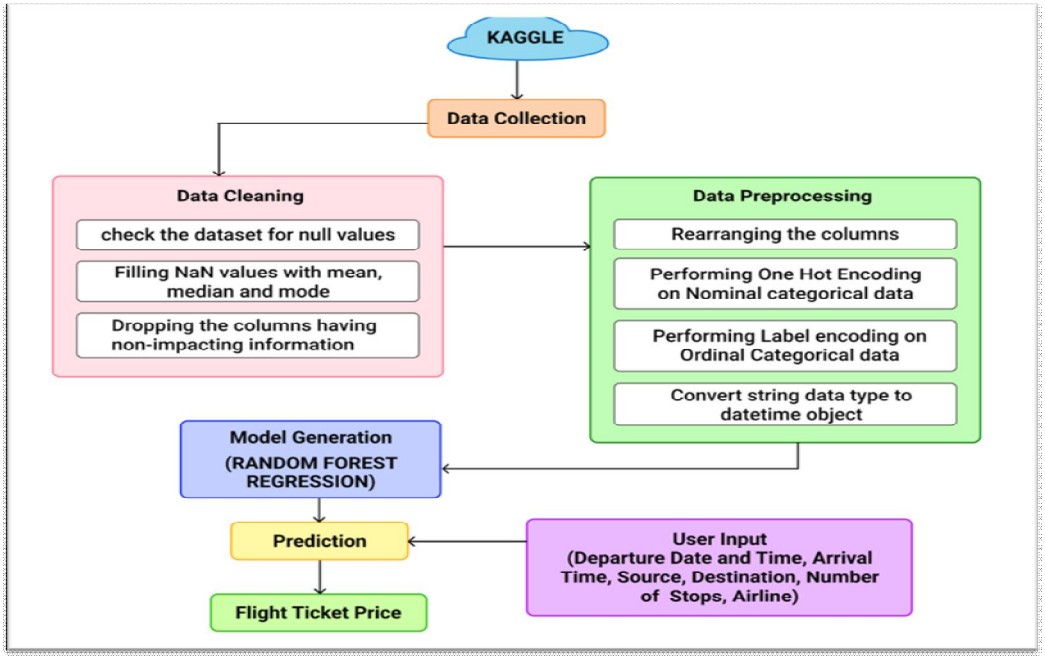


Fig1. Methodology block diagram

V. IMPLEMENTATION

A. *Model*

Random Forest Regression: Random Forest Regression is a supervised learning algorithm that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. It operates by building decision trees during training time and outputting the mean of the classes as the prediction of all the trees.[8]

B. *UI Development*

In this project, Flask framework has been used for the UI development. The main web page of the project takes the required inputs from the user in order to predict the price for the flight. The user inputs required are Departure date and Departure Time, Arrival time of the flight, Source and Destination of the journey, the number of halts during the whole journey and most importantly the airline company which we choose to travel with. After inputting all the fields, the user will click the Submit" button and then the form is submitted. Model enters the scenario at the backend after the submission of the form. The inputs take the help of the historical data and are analysed through supervised machine learning techniques resulting in the prediction of the ticket price. The routing of the pages is done based on the URLs. When the browser finds the '/' in the URL it redirects the user to the home page. After the submission of the form, the user is redirected to the '/result' URL i.e., to the result page where we can see the final result i.e., the prediction of the ticket price. The webpage body was built using HTML5 and the same was styled using CSS3.

VI. RESULTS AND DISCUSSIONS

We are using evaluation metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) and R squared Value for evaluating all the 3 models.

1) Mean Absolute Error (MAE) is the average of difference between the actual data value and the predicted data value. It is calculated as shown below:

$$MAE = (1/n) * \Sigma|y_i - x_i|$$

where:

- **Σ**: A Greek symbol that means "sum"
- **y_i**: The observed value for the i^{th} observation
- **x_i**: The predicted value for the i^{th} observation
- **n**: The total number of observations

[9]

2) Mean Squared Error is the average squared difference between the estimated values and the actual value.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

[10]

Where, n = Data set observations

Y_i = Observation values

$Y\hat{}_i$ = Predicted Values

3) Root Mean Squared Error is the root of MSE

$$\mathrm{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(S_i - O_i)^2}$$

[11]

Where, n = Data set observations

S_i = Predicted values

O_i = Observations

4) R squared value is used for measuring the accuracy of the model.

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

[12]

Where,
R^2 = coefficient of determination
Following are the values of evaluation metrics for the Random Forest Regression Model.

Table III
Values of Evaluation Metrics

Mean Absolute Error (MAE)	1172.6134
Mean Squared Error (MSE)	4044048.9764
Root Mean Squared Error (RMSE)	2010.9820
R Squared Value	81.0258
Accuracy	81%

