Flight Fare Prediction Using Random Forest Algorithm

Ch.Rajani[1], Y.Hemasri[2], A.Sowmya[3], P.V.Geethika[4]
[1] Professor, [2, 3 & 4] Student
[1]rajani.kadiyala@gmail.com, [2] hemasriyakkali@gmail.com, [3] akkisettysowmya@gmail.com, [4] bhuvanareddy2305@gmail.com
Department of Computer Science and Engineering,
Narasaraopeta Engineering College, Narasaraopet, Andhra Pradesh, India

ABSTRACT- The globe is full of transportation networks. primarily buses, trains, airplanes, etc. In addition to using the train system, some people also travel by bus and airplane. Generally speaking, ticket prices for airline travel are higher than for other modes of transportation. We can easily move from one place to another because of the journey distance, which makes them more expensive. Also, as a result of ignorance, most people spend more money and are unaware of when prices are high and low. Thus, we are able to forecast the flight cost in our project and determine when it is high and when it is low.Here, we may use the Random Forest machine learning technique to forecast the flight cost by taking into account the origin, destination, departure and arrival times, number of stops, airlines, and other factors. Using the Random Forest Algorithm, we can use this information to estimate flight fares and inform clients of either high or cheap ticket prices. High performance accuracy was demonstrated by our research in Random forest,the obtained accuracy is 82.82%.

KEYWORDS: Flight fare prediction, random forest, machine learning, linear regression, and decision tree regression.

I. INTRODUCTION

Machine Learning is a subset of Artificial Intelligence, while Deep Learning is a subset of both. Using training algorithms, Machine Learning can identify patterns in data and use those patterns to make predictions. There are several algorithms in Machine Learning. Various Machine Learning methods have been suggested [1] for the purpose of examining consumer behavior. We are able to forecast which service a consumer is most likely to purchase even though they do not adhere to any predetermined guidelines .In this context, the Random Forest algorithm has proven to be a powerful and effective tool for predicting flight fares**.**

By utilizing various materials, such as airlines, routes, departure and arrival times, sources, destinations, and so on, we can forecast the cost of the flight. By taking into account the aforementioned information, this model can forecast the cost of the flight and make it easier for the user to purchase a ticket in the future by indicating when the price is high and low.

There are several benefits to predicting flying costs using random forests. And to do this, we must first determine other customers' purchasing patterns. If a new customer's purchasing patterns coincide with those of the past consumers, we may be able to forecast their decision [2]. In order to anticipate the flight fare, this machine learning technique—random forest—is used. Users are able to travel at a reduced cost and make better judgments about their tickets by utilizing the random forest algorithm. As a result, travelers may enjoy their trip and save money on airfare.

The assessment system adjusts the charge according to the day of the week, the season, and the holidays so that the header and footer on the following pages change. Airlines' primary objective is to earn a profit, even when passengers search for the best deal.Additionally, businesses can enhance the customer experience if the purchase decision can be made in advance by suggesting [3] the client's favorite services.

This paper's primary goal is to forecast various airlines' flight costs using the data in the dataset. In this case, we worked with data from various airlines, and as a result, flight fares may also differ. clients can also use machine learning techniques, such as random forest, to anticipate the flight fare. Airlines occasionally make offers to their clients depending on specific events.

II. LITERATURE SURVEY

Abhinav Garg et al.[4] describes effectiveness of random forest and gradient boosting for pricing prediction of airline tickets is compared in this research. The outcomes demonstrate the good performance of both random forest and gradient boosting, with random forest just barely surpassing gradient boosting.

K. D.V.N.Vaishnavi et al.[5] functionality to flight price prediction based on random forests is presented in this research. The writers pull information from a variety of sources, such as third-party aggregators and airline websites. The suggested method functions well, as evidenced by the findings, which show an accuracy of about 80%.

Shubham Agarwal et al. [6], random forest is one of the more effective regression techniques. The results show that random forest outperforms other methods of regression.

Sri Sai Venkata Subba Rao et.al.[7] their functionality described a machine learning method for predicting airline ticket prices that makes use of XGBoost and random forest. The outcomes demonstrate the effectiveness of the suggested strategy.

Ms Jetty Benjamin et al.[8] attempted to use both machine learning and neural network techniques to tackle the prediction problem. On the basis of historical data and the climate, the classification model was constructed. The predicted outcomes of this classifier have an accuracy rating of 72%. This study suggests that increasing the number of microclimate characteristics can increase prediction accuracy.

Wohlfarth et al.[9] model for improving ticket purchase times was developed using a measurable research strategy, unique pre-processing techniques for customers (known as macked point processors), and facts mining techniques (bunching and organizing). This system can be used to compute unsupervised grouping by converting heterogeneous value arranging information into a new value of the arrangement directions. The value direction is separated into groups according to comparable estimating behavior. Value change designs are evaluated for behavior and processing using the advancement model. The optimal addresses were ascertained by an order computation based on trees.

In conclusion, a number of studies have proven that random forest outperforms other conventional machine learning methods in the common task of predicting flight costs. Feature engineering, data set size, and data quality are only a few of the variables that affect the prediction model's accuracy.

III. PROPOSED SYSTEM

Below are some of the criteria that our model is proposed to meet. .

Data set Analysis
Data Visualization
Pre processing
Model Creation and Evaluation
Accuracy Table

A . Dataset Analysis

Flight_fare.csv is the dataset[11] we used to make our forecasts, and it was obtained from Kaggle.com.

	A	B	C	D	E	F	G	H	I	J	K
1	Airline	e_of_Jour	Source	Destinatior	Route	Dep_Time	rrival_Tim	Duration	otal_Stop	ditional_In	Price
2	IndiGo	24/03/201	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 M	2h 50m	non-stop	No info	3897
3	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR	05:50	13:15	7h 25m	2 stops	No info	7662
4	Jet Airway	9/06/2019	Delhi	Cochin	DEL → LKC	09:25	04:25 10 J	19h	2 stops	No info	13882
5	IndiGo	12/05/201	Kolkata	Banglore	CCU → NA	18:05	23:30	5h 25m	1 stop	No info	6218
6	IndiGo	01/03/201	Banglore	New Delhi	BLR → NA	16:50	21:35	4h 45m	1 stop	No info	13302
7	SpiceJet	24/06/201	Kolkata	Banglore	CCU → BLI	09:00	11:25	2h 25m	non-stop	No info	3873
8	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	18:55	10:25 13 M	15h 30m	1 stop	In-flight m	11087
9	Jet Airway	01/03/201	Banglore	New Delhi	BLR → BOI	08:00	05:05 02 M	21h 5m	1 stop	No info	22270
10	Jet Airway	12/03/201	Banglore	New Delhi	BLR → BOI	08:55	10:25 13 M	25h 30m	1 stop	In-flight m	11087
11	Multiple ca	27/05/201	Delhi	Cochin	DEL → BOI	11:25	19:15	7h 50m	1 stop	No info	8625
12	Air India	1/06/2019	Delhi	Cochin	DEL → BLR	09:45	23:00	13h 15m	1 stop	No info	8907
13	IndiGo	18/04/201	Kolkata	Banglore	CCU → BLI	20:20	22:55	2h 35m	non-stop	No info	4174
14	Air India	24/06/201	Chennai	Kolkata	MAA → CC	11:40	13:55	2h 15m	non-stop	No info	4667
15	Jet Airway	9/05/2019	Kolkata	Banglore	CCU → BO	21:10	09:20 10 M	12h 10m	1 stop	In-flight m	9663

Fig. 1. flight_fare.csv

The dataset is organized into several columns, with the name of the airline operating the flight being represented by the column labeled "Airline" in Fig. 1. The day of departure is indicated by Date_of_journey, and the destination city is indicated by source. Destination, denoting the city where the journey will conclude, A route is a description of a course of travel that inludes the destination and beginning sites..

Dep_Time, moment that the source city is left, Arrival_Time, a variable that indicates the exact moment of arrival in the target city, Amount that indicates the journey's cost.

B. Visualization of Data

After conducting the necessary data analysis on the earlier information, we were able to draw certain conclusions that would be useful in developing the model, including the trend between a few columns within the dataset.

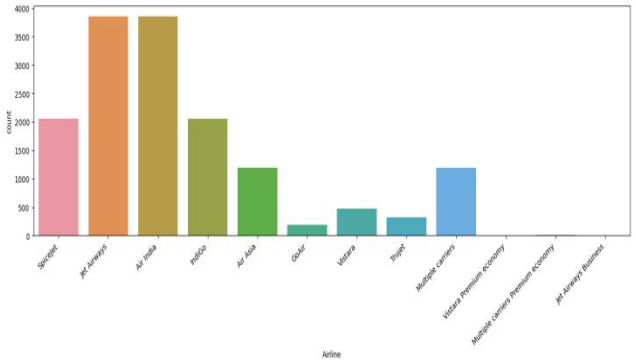

Fig. 2. Airline count

From the Fig. 2. observe that how many times an Airline company is mentioned inside the dataset. From the Fig.2 we can clearly see that Jet Airways is the clear winner. This means most of the users are using that specific airline for travelling.

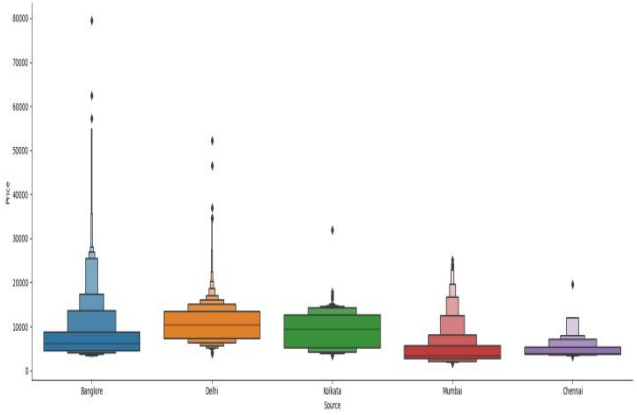

Fig. 3. Source count

Source count the insight is made about the count of each city as a source inside the dataset. As Fig.3 means this can be used to find from which city most of the users are travelling. Here we can see that Delhi is the clear winner which means from Delhi most of the users are travelling.

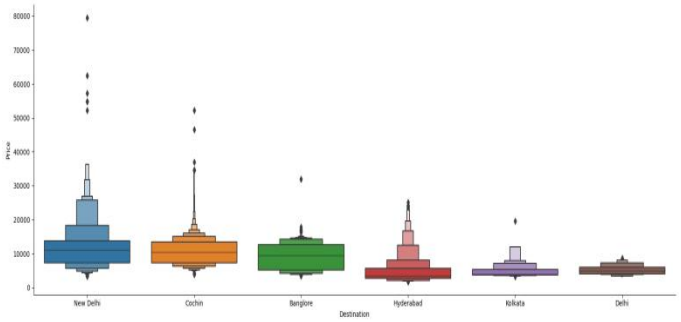

Fig. 4. Destination count

Destination count displays the number of cities that have been selected as destinations as described in the above Fig. 4. According to the dataset, we can infer from that the majority of users are traveling to Cochin.

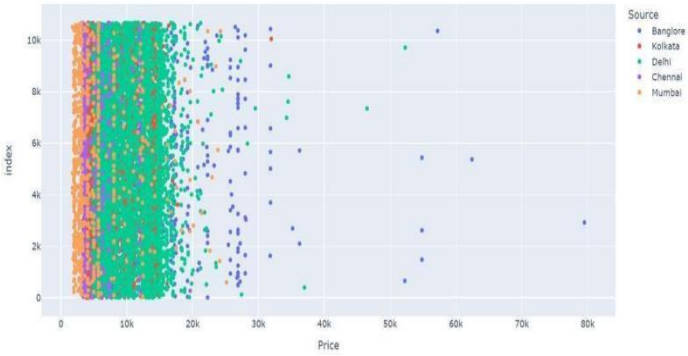

Fig. 5. Price difference

Price difference visualization in Fig. 5, we can see how the ticket fare varies as per the journey. From the insight we can clearly see that most of the users spent less that 30,000 INR for their journey. And there are less than 10% users who has spent more than 30,000 INR for their journey. Color of the plot changes according to the Source city too.

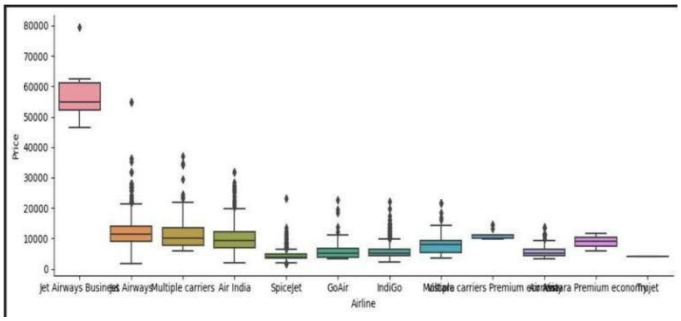

Fig. 6. Airways

Airways tells us about the relation between the Airline company and the Price of the journey. From the above Fig-6 we can see the median price of Jet Airways is the highest at 52,000 INR

C. Pre Processing Techniques

As using a dataset with large number of instances it will have some outliers and some noise in the data which can affect the performance of model. So to get rid of that data we have to perform data Pre-Processing which is a ML technique used to make the process ready for fitting it into a model. Pre-processing includes steps like replacing null values which means in our data if there is an instance with a null value in it will be replaced by any central tendency of the column like mean, mode or median, another technique like removing the outliers which means the data points which are far away from the other data clusters and will only disturb the data instead of making the model better.

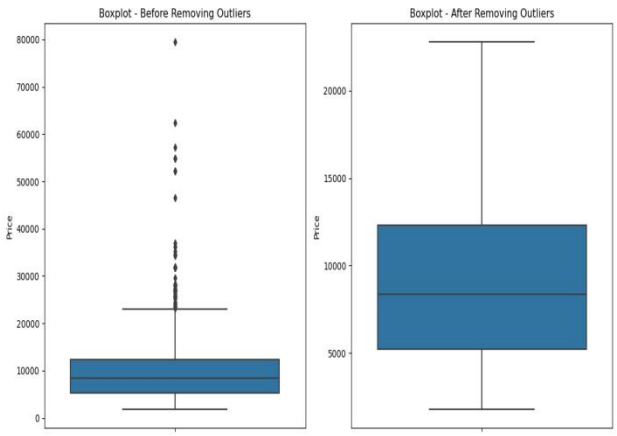

Fig . 7. Outliers

Outliers can also be replaced like null values or sometimes the instances with the null values will be entirely removed from the dataset as shown in the Fig. 7 as most of the people think that there is use of outliers.

	Total_Stops	Duration_hrs	Duration_mins	Day_of_Journey	Month_of_Journey	Dep_hr	Dep_min	Arrival_hr	Arrival_min	Airline_Air India	Airline_GoAir	Airline_In
0	0	2	50	24	3	22	20	1	10	False	False	
1	2	7	25	1	5	5	50	13	15	True	False	
2	2	19	0	9	6	9	25	4	25	False	False	
3	1	5	25	12	5	18	5	23	30	False	False	
4	1	4	45	1	3	16	50	21	35	False	False	

Fig. 8. Encoding

Another technique like encoding or one-hot encoding is also used in data pre-processing for the columns with categorical data. Encoding will convert the categorical data into numerical data Observed in the Fig.8 because most of the ML models works better with numerical data. This process is performed after analysis of data and the result of the data pre-processing will be used to generate the model to predict the data.

D. Feature Selection

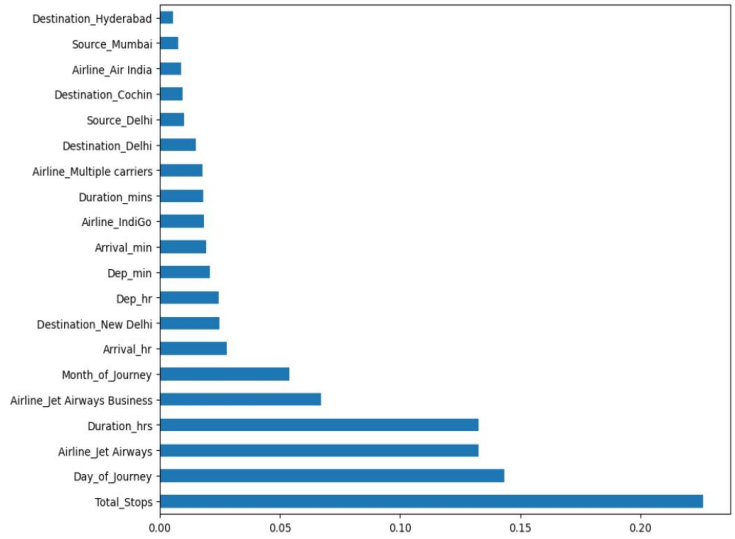

Fig. 9. Feature importance

From the graph in Fig. 9 “Total_Stops” is the most important feature. Looking closely at the high end of the feature important spectrum, there are no two features which are similar in their importance. However, at the lower end, we have four features which are almost similar in their importance. So, here may decide on reducing the dimensionality of the features but choose not to do so because the dimensionality of the data set is not too large and may model seems to handle it well.

E. Creation and Evaluation of Model

The next step after pre-processing the data is to create a model which better understands the data and predict the outcome more accurately. For this we have used few models like Logistic regression, Decision tree regression and Random forest to predict the outcome of a new data. After testing all three models with new data or test data the and examining the accuracies of all the models when tested with the testing data we came to a conclusion for the data we have Random forest classifier is performing better than others with an accuracy of 82.83%, which means out of every 100 instances our model can predict 83 instances correctly.The parameters of Random Forest are almost the same as those of a decision tree or stowing classifier model.

Example code:

Random forest code:-

```
from sklearn.ensemble
import RandomForestRegressor
model=RandomForestClassifier(n_estimators=200,
min_samples_split=3 max_features ="auto") model.fit(x_train, y_train)
y_pred = model.predict(x_test)
from sklearn.metrics import accuracy_score
ac =accuracy_score(y_pred, y_test)
mae=mean_absolute_error(y_pred, y_test)
print("Mean Absolute Error:" , mae)
mse=mean_squared_error(y_test, y_pred)
print("Mean Squared Error:" , mse)
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('The r2_score is', metrics.r2_score(y_test, y_pred))
```

#accuracy - 0.82

The table below shows the accuracy scores

of the three models which we have created.

F. Accuracy

One popular statistic for assessing how well a Machine Learning algorithm is performing is accuracy. Out of all the cases in the test dataset, it calculates the percentage of correctly classified instances.

TABLE I. ACCURACY OF DIFFERENT ALGORITHMS

ALGORITHM	ACCURACY
Linear Regression	62.50%
Decision Tree Regression	71.77%
Random Forest	82.82%
Lasso Regression	62.66%
Ridge Regression	62.12%
Polynomial Regression	52.07%

Linear Regression : In machine learning, one type of algorithm is called linear regression. It functions by determining the connection between one or more input variables and resultant variable. Linear predictor functions are used in the construction of these relationships getting 62.50%

Decision Tree Regression : This model belongs to the family of supervised learning models. Both regression and classification problems can benefit from its strong fit. It has a tree-like structure, as implied by its name, with decision nodes and leaf nodes getting 71.77%.

Random Forest Regression: Multiple decision trees are used by the Random Forest regressor to carry out regression tasks which results 82.82%. It is an illustration of group education. A supervised learning method called Random Forest regression and classification. The particular data that decision trees are trained on affects them. The decision tree that results from altering the training data may differ significantly, which may also affect the predictions. Because decision trees can't go back after they have split, they also tend to identify local optimal solutions, are computationally expensive to train.

Lasso Regression: The operator that combines the least absolute shrinkage and selection is called Lasso. Its utility makes it a popular choice for handling huge dimensional data in machine learning since it facilitates automatic feature selection resulting 62.66%.

Ridge Regression: To handle data multicollinearity, ridge regression is a linear regression technique. Multicollinearity, which occurs when two or more independent variables in a regression model have a high degree of correlation, can lead to instability and inflated standard errors for the regression coefficients reaching 62.12%.

Polynomial Regression**:**An approach for modeling the relationship between a dependent and independent variable is polynomial regression. In machine learning, it's also known as the multiple linear regression special case. Regression equation into Polynomial Regression, we must add certain polynomial terms to it and getting 52.07%.

F . Model assessment

This is an important stage of the study since it enables us to evaluate the accuracy and efficiency of our model. Test data is used in the evaluation of the model. In this instance, cross-validation was used to evaluate the model. This method creates k-subsets, or folds, from the data. The model is tested on the remaining folds after training on k-1 folds. This technique is performed k times, using each fold as a test set once. The average performance over all k-folds is the final outcome.

TABLE II. ACCURACY OF DIFFERENT ALGORITHMS

Evaluation	Existed[9]	Proposed
Mean Absolute Error (MAE)	1431.7940180939901	1212.7535713624798
Mean Squared Error (MSE)	7146863.527326601	4174259.223038839
Root Mean Squared Error (RMSE)	2673.3618399548163	2043.1003947527749
R Squared Value	0.6954008339903984	0.8220931639195391
Accuracy	78.77%	82.82 %

Based on the above table-II we can compare the predictive ability of a proposed model with an existing model[9], a number of metrics were utilized. With a mean absolute error (MAE) of 1212.75, the suggested model demonstrated an improvement in prediction accuracy compared to the old model's 1431.79 MAE. The mean squared error (MSE) demonstrated a similar pattern, with the suggested model considerably lowering the recorded 7,146,863.53 in the present model to 4,174,259.22. When compared to the current model, the root mean squared error (RMSE) decreased from 2673.36 to 2043.10, indicating improved prediction precision in the suggested model. Increasing from 0.6954 to 0.8221, the proposed model's R-squared value showed significant improvement in explaining the variability in the data. In summary, the suggested model improved the previous one with an 82.82% prediction accuracy compared to 78.77% for the latter. The combined effect of these metrics indicates that the suggested model performs better than the current one in a number of assessment criteria, demonstrating its greater efficacy and predictive accuracy**.** we can infer that the Random Forest algorithm model has an excellent accuracy of 82.82.

Root Mean Squared Error (RSME): The root mean squared error, or RSME, is the average squared difference between the predicted and actual values in a regression issue.

Mean Absolute Error (MAE): The Mean Absolute Error (MAE) is the absolute difference between the expected and actual numbers. Higher negative mean values suggest a better performing model.

Mean Squared Error (MSE): Mean squared error, or MSE, is a metric used to quantify the error level of a statistical model. The difference between the projected and actual values, squared, is averaged and assessed. The inaccuracy increases the model's value.

R-Squared: This statistic compares the regression model's fit to the data to a baseline model that consistently predicts the mean value. It displays the extent to which the model explains the variation in the data.

TABLE III. BEFORE AND AFTER K-FOLD

Classifiers	Actual accuracy	K-Fold accuracy
Linear	62.44	62.50
Decision	68.91	71.77
Random	82.70	82.82
Lasso	62.59	62.66
Ridge	62.12	62.33
Polynomial	52.07	52.07

G . Friendly User Interface

In the output screen of flight fare prediction as shown in the Fig. 9 predicts the price of airline which is desired according to the source, destination, date of journey, time and additional stoppage information which describes number of stops select the data provided in the drop down menu which is present at the right side of the field shows the result as your flight price in rupees.

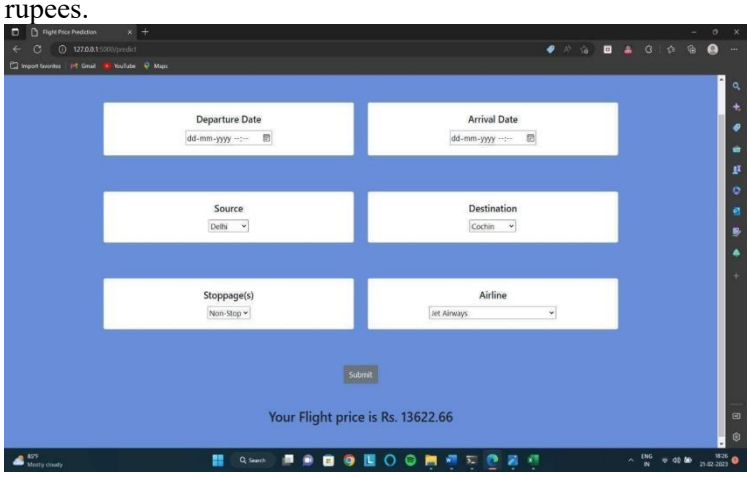

Fig. 9. Result

IV. CONCLUSION AND FUTURE SCOPE

The study's goal was to estimate a user's flight cost for a new instance. The accuracy varies for different algorithms. The accuracy for Random Forest algorithm is 82.82% when K-Fold cross validation is applied. The accuracy of Linear Regression algorithm is 62.50% when K-Fold cross validation is applied. The accuracy for Decision tree Regressor K-Fold cross validation is 71.77%. The accuracy of Lasso Regression algorithm is 62.66% when K-Fold cross validation is applied. The accuracy of Ridge Regression algorithm is 62.12% when K-Fold cross validation is applied. The accuracy of Polynomial Regression algorithm is 52.07% when K-Fold cross validation is applied.So finally we can conclude that the we get the highest accuracy for Random Forest algorithm and least accuracy for Polynomial Regressor.To develop more accuracy using machine learning algorithms and advanced techniques. The work can be extended and improved for the automation of Flight Fare analysisby using Machine Learning.

V. REFERENCES

[1] S. Naveen Prasath, Dr. Sathish Kumar M, and Ms. Sherin Eliyas, "A Prediction of Flight Fare Using K-Nearest Neighbors" 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022.

[2] I. C. Adetunmbi and Vishan Lal, Paul Stynes and CristinaMuntean. “An Investigation into Predicting Flight Fares in India using Machine Learning Models,"by https://www.researchgate.net/publication/373531278,Published in the Journal of Computer Science and its Applications 2023.

[3] K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," *2017 25th European Signal Processing Conference (EUSIPCO)*, Kos, Greece, 2017, pp. 1036-1039, doi: 10.23919/EUSIPCO.2017.8081365.

[4] Abhinav Garg, Abhishek Dixit, Abhinav Raj, Neeraj Arya.Airfare prediction model based on machine learning. International Journal Of Innovative Research In Technology IJIRT Access, 8, 168080-168090,2023

[5] K.D.V.N.Vaishnavi, L. Hima Bindu, M. Satwika.Airfare prediction using machine learning models. EPRA International Journal of Research and Development (IJRD) , 29(10), 3852-3859,2023.

[6] Shubham Agarwal, Ram Agrawal, Neha Singh, Kiran Adsul. Airfare prediction using machine learning models with user preference. International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) 78, 15-24,2022.

[7] N. Sri Sai Venkata Subba Rao, S. John Justin Thangaraj.Airline ticket price forecasting using machine learning algorithms. Journal of Section-A Research, 58(6), 964-978 ,2023

[8] Ms Jetty Benjamin A comparative study using machine learning for airfare prediction. 11(4), 460–467, National Conference on Emerging Computer Applications (NCECA)-2022.

[9] Alex Krantz and Jason Turley "Predicting Flight Prices with Machin Learning" by - This is a blog post that uses machine learning algorithms such as Gradient Boosting to predict flight fares. The post is available on Medium, 2021.

[10] Parth Kulkarni and Nirali Desai "Airfare Prediction using Machine Learning by - This is a research paper that uses machine learning algorithms such as Random Forest and Gradient Boosting to predict flight fares. The paper is available Research Gate,2022.

[11] Dataset Link: https://www.kaggle.com/datasets/riteshbagdi/flight-fare-prediction-dataset

[12] Thang Le Duc, Rafael García Leiva, Paolo Casari, and Per-Olov Ostberg.Machine Learning Methods for Reliable Resource Provisioning in Edge-Cloud Computing: A Survey. ACM Comput. Surv. 52, 5, Article 94,2019

[13] Fan Wu, Guihai Chen, Chengfei Lyu, Chun Hu, Zhihua Wu, and Renjie Gu. A thorough analysis of machine learning from server-based to client-based approaches. ACMComput. Surv. 54, 1, Article 6,2020.

[14] Muhammad Waheed, Saad Sajid Hashmi, Muhammad Usman, and Muhammad Ikram. He Xiangjianr. Internet of Things Security and Privacy: Risks and Reactions with Block chain and Machine Learning. ACM Comput. Surv. 53, 6; Article 122,2020.

[15] T. Janssen, “A linear quantile mixed regression model for prediction of airline ticket prices,” Bachelor Thesis, Radboud University,Published in the International Journal of Engineering and Advanced Technology, 2021.