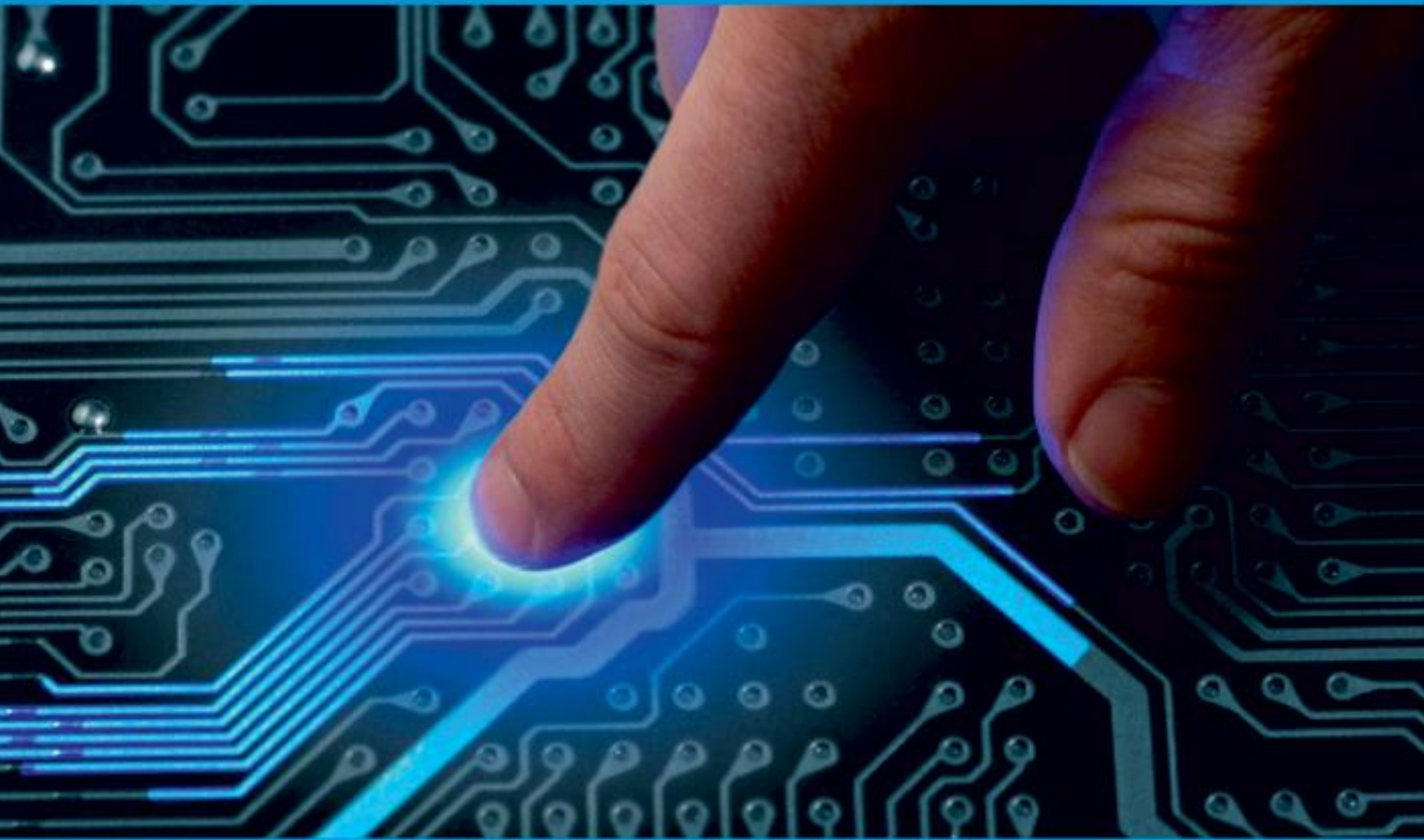
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

| DOI: 10.15680/IJIRCCE.2022.1004154|

# Implementation of Flight Fare Prediction Model using Data Science and Machine Learning Algorithms

**Kunal Patle, Khush Pachghare, Dhaval Kalamkar, Kajal Sardar, Dr. M. A. Pund**

Department of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera, Amravati, India

**ABSTRACT***:* What is the best time to buy a flight ticket? The airline implements dynamic pricing for the flight ticket. According to the survey, flight ticket prices change during the morning and evening time of the day. Also, it changes with the holidays or festival season. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. 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. The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices. Then using this information, build a system that can help buyers whether to buy a ticket or not.

**KEYWORDS***:* Machine Learning, Data Science, Random Forest Algorithm, Extra Tree Regressor, efficiency and Accuracy.

## I. INTRODUCTION

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. 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 take-off day. Since they trust that airfare will be most likely high when the date of buying a ticket is closer to the take-off date, yet it is not generally true. Purchaser may finish up with the paying more than they ought to for a similar seat. Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow and it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. In this project, we are going to predict the fare of the flights using the past datasets. Here we are provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities. Using this data, we will try to predict the price of the flights.

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 travellers 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 continuously expanding.

## II. RELATED WORK

Proposed study [1] Airfare price prediction using machine learning techniques, For the research work a dataset consisting of 1814 data flights of the Aegean Airlines was collected and used to train machine learning model. Different number of features were used to train model various to showcase how selection of features can change accuracy of model.In case study by William groves an agent is introduced which is able to optimize purchase timing on behalf of customers. Partial least square regression technique is used to build a model.In a survey paper by supriya rajankar a survey on flight fare prediction using machine learning algorithm uses small dataset consisting of flights between Delhi and Bombay. Algorithms such as K-nearest neighbours (KNN), linear regression, support vector machine (SVM) are applied.

| DOI: 10.15680/IJIRCCE.2022.1004154|

Research done by Santos analysis is done on air fare routes from Madrid to London, Frankfurt, New York and Paris over course of few months. The model provides the accepted number of days before buying the flight ticket.

Tianyi wang proposed framework where two databases are combined together with macroeconomic data and machine learning algorithms such as support vector machine, XGBoost are used to model the average ticket price based on source and destination pairs. The framework achieves a high prediction accuracy 0.869 with the adjusted R squared performance metrics in the research a desired model is implemented using the Linear Quantile Blended Regression methodology for San Francisco–New York course where each day airfares are given by online website. Two features such as number of days for departure and whether departure is on weekend or weekday are considered to develop the model.

## III. PROPOSED SYSTEM ARCHITECTURE

In this paper, we are proposing to develop a web application which can provide prediction result with speedy calculations and trying to achieve highest accuracy possible.
To above objective the below system architecture is proposed.

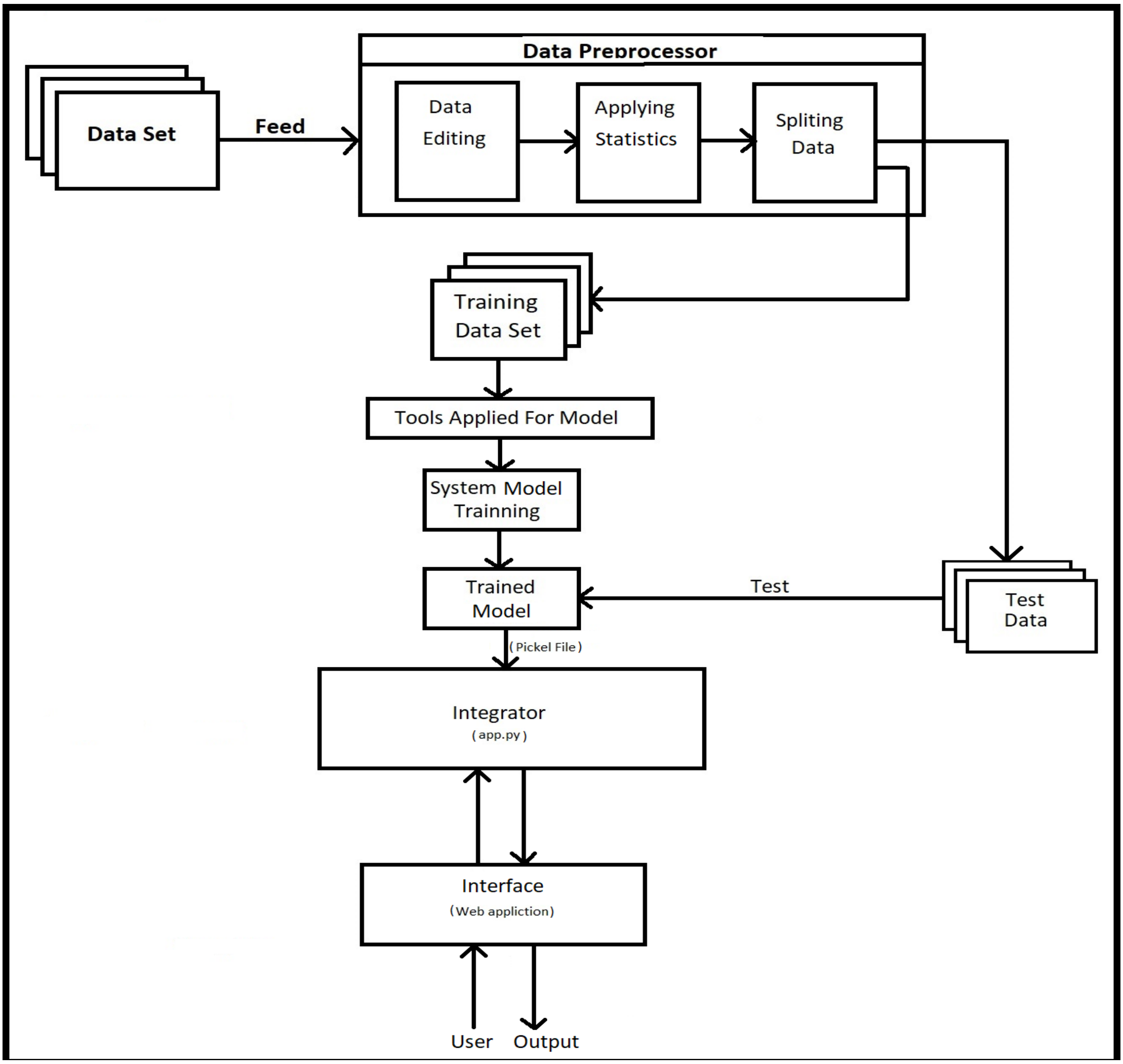


Figure 1: Flow chart – Work Flow

Data Set: The data set is the collection of related states of information that is composed of separate elements but can be manipulated as a single unit by a computer. Data set was acquired from Kaggle which contains past airline data from March to June in year 2019. contains fields like departure, date of departure, date of arrival, airlines, ticket price etc.

Data Pre-processing: The data is feed to data pre-processor unit in next phase. This unit comprises of three subunits

1.Data Editing

2.Appling Statistics

3.Splitting Data

I. The data is cleansed all the fields and records check for the presence of the null value in order to deal in with null value appropriate measures are taken.

II. Various statistical tools like mean median and mode were used to explore the data set after their correlation was used to find out the correlation among the various fields.

III. In this we split the data set into two different data set one is called training data set and other is called test data.

Training data: Training data set is used to train a model and test date set is used to train the model. Tools Application for Model: This scikit-learn is incorporated in our environment from this library we use random forest predictive algorithm for building a model.

System Model Training: Now training dataset is passed to the built model for training itself.

Trained Model: Now this train model is being tested on the test data set and it produces a certain accuracy finally of pickle file is generated which hold our current model.

Integrator: Integrator work as an intermediate platform Between user interface and trained model or the backend and frontend. "app.py" is the main file in this unit.

Interface: User interface is used to accept the query from the user and display output on the screen of a system.

## IV. IMPLEMENTATION

For this project, we have implemented the machine learning life cycle to create a basic web application which will predict the flight prices by applying machine learning algorithm to historical flight data using python libraries like Pandas, NumPy, Matplotlib, seaborn and sklearn.

Data selection is the first step where historical data of flight is gathered for the model to predict prices. Our dataset consists of more than 10,000 records of data related to flights and its prices. Some of the features of the dataset are source, destination, departure date, departure time, number of stops, arrival time, prices and few more. In the exploratory data analysis step, we cleaned the dataset by removing the duplicate values and null values. If these values are not removed it would affect the accuracy of the model. We gained further information such as distribution of data. Next step is data pre-processing where we observed that most of the data was present in string format. Data from each feature is extracted such as day and month is extracted from date of journey in integer format, hours and minutes is extracted from departure time. Features such as source and destination needed to be converted into values as they were of categorical type. For this One hot-encoding and label encoding techniques are used to convert categorical values to model identifiable values.

Feature selection step is involved in selecting important features that are more correlated to the price. There are somebe selected and passed to the group of models. Random forest basically uses group of decision trees as group of models. Random amount of data is passed to decision trees and each decision tree predicts values according to the dataset given to it. From the predictions made by the decision trees thefeatures such as extra information and route which are unnecessary features which may affect the accuracy of the model and therefore, they need to be removed before getting our model ready for prediction. After selecting the features which are more correlated to price the next step involves applying machine algorithm and creating a model. As our dataset consist of labelled data, we will be using supervised machine learning algorithms also in supervised we will be using regression algorithms as our dataset contains continuous values in the features. Regression models are used to describe relationship between dependent and independent variables. Following are the algorithm that are used in this project,

**Decision Tree**: Decision trees are basically of two types classification and regression tree where classification is used for categorical values and regression is used for continuous values. Decision tree chooses independent variable from dataset as decision nodes for decision making. It divides the whole dataset in different sub-section and when test data is passed to the model the output is decided by checking the section to which the datapoint belong to. And to whichever section the data point belongs to the decision tree will give output as the average value of all the datapoints in the sub-section

**Random Forest**: Random Forest is an ensemble learning technique where training model uses multiple learning algorithms and then combine individual results to get a final predicted result. Under ensemble learning random forest

| DOI: 10.15680/IJIRCCE.2022.1004154|

falls into bagging category where random number of features and records willaverage value of the predicted values if considered as the output of the random forest model.

**Performance metrics** are statistical models which will be used to compare the accuracy of the machine learning models trained by different algorithms. The sklearn. Metrics module will be used to implement the functions to measure the errors from each model using the regression metrics. Following metrics will be used to check the error measure of each model.

R 2 (Coefficient of determination) It helps you to understand how well the independent variable adjusted with the variance in your model.

$$R^2 = 1 - \frac{\Sigma(y' - Y)^2}{\Sigma(y - Y)^2}$$

The value of R-square lies between 0 to 1. The closer its value to one, the better your model is when comparing with other model values.

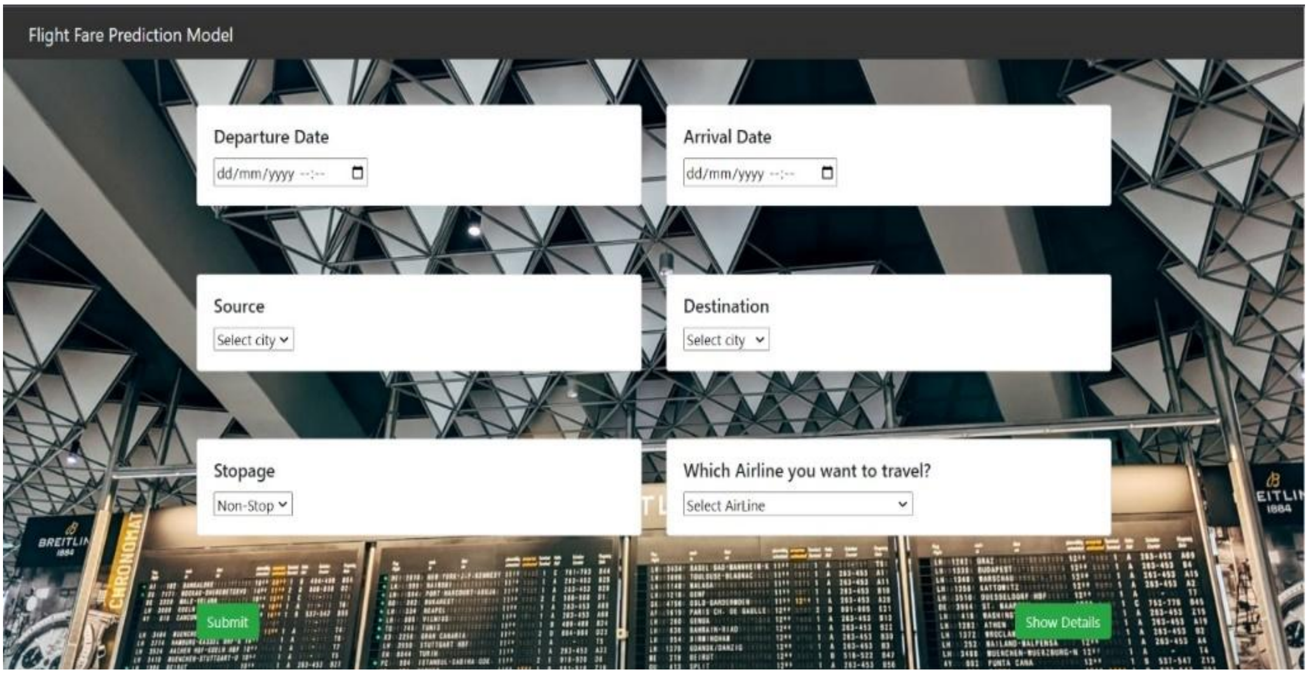
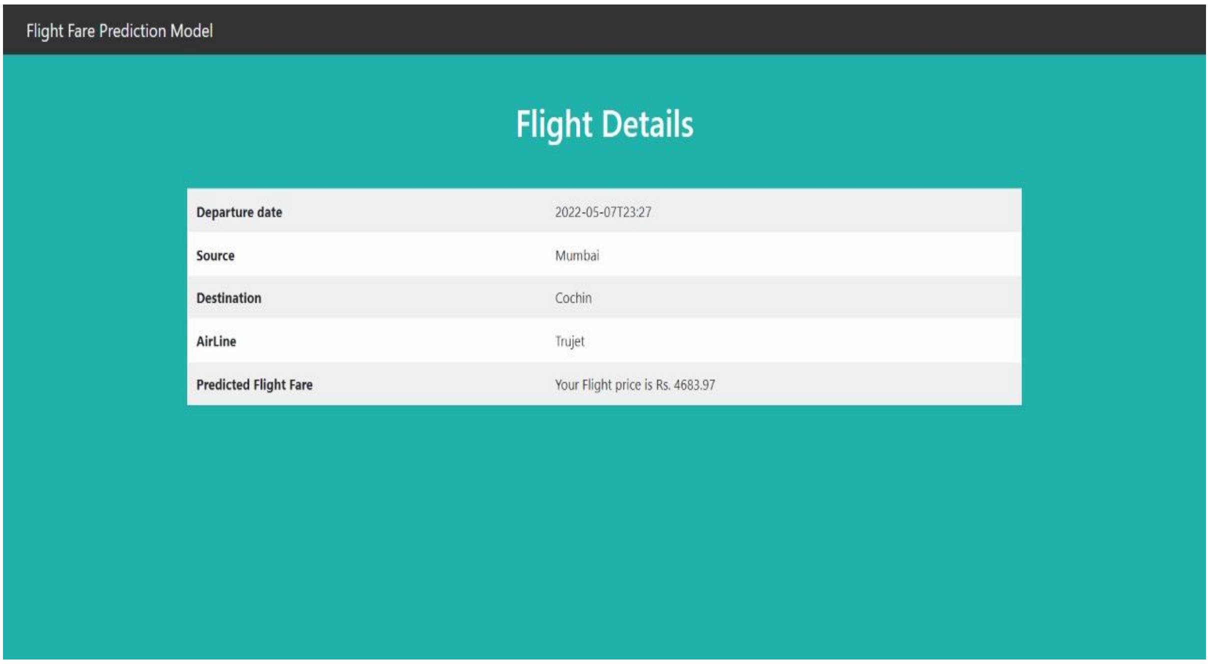


Figure 2: User Interface



Flight Fare Prediction Model

Flight Details

| | |
|---|---|
| Departure date | 2022-05-07T23:27 |
| Source | Mumbai |
| Destination | Cochin |
| AirLine | Trujet |
| Predicted Flight Fare | Your Flight price is Rs. 4683.97 |

Figure 3: Predicted Result

## V. CONCLUSION

A study was conducted to find out how to predict airplane ticket prices. An extensive dataset collection was carried out and Random Forest Regression Machine Learning model was used for deployment. Using visualization, the features which influenced airfare prices were identified. With experimental analysis, it can be concluded that Random Forest Regression model achieves good accuracy in predicting airfare prices. 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 experiments 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.

## REFERENCES

[1] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine larning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras" Proc. 12th Int. IEEE Conf. Intell. Transp. Syst. pp. 1-6 Oct. 2009.

[2] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.

[3] J. Santos Dominguez-Menchero, Javier Rivera and Emilio TorresManzanera "Optimal purchase timing in the airline market".

[4] Supriya Rajankar, Neha sakhrakar and Omprakash rajankar "Flight fare prediction using machine learning algorithms" International journal of Engineering Research and Technology (IJERT) June 2019.

[5] Tianyi wang, samira Pouyanfar, haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"

[6] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"

[7] Wohlfarth, T.clemencon, S.Roueff "A Dat mining approach to travel price forecasting" 10th international conference on machine learning Honolulu 2011.

[8] medium.com/analytics-vidhya/mae-mse-rmse-coefficient-ofdetermination-adjusted-rsquared-which-metric-is-bettercd0326a5697e article on performance metrics.

[9] www.keboola.com/blog/random-forest-regression article on random forest.

[10]https://towardsdatascience.com/machine-learning-basics-decisiontree-regression1d73ea003fda article on decision tree regression

[11] Frank, E., Hall, M. A., Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[12] Lawson, J. (2015). Data Science in Higher Education: A Step-by-Step Introduction to Machine Learning for Institutional Researchers. CreateSpace Independent Publishing Platform.

[13] R. M. Riensche et al., “Serious Gaming for Predictive Analytics,” in AAAI Spring Symposium on Technosocial Predictive Analytics. Association for the Advancement of Artificial Intelligence (AAAI), San Jose, CA, no. Zyda, pp. 108-113, 2009.

[14] N. Chinchor, J. Thomas, and P. Wong, “Multimedia Analysis+ Visual Analytics= Multimedia Analytics,” IEEE Computer Graphics, 2010.

[15] Breiman, L. (2001). Random Forest. Machine Learning 45, 5-32.

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ☎ 6381 907 438 ✉ ijircce@gmail.com

Scan to save the contact details

www.ijircce.com