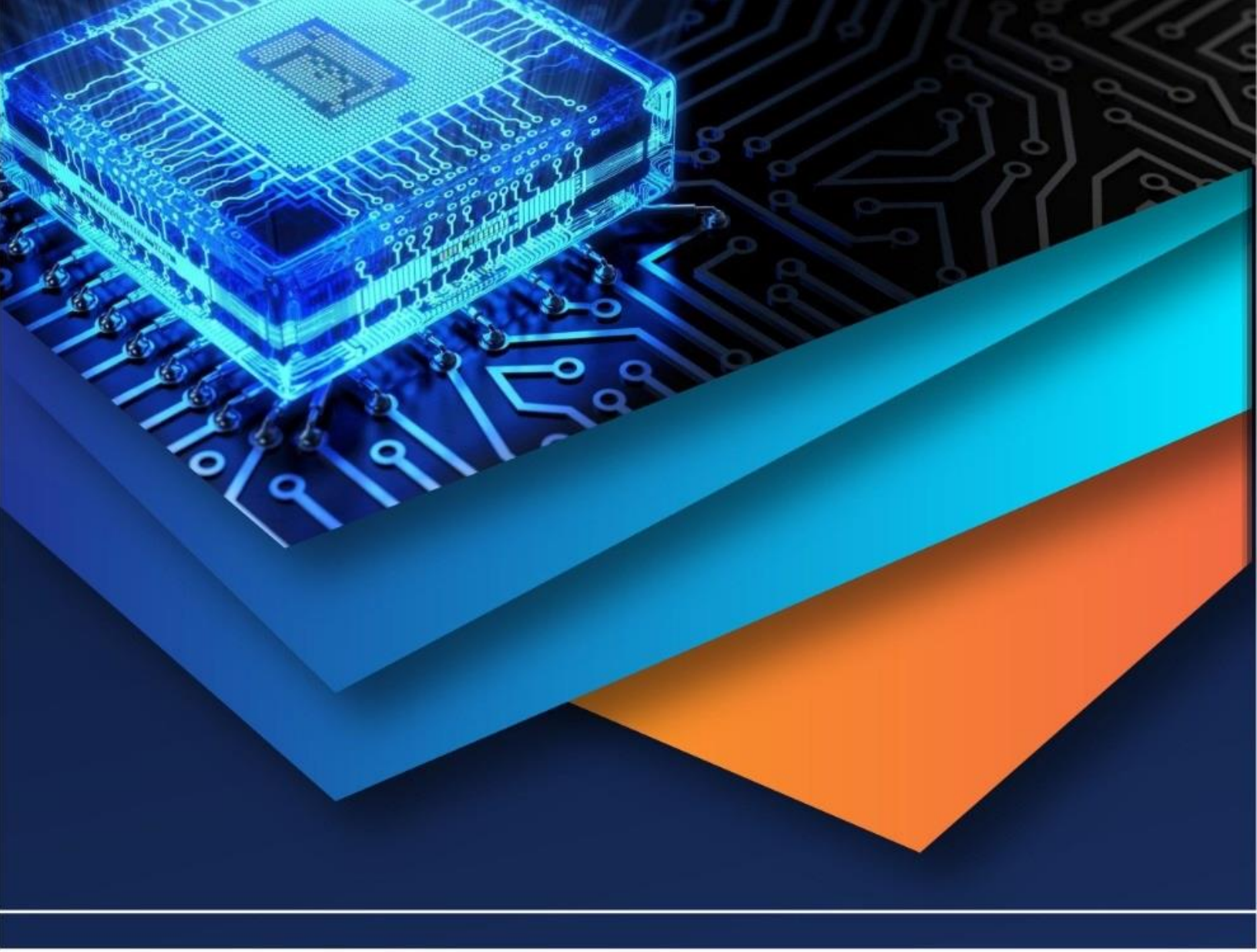
From the table above, we can conclude that the random forest Regression model provides a good accuracy of 81%.

VII. CONCLUSION AND FUTURE SCOPE

For this paper, an extensive study was carried out with dataset collection from Kaggle and Random Forest Machine Learning model was used for deployment. Using visualization, we were able to determine the features which influence airfare prices the most. With experimental analysis, it can be concluded that Random Forest Regression model achieves good accuracy.
The future aim is to work more on the feature selection and model accuracy. We also plan to extend the study by working with larger datasets and greater number of experimentations on the same to procure more accurate airfares which will in turn help users to get an estimated cost of their next airplane travel and can benefit them to make the best deal. We also plan to level up web applications' user interface to provide a premium user experience. We can also consider various other crucial features that affect airplane ticket prices like public holidays, number of luggage, number of hours till departure, crude oil price, etc. in order to get best results. In the near future, there is also a plan to host the web application.

REFERENCES

[1] Tom Chitty, CMBC Business News, "This is how airplanes price tickets", August 3, 2018. Available: https://www.cnbc.com/2018/08/03/how-do-airlines-price-seat-tickets.html
[2] Moira McCormick, BlackCurce, "Behind the Scenes of Airline Pricing Strategies", September 19, 2017. Available: https://blog.blackcurve.com/behind-the-scenes-of-airline-pricing-strategies.
[3] K. Tziridis, Th. Kalampokas, G. A. Papakostas, "Airfare Prices Prediction Using Machine Learning Techniques", 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017.
[4] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen, "A Framework for Airfare Price Prediction: A Machine Learning Approach", 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), September 9, 2019.
[5] Tao Liu, Jian Cao, Yudong Tan, Quanwu Xiao, "ACER: An Adaptive Context- Aware Ensemble Regression Model for Airfare Price Prediction", 2017 International Conference on Progress in Informatics and Computing (PIC), December, 2017.
[6] Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar, "Predicting the price of a flight ticket with the use of Machine Learning algorithms", international journal of scientific & technology research volume 8, December, 2019.
[7] Juhar Ahmed Abdella, Nazar Zaki and Khaled Shuaib, "Automatic Detection of Airline Ticket Price and Demand: A Review", 13th International Conference on Innovations in Information technology (IIT), January 10, 2019.
[8] Chaya Bakshi, Medium, "Random Forest Regression", June 9, 2020, Available: https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
[9] Zach, Statology, "How to calculate mean Absolute Error in Python", January 8. 2021, Available: https://www.statology.org/mean-absolute-error-python/
[10] Wikipedia, "Mean Squared error", Available: https://en.wikipedia.org/wiki/Mean_squared_error
[11] Science Direct, "Root-Mean Squared Error", Available: https://www.sciencedirect.com/topics/engineering/root-mean-squared-error/
[12] NCL.AC.UK, "Coefficient of Determination, R-squared", Available: https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY