Volume: 12 Issue: IX Month of publication: September 2024

DOI: https://doi.org/10.22214/ijraset.2024.64350

Comparison of Air Fare Prediction Using Various Machine Learning Techniques

Lakshmi Srinivas Dendukuri[1], Shaik Jakeer Hussain[2]
Department of Electronics and Communication Engineering, Vignan's Foundation for Science, Technology and Research, Guntur, Andhra Pradesh, India

Abstract: Due to a rise in air service links globally, air travel is now a popular, significant, and faster form of transportation. Airline rates are thought to be impacted by several factors and are subject to frequent fluctuation, making estimation of these rates a challenging but essential task. The airline uses a pricing structure for the plane ticket. The cost of airline tickets varies based on the time of day, according to the survey. Even the weekends, the tourist season, and the pageant season have an impact on this. An aircraft ticket's price is influenced by several different elements. The buyer only has access to a limited amount of information, which is truly insufficient to estimate flight pricing, but the seller is aware of all the requirements. In this study, we use machine learning regression approaches such as K Nearest Neighbourhood, Decision Tree, and Random Forest to estimate flight fares based on basic data including source, destination, departure time and date, arrival timings, halts, and airline type. The results of the investigation show that the Random Forest Regression Model yields extremely optimal results.

Keywords: Machine Learning (ML), K Nearest Neighbourhood, Decision Tree and Random Forest.

I. INTRODUCTION

Everyone knows of that the vacations always need a well-deserved break, and organizing the trip may be an exhausting work [1]. The global proliferation of the internet and e-commerce has led to a massive growth in the commercial aviation industry, transforming it into an organized marketplace. As a result, several tactics such as consumer profiling, commercial marketing, and social considerations are utilised to establish ticket prices for airline corporate finance [2]. When tickets are ordered months ahead of time, airfares are frequently affordable, but they soar when tickets are bought last minute. However, the number of hours till departure might not be the only element that influences airline prices, and to name a few.

To avoid the effects of its most severe charge, the airline ticket purchasing procedure is to acquire a ticket several days ahead to aircraft take-off [3]. The majority of aircraft routes oppose this practise. Plane companies may reduce ticket prices because they need to establish a business and when tickets are more difficult to come by. They might be able to get as much bang for their buck. As a result, the price may be affected by a variety of factors. To predict expenses, this company use artificial intelligence to provide potential flight paths after a period of time. Every organisation has the authority and ability to adjust ticket prices at any moment. Somebody who book plane ticket on a daily basis might be able to forecast that optimal moment to buy a ticket in order to get all the greatest price. For financial management, several airlines modify ticket pricing. Explorers can save money by buying the ticket at the lowest price. Those who fly by plane on a regular basis are familiar with price swings. For the implementation of different assessing systems, airways utilise sophisticated financial management rules. As a consequence, the assessing system adjusts the price based on the time, seasons, and holiday dates to update top and bottom for subsequent page. Because eventual goal of the airlines helps to make a profit, while the client is looking the best deal. Purchasing tickets much in advance of the intended departure is a common attempt by buyers to avoid price hikes as the arrival day draws near. Unfortunately, it isn't true. The consumer may end up donating more than is necessary.

The speed at which prices can fluctuate is evident for anyone who has ever bought a plane ticket [3]. Aircraft employ novel sales management technologies to execute a distinctive price strategy. The price of a ticket can go up or down, and the lowest ticket that is still available may not always be the same. This method of valuation modifies the toll based on the time of day, for example, mornings, noon, or evenings. There may also be seasonal pricing differences during the winter, summer, and holiday seasons. The customer is searching for the greatest offer, whereas the company's main goal is to maximize income. Typically, purchasers aim to secure tickets well in advance of the travel date. People sometimes assume that prices will increase as the departure date draws near, but this isn't necessarily the case. Customers can wind up paying more than they ought to for an equivalent seat. Airlines try to keep flight ticket prices low in order to increase earnings in today's market. The majority of frequent flyers are aware of the optimum times to get discounted tickets.

Many clients who are just not competent at booking tickets, on the other hand, slip into the industry's discount scam, prompting them to waste their money. Airlines' primary purpose is to generate an income, while customers are shopping for the greatest deal. Buyers usually attempt to buy tickets long in advance of the travel date to avoid price hikes as the arrival day approaches. Because the intricacy of airline fare structures its indeed incredibly hard for the client to purchase a plane ticket at a quite cheap price since the cost fluctuates continually. When a market is needed and tickets are difficult to come by, airlines might reduce ticket costs. These strategies take into account of economic, promotional, economic, and social elements that affect final airline price. They could try to maximise the earnings. As just a conclusion, numerous factors may impact expenses. As the airlines' pricing models are so complicated, prices vary often, making it impossible for passengers to purchase tickets at extremely cheap costs. Over the previous couple of decades, consumer and airline surveys have increased steadily. Establishing a cheap price or a favourable time to acquire a ticket is a crucial question from of the customer's perspective. Passengers struggle extremely hard to obtain great & lowest tickets bargain due to the airline industry's complicated pricing methodology. Deep Learning and ML-based approaches and modals had created to overcome the challenge, as well as considerable study. This study discusses a Machine Having to learn Airline Price Prediction System that use Random Forest Regression to forecast flight ticket pricing.

II. LITERATURE REVIEW

In, K. Tziridis et al., devised a technique for predicting airline price [4]. The chapter starts with some basic knowledge about ML before moving on to the research methods, which includes four main stages of Feature Extraction that impacts flight costs, data collection from Airlines, variety of an exact Machine Learning Regression model, and assessment. They acquired airline information from internet for a particular Greek airplane firm and demonstrated that previous price data may be used to estimate trip pricing. Overall results from the experiment demonstrate that machine learning models are just a good model for estimating flight costs. Data gathering and extraction of features, out of which they extracted some beneficial findings, are also essential aspects in ticket price prediction. Based on the results of the tests, they were able to determine which characteristics have the most impact on price prediction. Other characteristics, in addition to the ones chosen, might help enhance prediction performance. The technique might be expanded in the long term to anticipate ticket rates for the whole travel plan of the United States. Naresh Alapati et al., developed a model for air fare prediction using Random Forest classifier. The classifier provided good results and also the authors have displayed the univariate distribution of data [5]. Naveen Prasanth et al., also predicted the air fare using the K Nearest Neighbourhood regression technique [6]. Ankita Panigrahi constructed a frame work that predicts the air fare by using Neural Networks, Decision Tree, Linear Regression and Random Forests classifiers [7]. Ratnakanth developed a model for prediction of air fare using deep Neural Network architecture. The data is also visualized and produces better results using random forest and gradient boosting techniques [8].

III. METHODOLOGY

A. Data Collection

The datasets of both test data and training were obtained out from Kaggle data source. They include both detailed and general information about Airlines of the India for the era 2018. The database giving critical information on several factors that influence flight fares, such as departure and arrival locations, departure and arrival times, flight route, the number of stops throughout the journey, and the cost of the ticket based on these variables.

B. Data Preprocessing

We converted the date when the user wants to travel, time of departure, and estimated time of arrival from string type of data to separate date and time format and then extract the integer data from them and from the time of departure and time of arrival attributes, we took hours and minutes data separately. One hot encoding was a technique that converts text type of data into binary data, making them acceptable for use in machine learning systems. One hot encoding approach was used on actual qualitative data characteristics such as the user- selected airline company, source, destination. 'Label encoding' assists us in converting the labelled data into integer data so that the dataset may be used. The label encoding approach was used on label type of data such as the "overall count of stops on the voyage." At the end, the columns will be reorganized.

C. Data Cleaning

The blank data with in trained data will be deleted. Only couple of entries that were useless for such as feature selection procedure has been omitted.

The fields of characteristics containing qualitative facts were removed in the database just after corresponding fields carrying integer data generated out from the previous data which is saved for forecasting. As a result, an adequately trained dataset will be obtained.

D. Classifier

Classifiers are used in developing models by training the network or model with features and labels. There are many classifiers in predicting the value from its features. In this work, we used Decision Tree, K Nearest Neighbourhood and Random Forest classifiers for predicting the fight price based on the features like time of travel, date of travel and various features.

1) Decision Tree

Decision Tree was among the most extensively used and effective ways of supervised learning. It can tackle Regression and Classification issues, with the later proving most useful. This is indeed a classifier that has 3 different kinds of elements inside a tree - like structure. The Source Node represents the complete data and may be partitioned. Branch nodes indicate decision criteria, whereas interior nodes reflect set of data attributes. Ultimately, these Leaf Nodes describe the output. This method is highly helpful when handling various decision-making problems.

2) Random Forest

Random Forest was a very well-known supervised ML technique. It's indeed useful to machine learning regression and classification issues. This is built on the notion of supervised methods, and that is the method of mixing numerous classifiers to tackle a issue and to increase effectiveness. Random Forest is indeed a classification algorithm that employs a variety of decision trees on distinct segments in a given data and mean the overall findings to increase the datasets forecasting accuracy.

3) K- Nearest Neighbourhood

KNN is an acronym of "Kth-Nearest Neighbour." This is a supervised machine learning method. This technique can tackle classification as well as regression issues. The Kth-Nearest Neighbour method assumes closeness among both incoming and existing cases. It keeps the incoming data in the group which is closer to the original groups. The K- Nearest Neighbour methodology would be utilized in both classification and regression problems; however, this is more commonly utilised for classification tasks. The sign 'K' denotes the count of nearby peers surrounding a distinct uncertain parameter which could get estimated or categorised. This is commonly known as a slow learning technique as it does not instantly understand out from trained data set; instead, it maintains the database and then commits an act on it during classification.

IV. RESULTS

We are now evaluating our regression models with the metrics like Root Mean absolute error, Mean squared error, mean absolute error.

A. Mean Absolute Error

The mean itself indicates the average of the difference between the actual and the predicted values present in our data sets. It can be evaluated from the formula:

$$\text{Mean}=\frac{1}{n}\sum | y_i - x_i| \quad -(1)$$

n=how many numbers of observations.

y_i indicates the observed value for i^{th} observation,

x_i indicates the predicted value for i^{th} observation.

B. Mean Squared Error

Mean squared error can be defined as the average of squared difference between the actual and predicted data values.

$$\text{MSE}=\frac{1}{n}\sum_{i=1}^{n} |a_i - b_i|^2 \quad -(2)$$

a_i is observed value for i^{th} observation

b is the predicted value for i^{th} observation
n is number of observations used from data sets.

C. Root Mean Squared Error

Root mean squared error is defined as the square root of average of squared difference between the actual and predicted values. RMSE can be calculated as:

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}|a_i - b_i|^2} \qquad \text{-(3)}$$

D. R Squared Value

R squared value can be defined as how accurately we trained our model.

$$R^2=\frac{1-\sum(a_i-b_i)^2}{\sum(a_i-b_i)^2} \qquad \text{-(4)}$$

Following are the values for metrics evaluation for the three regression models such as random forest, k-nearest neighbour, decision tree model. In figure 1, the boxplot represents the features and their mean, median values. In figure 2, the graph represents the test input and predicted output values. Most of the test and predicted values are similar providing less error in prediction by the model.

TABLE I: Metric values evaluated for k-Nearest Neighbour regression model.

Mean Absolute Error	1100.252
Mean Squared Error	4113794.119
Root Mean Squared Error	2028.2490
R Squared value	0.809
Accuracy	80%

TABLE II: Metric values evaluated for Decision Tree model.

Mean Absolute Error	756.269
Mean Squared Error	2668391.201
Root Mean Squared Error	1633.521
R Squared value	0.871
Accuracy	87%

TABLE III: Metric values evaluated for Random Forest model.

Mean Absolute Error	662.06
Mean Squared Error	2302761.683
Root Mean Squared Error	1517.483
R Squared value	0.889
Accuracy	89%

From the result analysis it is clear that Random Forest is outperforming than other regression models with accuracy of 89%. Air India and Jet Airways are full-service airlines that are usually expensive owing to the multiple services they facilitate. Cheaper-cost airlines such as Indigo and SpiceJet provide reasonable and comparable fares. The advantage of this study is that travellers may examine the peak cost to low cost in order to understand the pricing variations across airlines. Passengers may post their flight reviews on each airline's website so that other passengers might benefit from them.

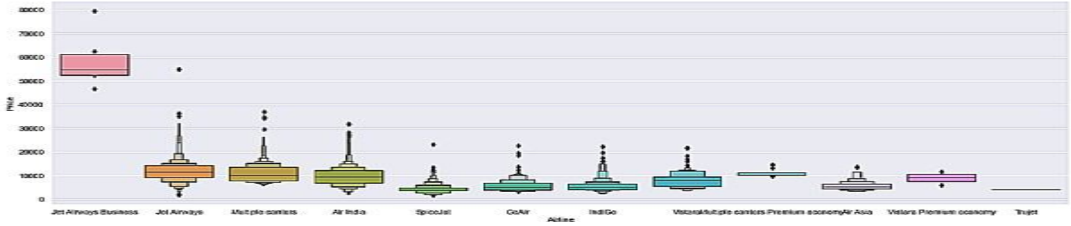


Fig: 1 Representation of various features in boxplot.

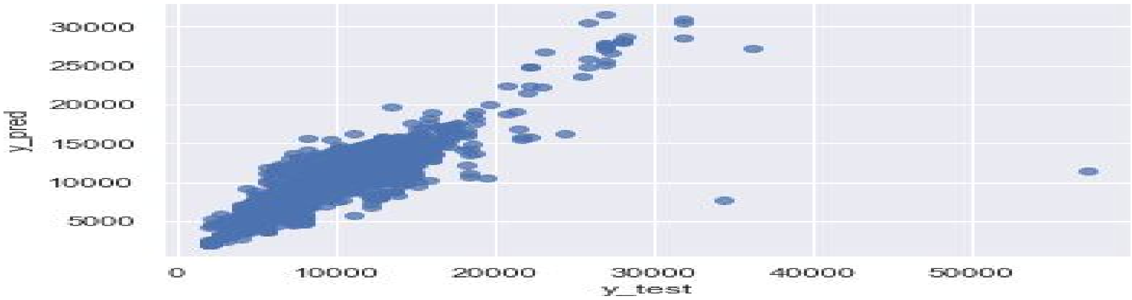


Fig: 2 Prediction value and test value graph.

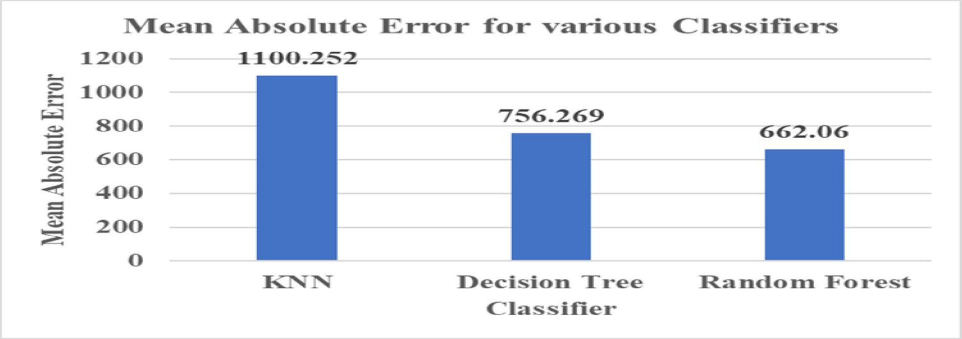


Fig: 3 Mean Absolute Error for classifiers.

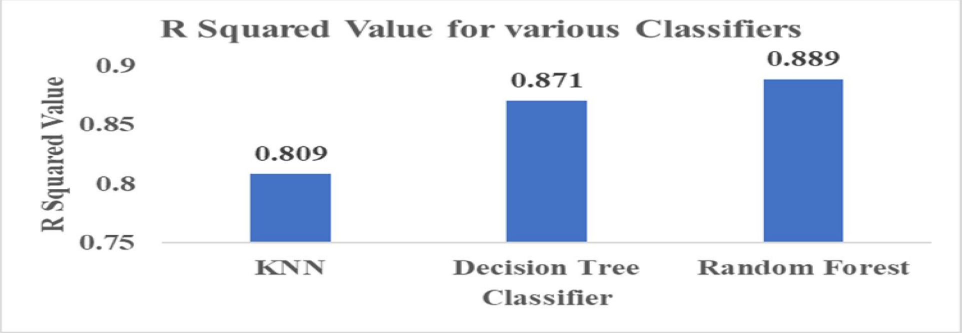


Fig: 4 R squared value of classifiers.

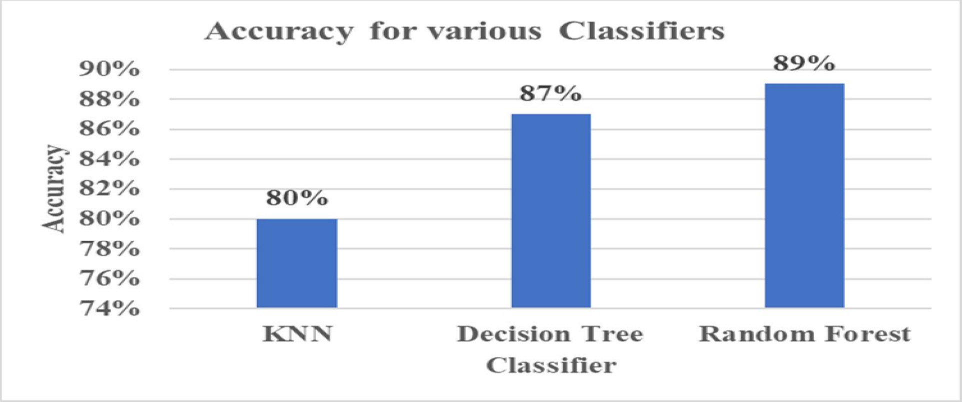


Fig: 5 Accuracy of classifiers.

From figures, 4,5 and 6, it is understood that the Random Forest classifier provides better accuracy compared with K Nearest Neighbourhood and Decision Tree classifiers.

The mean absolute error is less and R squared value and accuracy are high using Random Forest Classifier. The Random Forest graph is drawn using Y-test data, and it demonstrates that the cost was increasing. According to the projection, the users purchased the highest priced ticket are approximately thirty thousand, and the majority people are purchasing tickets ranging from three thousand to seventeen thousand. The benefit of this research is that consumers will have a better understanding of the regular price offers in festivals and user can select the best trip fare.

V. CONCLUSION

An extensive study was conducted for this paper, with datasets being collected from Kaggle. We are trying to determine the factors that have the greatest impact on airfare prices by using visualisation. We are using r2 score for predicting accurate airline fares and providing accurate value of airfare price. We have calculated RMSE, R^2 value, MSE in order to know how our regression model will fit to our data. Based on the results of this experiment, it is possible to sum up that the Random Forest has a high level of precision. As a result, the Random Forest performs admirably in predicting airfare price. Further data, like real seat availability, might be gathered, improving the accuracy of the anticipated outcomes. The objective is to concentrate high on selection of parameters and precision. We would like to enhance research by dealing with huge database and numerous trials to achieve more accurate flight prices, allowing customers to get an anticipated cost before their next flying journey and assist them make good bargain.

REFERENCES

[1] Tom. Chitty, “This is how airlines price tickets,” Retrieved from https://www.cnbc.com/2018/08/03/how-do-airlines-price-seat-tickets.html. Accessed on 23rd July 2024.

[2] McCormick. M, “Behind the scenes of airline pricing strategies,” Retrieved from https://blog.blackcurve.com/behind-the-scenes-of-airline-pricing-strategies. Accessed on 24th July 2024.

[3] B. Smith, J. Leimkuhler, R. M. Darrow, “Yield management at American airlines,” Interfaces 22, vol. 1, pp: 8-31, 1992.

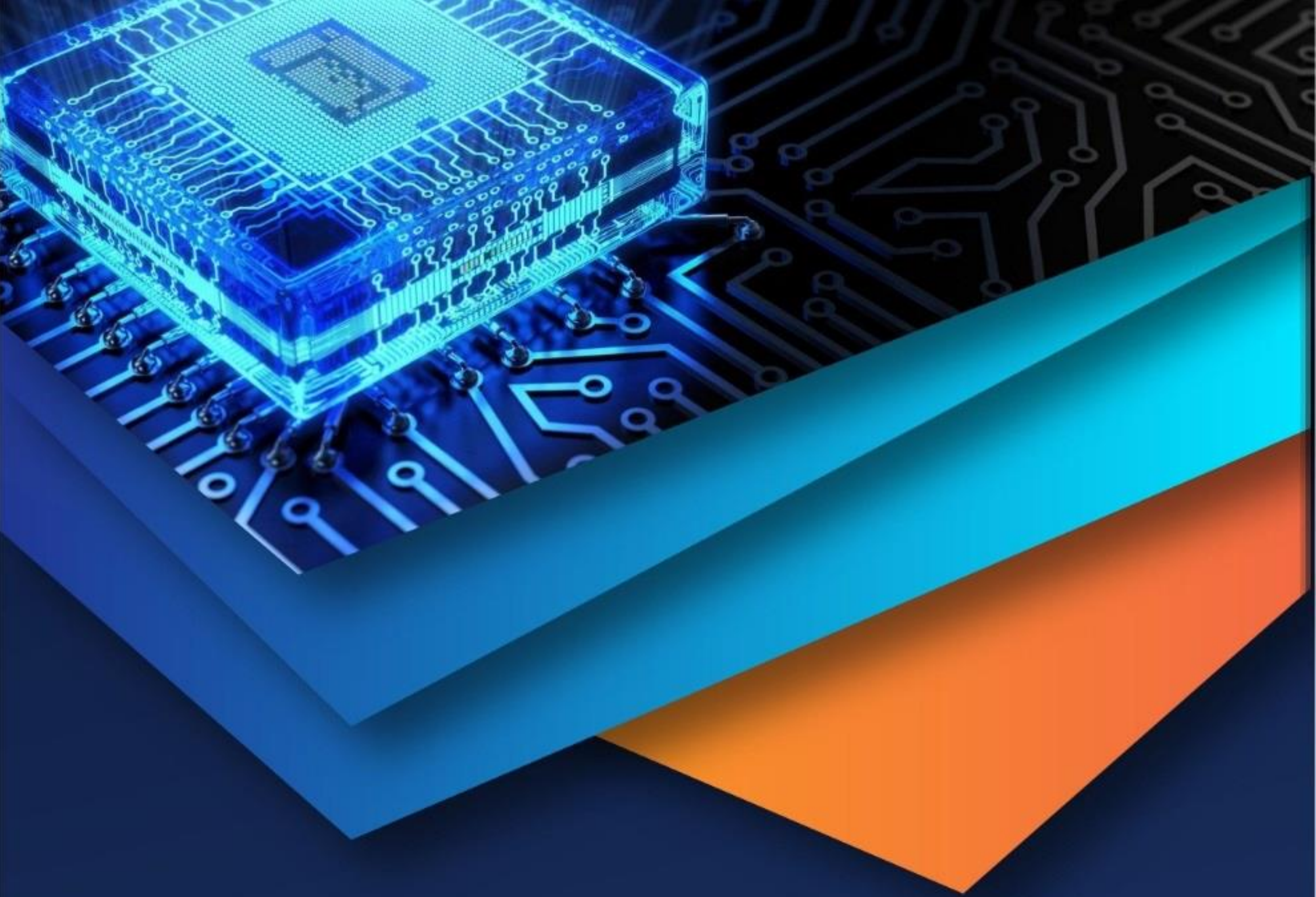
[4] K. Tziridis, Th. Kalampokas, G. A. Papakostas, “Airfare Prices Prediction Using Machine Learning Techniques,” In Proc: 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, 2017, pp: 1036-1039.

[5] Alapati. Naresh, B. V. V. S. Prasad, Aditi Sharma, G. R. P. Kumari, S. V. Veeneetha, N. Srivalli, T. Udaya Lakshmi, D. Sahitya. "Prediction of Flight-fare using machine learning," In Proc: International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP), Uttarakhand, India, 2022, pp. 134-138.

[6] Prasath, S. Naveen, Sherin Eliyas. "A Prediction of Flight Fare Using K-Nearest Neighbors," In Proc” 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1347-1351.

[7] Panigrahi. Ankita, Rakesh. Sharma, Sujata. Chakravarty, Bijay. Paikaray, Harshvardhan. Bhoyar, "Flight Price Prediction Using Machine Learning," In ACI@ ISIC, Savannah, United States, 2022, pp. 172-178.

[8] Ratnakanth. G, "Prediction of flight fare using deep learning techniques" In Proc: 2022 International Conference on Computing, Communication and Power Technology (IC3P), Vishakapatnam, India, 2022, pp. 308-313.



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL
IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY