iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 Issue: IV Month of publication: April 2023

DOI: https://doi.org/10.22214/ijraset.2023.49949

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue IV Apr 2023- Available at www.ijraset.com*

# Development of Machine Learning Based Flight Price Prediction System for Indian Market

Aarish Siddiqui[1], Shubham Yadav[3], Suraj Yadav[4], Akash Singh[5]
*Department of Information Technology, Shree L.R Tiwari College of Engineering, (Mumbai University) Mumbai, India*

***Abstract:** Airlines usually keep their price strategies as commercial secrets and information is always asymmetric, it is difficult for ordinary customers to estimate future flight price changes. However, a reasonable prediction can help customers make decisions when to buy air tickets for a lower price. Flight price prediction can be regarded as a typical time series prediction problem. When you give customers a device that can help them save some money, they will pay you back with loyalty, which is priceless. Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. Features are extracted from the collected data to apply Random Forest Machine Learning (ML) model. Then using this information, we are intended to build a system that can help buyers whether to buy a ticket or not. We have used Random Forest Algorithm which is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. With that said, random forests are a strong modelling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.*

***Keywords:** Airline strategies, Airfare price prediction, Airfare changes, Random Forest algorithm, predictive model.*

## I. INTRODUCTION

With the worldwide growth of internet and E-commerce, commercial aviation industry has witnessed a tremendous growth and has become a regulated marketplace. 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. 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. Any individual who has booked a flight ticket previously knows how dynamically costs change. Aircraft uses advanced strategies called Revenue Management to execute a distinctive valuing strategy [1]. The least expensive accessible ticket changes over a period the cost of a ticket might be high or low. This valuing method naturally modifies the toll as per the time like morning, afternoon or night. Cost may likewise change with the seasons like winter, summer and celebration seasons. The extreme goal of the carrier is to build its income yet on the opposite side purchaser is searching at the least expensive cost. Purchasers generally endeavor to purchase the ticket in advance to the takeoff day. Since they trust that airfare will be most likely high when the date of buying a ticket is closer to the takeoff date, yet it is not generally true. Purchaser may finish up with the paying more than they ought to for a similar seat.

### A. Motivation

Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket.

### B. Problem Formulation

A report says India's affable aeronautics industry is on a high- development movement. India is the third-biggest avionics showcase in 2020 and the biggest by 2030. Indian air traffic is normal to cross the quantity of 100 million travelers by 2017, whereas there were just 81 million passengers in 2015.

Agreeing to Google, the expression "Cheap Air Tickets" is most sought in India. At the point when the white collar class of India is presented to air travel, buyers searching at modest costs. The rate of flight tickets at the least cost is continuouslyexpanding.

## II. PROPOSED FRAMEWORK

Our suggested approach makes use of datasets to forecast airfare at the business segment levels. Fig 1 depicts a high-level view of the project framework's primary components. During the data pre-treatment stage, all databases are cleaned to remove any potentially erroneous examples, then converted and integrated depending on market group. The feature extractor extracts and generates handmade attributes that are intended to describe the segment of the market. The goal of adaptive filtering modules is to improve accurate channels by assessing the utility of the characteristics and removing any unnecessary characteristics. Finally, we use the selected criteria to build our forecasting techniques, that result in the finished product of the projected airline cost of the ticket.

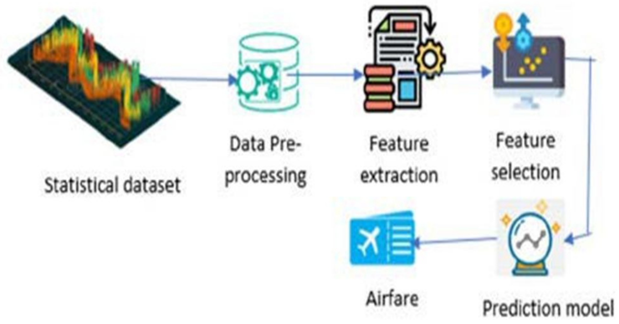


Fig. 1. Proposed framework for airfare price prediction.

Use case diagram for proposed system

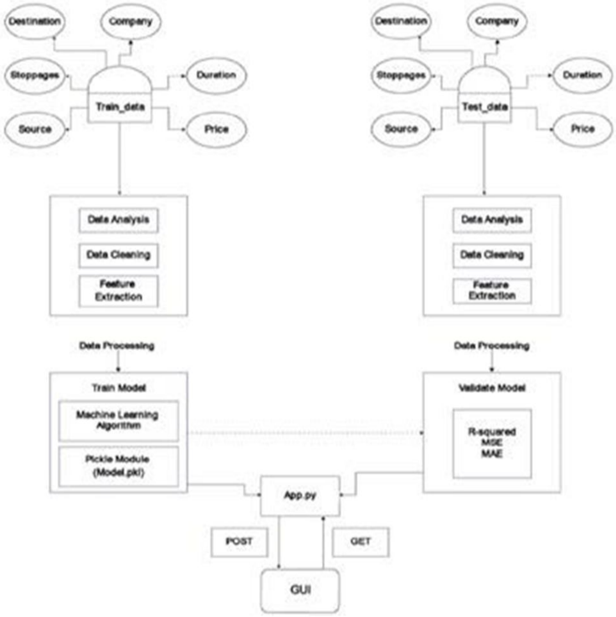


Fig 2. Use Case Diagram

## III. METHODOLOGY

Random Forest Algorithm can be used for both Classification and Regression problems in ML. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random-forest classifier:

*1)* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
*2)* The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest :

*a)* It takes less training time as compared to other algorithms.
*b)* It predicts output with high accuracy, even for the large dataset it runs efficiently.
*c)* It can also maintain accuracy when a large proportion of data is missing.

*3)* The below diagram explains the working of the Random Forestalgorithm:

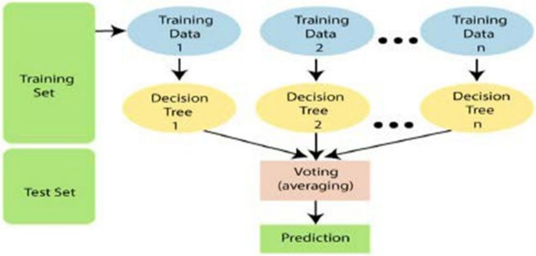


Fig 3. Random Forest Algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. The Working process can be explained in the below steps and diagram shown above:

*a)* Step-1: Select random K data points from the training set.
*b)* Step-2: Build the decision trees associated with the selected data points (Subsets).
*c)* Step-3: Choose the number N for decision trees that you want to build.
*d)* Step-4: Repeat Step 1 & 2.
*e)* Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

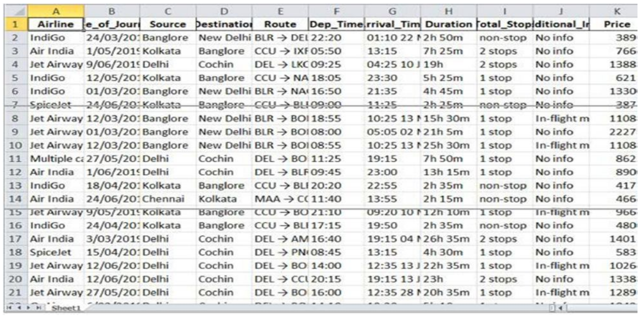
## IV. IMPLEMENTATION DETAILS

### *A. Dataset*

We are using the dataset which consist of flight tickets for various airlines between the months of March and June of 2019 and between various cities. Size of training set: 10683 records and Size of test set: 2671 records.

### *B. Feature Selection*

In the feature selection, we can find out the best feature which will contribute and have good relation with target variable. Some of the feature selection methods are Heatmap, Feature Importances. In this we can find out which feature of data is independent and dependent, also to find out which feature contains the most importance we can do this by using the Extra tree Regressor.

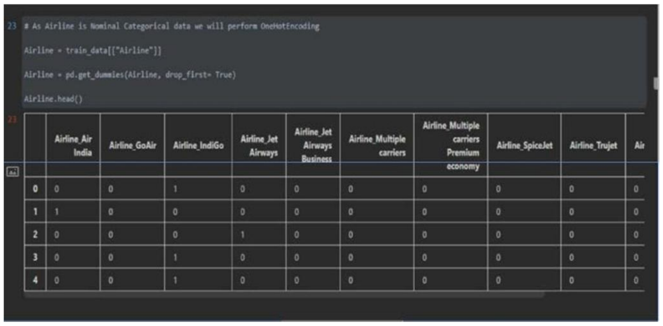


| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Airline | e_of_Jouri | Source | Destinatio | Route | Dep_Time | rrival_Tim | Duration | Total_Stop | ditional_Ir | Price |
| 2 | IndiGo | 24/03/20: | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 M | 2h 50m | non-stop | No info | 389 |
| 3 | Air India | 1/05/2019 | Kolkata | Banglore | CCU → IXF | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 766 |
| 4 | Jet Airway | 9/06/2019 | Delhi | Cochin | DEL → LKC | 09:25 | 04:25 10 J | 19h | 2 stops | No info | 1388 |
| 5 | IndiGo | 12/05/20: | Kolkata | Banglore | CCU → NA | 18:05 | 23:30 | 5h 25m | 1 stop | No info | 621 |
| 6 | IndiGo | 01/03/20: | Banglore | New Delhi | BLR → NA | 16:50 | 21:35 | 4h 45m | 1 stop | No info | 1330 |
| ~~7~~ | ~~SpiceJet~~ | ~~24/06/20:~~ | ~~Kolkata~~ | ~~Banglore~~ | ~~CCU → BLI~~ | ~~09:00~~ | ~~11:25~~ | ~~2h 25m~~ | ~~non-stop~~ | ~~No info~~ | ~~387~~ |
| 8 | Jet Airway | 12/03/20: | Banglore | New Delhi | BLR → BOI | 18:55 | 10:25 13 M | 15h 30m | 1 stop | In-flight m | 1108 |
| 9 | Jet Airway | 01/03/20: | Banglore | New Delhi | BLR → BOI | 08:00 | 05:05 02 M | 21h 5m | 1 stop | No info | 2227 |
| 10 | Jet Airway | 12/03/20: | Banglore | New Delhi | BLR → BOI | 08:55 | 10:25 13 M | 25h 30m | 1 stop | In-flight m | 1108 |
| 11 | Multiple ca | 27/05/20: | Delhi | Cochin | DEL → BO | 11:25 | 19:15 | 7h 50m | 1 stop | No info | 862 |
| 12 | Air India | 1/06/2019 | Delhi | Cochin | DEL → BLF | 09:45 | 23:00 | 13h 15m | 1 stop | No info | 890 |
| 13 | IndiGo | 18/04/20: | Kolkata | Banglore | CCU → BLI | 20:20 | 22:55 | 2h 35m | non-stop | No info | 417 |
| 14 | Air India | 24/06/20: | Chennai | Kolkata | MAA → CC | 11:40 | 13:55 | 2h 15m | non-stop | No info | 466 |
| 15 | Jet Airway | 9/05/2019 | Kolkata | Banglore | CCU → BO | 21:10 | 09:20 10 M | 12h 10m | 1 stop | In-flight m | 966 |
| 16 | IndiGo | 24/04/20: | Kolkata | Banglore | CCU → BLI | 17:15 | 19:50 | 2h 35m | non-stop | No info | 480 |
| 17 | Air India | 3/03/2019 | Delhi | Cochin | DEL → AM | 16:40 | 19:15 04 M | 26h 35m | 2 stops | No info | 1401 |
| 18 | SpiceJet | 15/04/20: | Delhi | Cochin | DEL → PN | 08:45 | 13:15 | 4h 30m | 1 stop | No info | 583 |
| 19 | Jet Airway | 12/06/20: | Delhi | Cochin | DEL → BO | 14:00 | 12:35 13 J | 22h 35m | 1 stop | In-flight m | 1026 |
| 20 | Air India | 12/06/20: | Delhi | Cochin | DEL → CCL | 20:15 | 19:15 13 J | 23h | 2 stops | No info | 1338 |
| 21 | Jet Airway | 27/05/20: | Delhi | Cochin | DEL → BO | 16:00 | 12:35 28 M | 20h 35m | 1 stop | In-flight m | 1289 |

Fig 4. Dataset

### *C. Data Pre-Processing*

Before building model, the data should be properly preprocessed and converted to quality, clean data even the resulting machine learning model will be of great quality. The data pre-processing includes three main parts that is data integration, data cleaning, data transformation. In data integration the data collected from various sources are integrated. In data cleaning process the data containing the null values, unnecessary rows with null values are being cleared. The data transformation includes the feature scaling. A good data pre-processing in machine learning is the most important factor that can make a difference between a good model and a poor machine learning model. So, we need to do pre-process the data to get the perfect accuracy.



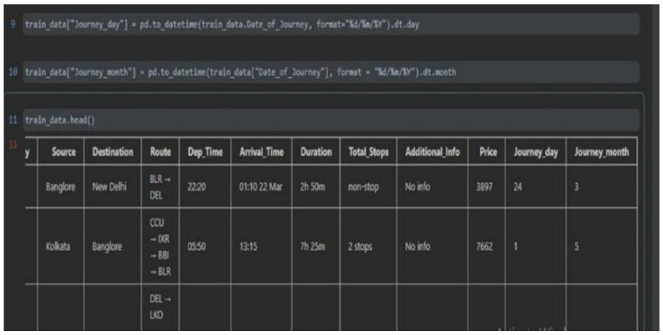
```
# As Airline is Nominal Categorical data we will perform OneHotEncoding

Airline = train_data[["Airline"]]

Airline = pd.get_dummies(Airline, drop_first= True)

Airline.head()
```

| | Airline_Air India | Airline_GoAir | Airline_IndiGo | Airline_Jet Airways | Airline_Jet Airways Business | Airline_Multiple carriers | Airline_Multiple carriers Premium economy | Airline_SpiceJet | Airline_Trujet | Air |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig 5. Pre-processing

### *D. Feature Extraction*

In this the features which are provided in dataset are converted into the easier form that is it means transforming raw data into a feature vector which helps in building the model. Making the data or the feature in its easier format would help to train our model easily also it provides a good accuracy rate.



```
train_data["Journey_day"] = pd.to_datetime(train_data.Date_of_Journey, format="%d/%m/%Y").dt.day
```

```
train_data["Journey_month"] = pd.to_datetime(train_data["Date_of_Journey"], format = "%d/%m/%Y").dt.month
```

```
train_data.head()
```

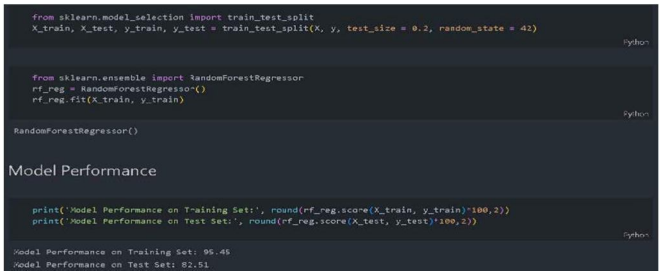
| y | Source | Destination | Route | Dep_Time | Arrival_Time | Duration | Total_Stops | Additional_Info | Price | Journey_day | Journey_month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 Mar | 2h 50m | non-stop | No info | 3897 | 24 | 3 |
| | Kolkata | Banglore | CCU → IXR → BBI → BLR | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 7662 | 1 | 5 |
| | | | DEL → LKO | | | | | | | | |

Fig 6. Feature Extraction

### *E. Fitting the Model*

For fitting of the model we split dataset into train and test set in order to prediction. After splitting, you will train the model on the training set and perform the predictions on the test set. After training, check the accuracy using actual and predicted values. Once the training of the model is done, we can store that model using pickle file so that we can reuse it again. Thus, you can predict the price.

### *F. Model Performance*



```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

Python

```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor()
rf_reg.fit(X_train, y_train)
```

Python

```
RandomForestRegressor()
```

Model Performance

```
print('Model Performance on Training Set:', round(rf_reg.score(X_train, y_train)*100,2))
print('Model Performance on Test Set:', round(rf_reg.score(X_test, y_test)*100,2))
```

Python

```
Model Performance on Training Set: 95.45
Model Performance on Test Set: 82.51
```

Fig 7. Model Performance

## V. CONCLUSION

The various machine learning algorithms such as Linear Regression (LR), Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors are used in the previous systems for predicting the flight ticket price. In our ML based system, we used the Random Forest Algorithm which gives more accuracy in predicting the airfare. Features such as departure time, the number of days left for departure and time of the day etc, will be used for predicting the flight ticket price. This system also helps the customers to buy the flight ticket at lower price. The system is easy to use and it gives more accuracy in prediction. Also, the time required for prediction is less and which helps the customers to get price quickly. This saves the time of the customers. Also, getting the flight ticket price in advance will help the customer in decision making whether to buy the ticket or not according to their convenience. In future work the system will be more expanded by adding more routes and more work need to be done on the GUI enhancement. We also plan to level up web applications' other input and data validations 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, crude oil price, etc. in order to get best results.

## REFERENCES

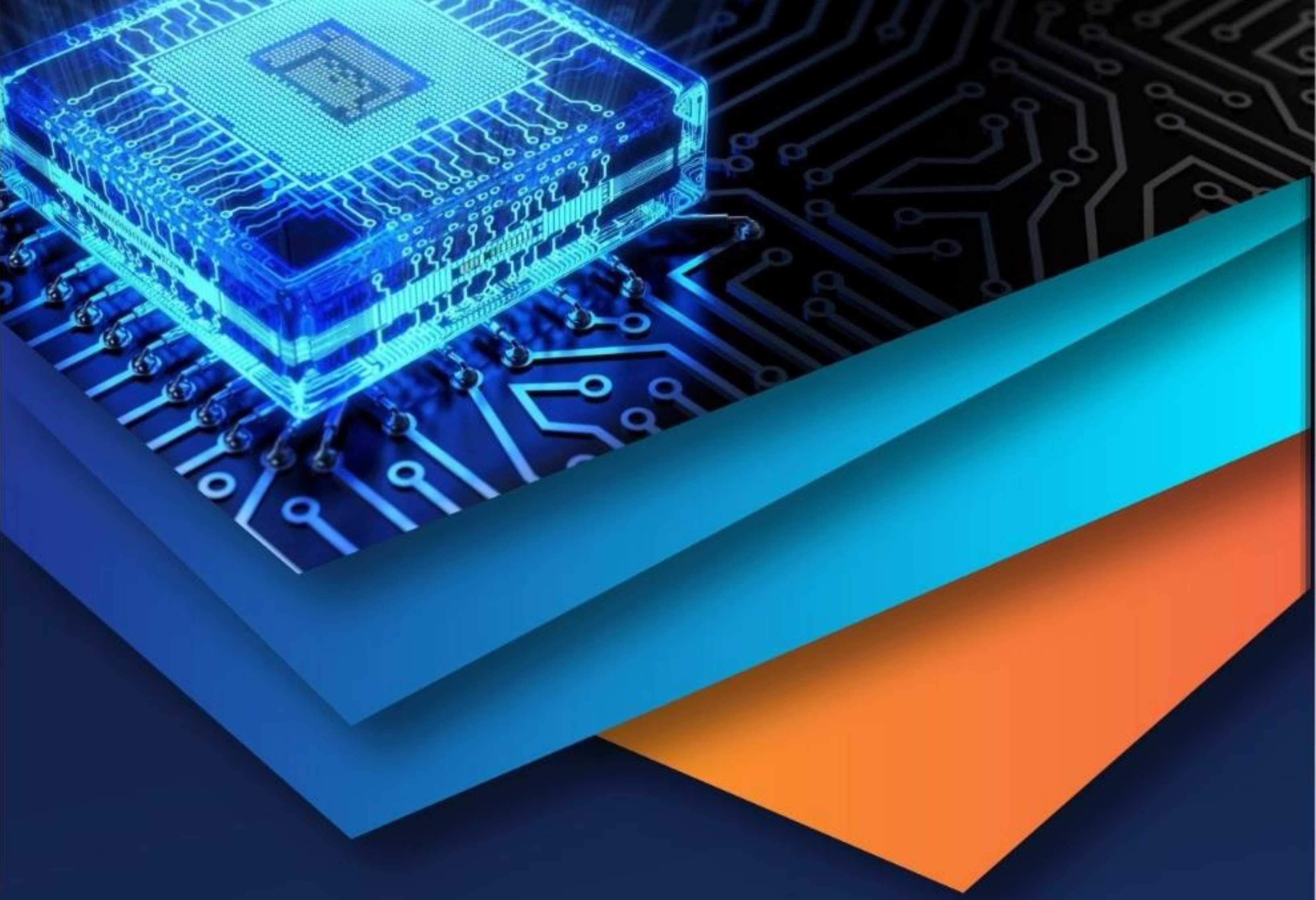
[1] Tom Chitty, CMBC Business News, "This is how airplanes price ticket https://www.cnbc.com/2018/08/03/how-do-airlines-price-seat- tickets.html.

[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.



10.22214/IJRASET

INDEX COPERNICUS

45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

ISI

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY