RESEARCH ARTICLE OPEN ACCESS

THE PREDICTION PRICES FOR AIRLINES WITH AN ARTIFICIAL INTELLIGENCE SYSTEM OF LEARNING

BADAM ANOOP,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
POOJA PATIL,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
ENNU ROHITH REDDY,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
NITISH SHARMA,UG scholar,CSIT,Sri Indu College Of Engineering & Technology(A)
Mr Shek Shakeel,Asst.Prof.CSIT,Sri Indu College Of Engineering & Technology(A)

ABSTRACT

The Flight Fare Prediction System is a comprehensive solution aimed at accurately forecasting flight ticket prices, providing travelerswithvaluableinsightsforbetterplanninganddecision-making.Nowadays,airlineticketpricescanvarydynamically forthesameflight.Fromthecustomer'sperspective,theywanttosavemoney,soIhaveproposedamodelthatpredictsthe approximateticketprice.Thissystemleveragesmachinelearningalgorithmsandhistoricalflightdatatogenerateaccuratefare predictions. The system utilizes a vast dataset comprising historical flight fares, including factors such as travel dates, destinations, airlines, departuretimes, andvariousotherrelevantvariables.Byanalyzingthisdatausingadvancedmachine learningtechniques,the systemlearnspatternsandrelationships,enablingittomakereliablepredictionsaboutfuture flight fares. An ensemble of machine learning algorithms, including regression-based models like Random Forest, Gradient Boosting, and SupportVectorRegression,isemployedtocapturecomplexpatternsandrelationshipswithinthedata.Thissystemwillgivepeoplean ideaofthetrendsthepricesfollow and also providethepredictedvalueoftheprice,whichtheycancheckbeforebooking flights to save money. This kind of system or service can be provided to customers through flight booking companies to help them book tickets.

KEYWORDS-*FlightFarePrediction, MachineLearning, HistoricalFlightdata, RandomForest.*

I. INTRODUCTION

Every one knows that holidays always call for a much-needed vacation and planning the travel itinerary becomes atime-consuming task. The commercial aviation business has grown tremendously and has become a regulated marketplace as a result of the worldwide growth of the Internet and E-commerce. Hence, for Airline revenue management, different strategies like customer profiling, financialmarketing, andsocialfactorsareusedforsettingticketfairs.Whenticketsarebookedmonthsinadvance,airfares are often reasonable, but when tickets are booked in a hurry, they are often higher. But, the number of days/hours until departure isn't the only factor that decides flight fare, there are numerous other factors as well. Customers find it quite difficult to obtain a perfect and lowest ticket deal due to the aviation industry's complex pricing methodology. Machine Learning and Deep Learning-basedtechnologiesandmodelshavebeencreatedtoovercomethischallenge,andsubstantialresearchisalsohappening.Thisstudydiscusses a Machine Learning-based Flight FarePredaction System that employ sRandom Forest Regression to predictairlineticket pricing. Various features that influence prices arealso studied along withthesystem's experimental analysis. Section II included literaturereviewthatlookedattechnicalpapersaswellassomecurrentmodelsandsystems.Differencesinthefeaturesconsideredarealsomappeddown,InSectionIII,theproposedsystemisdescribedindetailalongwiththeworkflowanditsfeatures.InSectionIV,theresultsaswellasvariouscomparisonsbetweenfindingsarereported.InSectionV,conclusionsareprovided.InSectionVI, prospective advancements for further research.

II. RELATEDWORK

The paper begins with some broad information regarding machine learning, after which the authors further proceed to the methodology. The methodology consists of four-phase process that influences flight prices, collection of data from flight fare prediction dataset by MH, selection and evaluation of an accurate ML Regression model. Key to its success is the integrity of the system where consumers accurately represent the segments for which prices have been differentially determined.

[2] Today,airlinespricetickets“asmuchasthecustomerandmarketwillbear,”accordingtoconsultantandformerairlineplanning executive. Airlines also profile their customers to help them adjust prices.
[3] This often means placing passengers into one of two groups: leisure or business. And the way each group is priced is very different. Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the marketsegmentlevel,however,isstillverylimited.Wedefinetheterm marketsegmentasthemarket/airportpairbetweentheflight origin and the destination.

[4] Theairlinedatasetincludedthefollowingeightcharacteristics:departureandarrivaltimes, typeof airline,numberof stopages, source, destination and additional_information. The authors performed prediction using regression Machine Learning models that including, LGBM Regressor, Random Forest Regression Tree, and Decision Tree Regressor. Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

[5] Allinpresentareviewofdeeplearningandsocialmediadata-basedAirlineticketpricepredictionmodel.Theauthorsintroduce thecurrentairlineticketpricingsituationwiththefactorsthataffectticketprices.ARandomForestoperatesbyconstructingseveral decision trees during training time and outputting the mean of the classes as the prediction of all the trees. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

[6] RandomForest Modelis used for developmentsince itoutperforms othermodels such asLGBM DecisionTree Regressor and NeuralNetworkintermsofdataperformancewithaRsquaredscoreof0.91,thispredictionframeworkhasagoodlevelofaccuracy.

III. METHODOLOGY

Thefollowingstepsareinvolved inthe proposed architectureofourproject.

DataCollection

The act of obtaining, acquiring, and combining the data that will be used to develop, test, and verify a machine learning model is knownasdatacollectioninmachinelearning.Thisstepplayscrucialroleinimplementation.Heredataiscollectedfromflightfare dataset which is imported from Kaggle. The dataset consists of both categorical data and numerical data. The categorical data includessource, destination,typeofairline,additionalinfoandnumericaldataincludesarrivalanddeparturedates,numberofStops. There are 11 columns (each represents a feature) and 10683 rows in this large dataset.

DataPreprocessing

Data preprocessing means nothing but cleaning data, which can be used for model training and testing. By this step we can make our data useful for model training purpose. Data preprocessing involves cleaning, transforming, and preparing the data for data analysis. The sub steps involved in the data preprocessing are:

DataCleaning: Inthisstepthenullvaluesareremoved, missingvaluesareremovedandif anyduplicatesarepresentthatarealso removed.

Feature selection and engineering: In this step the features of our model are extracted and all the relevant features are used for modeltraining.Indatasetitcontainsdateofjourney,arrivaldate, departuredatecolumnsandallthenumericalvaluesare extracted asDeparturehour,departureminutes,arrivalhour,arrivalminutes,journeyday,journeymonth.Asdatasetcontainsbothcategorical andnumericalfeatures,byusing'Onhotencoding'methodfor nominalcategoricaldataand'labelencoding'for ordinalcategorical datawasusedtoconvertthecategoricalvaluestonumericalvalues.Thedatasetconsistsofcategoricalvariableslikeairline,source, destination, route, total number of stops and additional info.

Data Splitting

Thisstepinvolvessplitting our dataintotwopartsfor trainingandtestingpurpose. Formodeltraining80percent of data wasused by using Random Forest regressor model was trained. The machine learning algorithms are:

LGBM Regressor

LGBMstandsforLightGradientBoostedMachine.Itisagradientboostingframeworkbasedondecisiontreesthatcanbeused for variousmachinelearningtaskssuchasregression, classificationandranking. LGBMRegressor isaclassinlightgbmpackage that can be used to train and predict regression models.

DecisionTreeRegressor
DecisiontreeregressorisaclassinSklearntreemodulethatcanbeusedtotrainandpredictregressionmodels.Itisadecisiontree-based algorithm that recursively partitions the input data based on the values of the input features, forming a tree-like structure. It initiallychoosesindependent variablefromdataset asdecision nodesfor decision makingandthenitdividestheentiredataset into sub-sectionsandwhentest dataispassedto themodel theoutput isdecidedbychecking thedatapoint belongsto thedecision tree will give output as the average value of all the datapoints in the sub-section.

RandonForest Regressor
Random Forest regressor uses multiple decision trees to perform regression tasks. It is an example of ensemble learning. Random forest is a Supervised Learning algorithm which uses ensemble learning approach for classification and regression. Decision trees are sensitive to the specific data on which they are trained. If the training data is changed the resulting decision tree can be quite different and in turn the predictions can be quite different.Also, Decision trees are computationally expensive to train, carry a big riskofoverfitting,andtendtofindlocaloptimalbecausetheycan'tgobackaftertheyhavemadeasplittoaddresstheseweaknesses, we turn to Random Forest.

Modelevaluation

This is animportantstep inour project, as ithelpsus to measure theperformance and accuracy of our model. Test datais used for model evaluation. Here, we employed Cross-validation for model evaluation. This method divides the data into k-subsets, called folds. the model is trained on k-1 folds and tested on the remaining fold. this process is repeated k times, so that each fold is used as a test set once. The average performance across all k-folds is reported as the final result. The metrics that are used for model evaluation purpose are:

RootMeanSquaredError(RSME):Itgivestherootoftheaveragesquareddifferencebetweentheactualvaluesandthepredicted values for a regression problem.

MeanAbsoluteError(MAE):Itgivestheabsolutedifference betweentheactualvaluesandpredictedvalues.Thehighernegative mean values indicate the better performance of model.

R-Squared: This metric measures how well the regression model fits the data, by comparing it to abaselinemodel that always predict the mean value. It shows how much variation in the data is explained by the model.

ModelArchitecture
Thearchitectureiscraftedwithessentialcomponentstooptimizeprediction.Ourproposedsystemanalyseshistoricaldatatoidentify patterns, seasonal trends and additional information that influence flight fares.

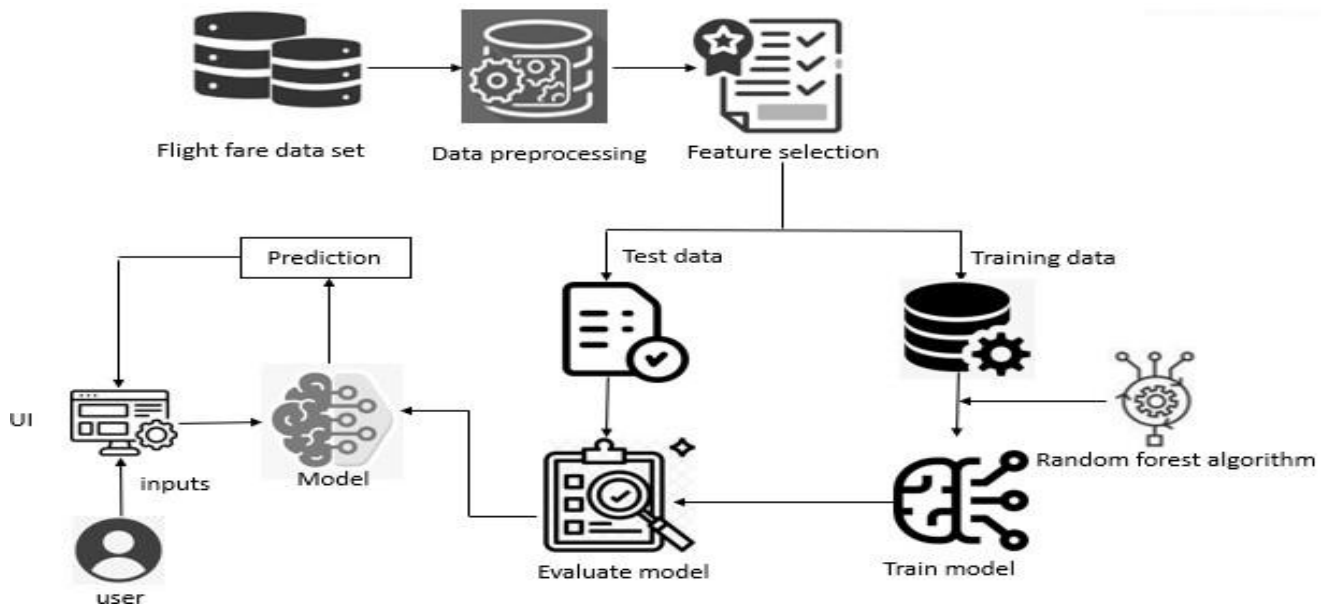


Fig.1:SystemArchitecture

IV. IMPLEMENTATIONAND RESULTS

1. Importingtherequiredlibraries.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_validate
from sklearn.metrics import mean_squared_error as mse
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_percentage_error
from math import sqrt
from scipy.stats import randint as sp_randint
from sklearn.linear_model import Ridge
from sklearn.linear_model import Lasso
from sklearn.linear_model import LinearRegression
from lightgbm import LGBMRegressor
from xgboost.sklearn import XGBRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import KFold
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import RandomizedSearchCV
from prettytable import PrettyTable
```

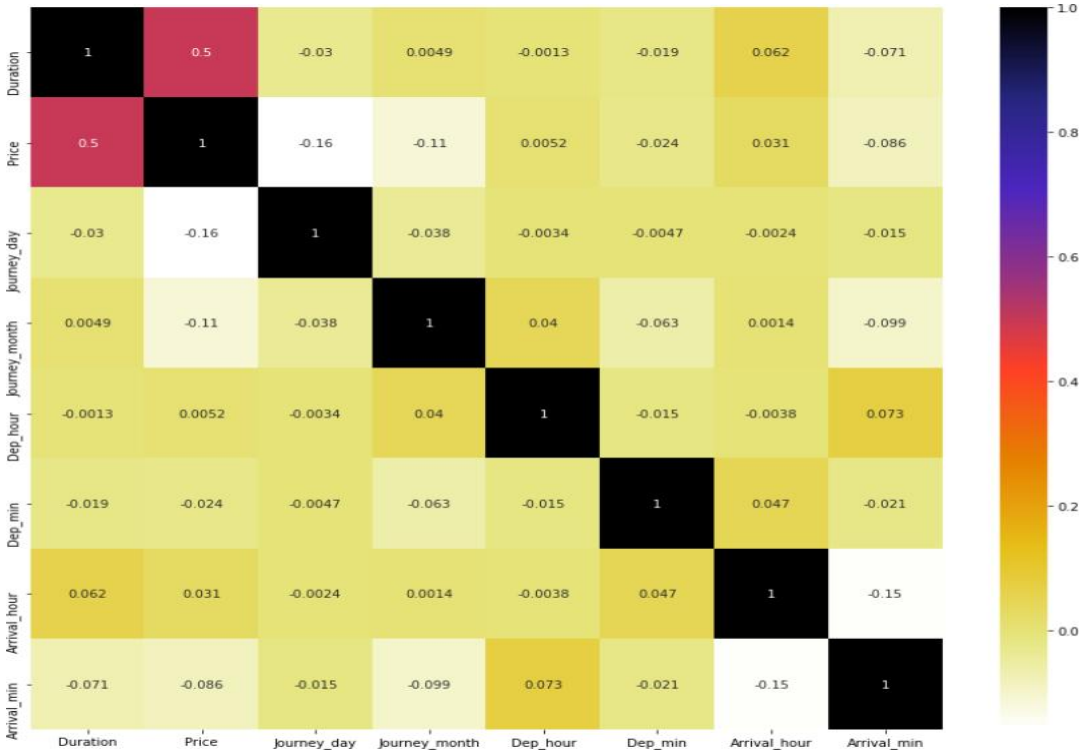
2. Readthedataset:
Ourdatasetformatmightbein.csv,excelfiles,.txt,.json,etc.We canread thedatasetwiththehelpofpandas.

```
df = pd.DataFrame(pd.read_excel("Dataset_Bonus_project.xlsx"))
pd.pandas.set_option('display.max_rows',None)
pd.pandas.set_option('display.max_columns',None)
df.head()
```

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

3. Datapreprocessing:
Thedf.isnull()methodisusedtoverifythatnovaluesare present. Weemploythesum()functiontoaddup thosenull values. Two null values were discovered in our dataset, we discovered. We thus start by investigating the data.

4. UsingaHeatMaptocheckthecorrelation:
Herewareusingaheatmaptocheckthecorrelationinth isinstance.Usingdifferentcolourcombinations,itdisp laysthe data as 2-D coloured maps. Instead of numbers, it will be plotted on both axes to describe the relationship variables.



5. By comparing all the models (LGBM Regressor, Decision Tree Regressor, Random Forest Regressor), we can conclude that Random Forest Regressor performs the best.

Model Name	Tr. RMSE	Tr. R-Squared	Te. RMSE	Te. R-Squared
Decision Tree Regressor	1480.8751646292635	89.32359295988586	2050.6082679556803	75.17121421662682
Random Forest Regressor	1020.1994631776361	95.2064826543969	1594.1167209563912	81.01685133811701
LGBM Regressor	1320.1604989356903	91.1690057894266	1522.1802516825198	80.0739576571085

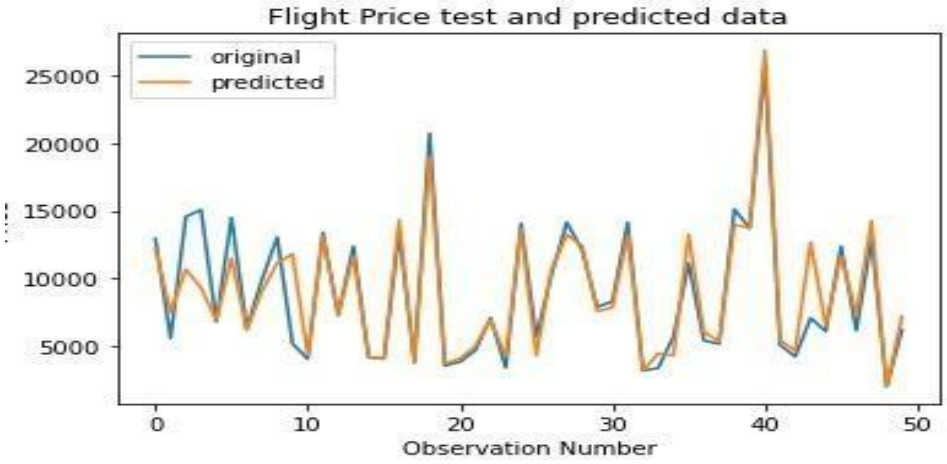


Fig2:FlightPriceTestandPredictedData

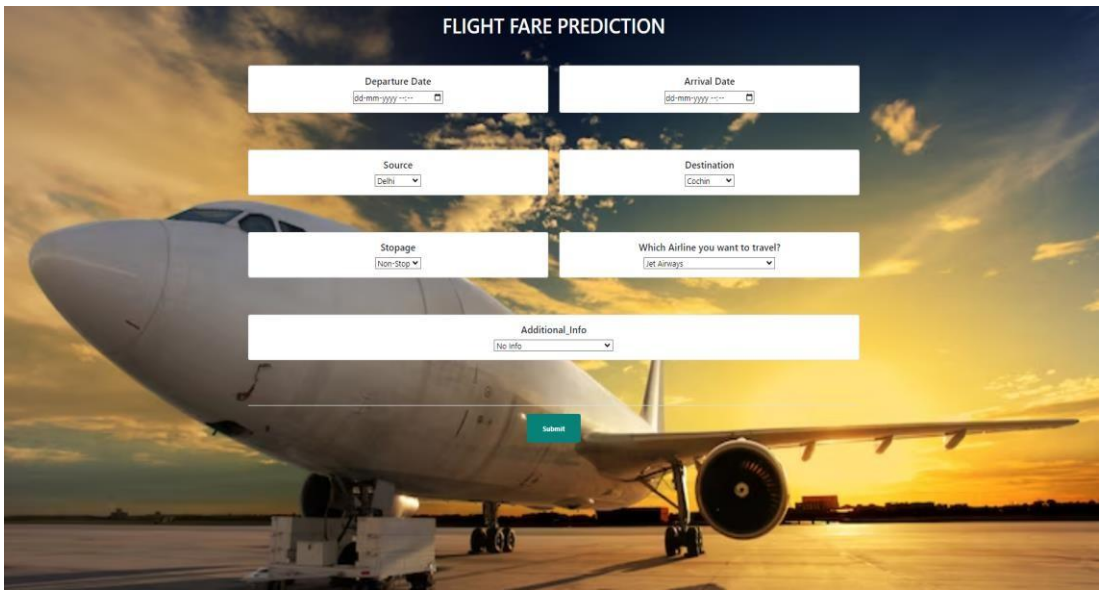


Fig.3:HomePage

Figure 3 By using flask the machine learning model wasy deployed. Flask is a python web frame work that allows you to build lightweightandflexiblewebapplicationsquicklyandeasily. Thewebpagecollectsinformationanditpredictsthefareof ticket.It collects information like departure date, arrival date, source, destination, number of stops, type of airline and additional information(meal included or not, no info, change airports, etc).

V. FUTURESCOPE

- Optimal date recommendation: Itmeanssuggesting thedate touserson which datethe flightpriceswillbe minimum.
- Real-Time Updates: Dealing with real-time datafor dynamicpricingadjustments based on factorslike weather, demand, and airline policies.
- Integration: Partnering with airlines, travel agencies, and online booking platforms to provide pricing as a value-addedservice.

VI. CONCLUSION

In conclusion, the main aim of our project flight fare prediction using machine learning is to predict the prices. we have created a User Interface for theentire process which includes arrival date, departuredate, source, destination, etc. Our flight fareprediction project using machine learning has successfully produced a reliable and user-friendly system. We collected, preprocessed, and extracted features from flight fare data, traineda robust random forest model andevaluated its performance. Thisweb application we developed empowers travelers to make informed decisions by predicting flight prices based on their input.

VII. REFERENCES

1. K.Tziridis,Th.Kalampokas,G.A.Papakostas,"AirfarePricesPredictionUsingMachineLearningTechniques",25thEuropeanSignal Processing Conference (EUSIPCO), IEEE, October 26, 2017.
2. N.BrownandJ.Taylor,AirFare:Stories,Poems&EssaysonFlight.SarabandeBooks,2004.
3. J.C.Lok,PredictionFactorsInfluenceAirlineFuelPriceChangingReasons.2018.
4. A.KakaraparthiandV.Karthick,"Asecureandcost-effectiveplatform foremployeemanagement system usinglightweight standalone frameworkoverDiffieHellman'skeyexchangealgorithm,"ECSTrans.,vol.107,no.1,pp.13663–13674,Apr.2022.
5. Wikipedia,"MeanSquarederror",Available:https://en.wikipedia.org/wiki/Mean_squared_error.
6. G. Ataman and S. Kahraman, "Stock Market Prediction in Brics Countries Using Linear Regression and Artificial Neural Network Hybrid Models," The Singapore Economic Review. pp. 1–19, 2021. doi: 10.1142/s0217590821500521.
7. B. Panwar, G. Dhuriya, P. Johri, S. S. Yadav, and N. Gaur, "Stock Market Prediction Using Linear Regression and SVM," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). 2021. doi: 10.1109/icacite51222.2021.9404733.
8. P.PureyandA.Patidar,"StockMarketClosePricePredictionUsingNeuralNetworkandRegressionAnalysis, "InternationalJournal of Computer Sciences and Engineering, vol. 6, no. 8. pp. 266–271, 2018. doi: 10.26438/ijcse/v6i8.266271.
9. WilliamGrovesandMariaGini"Anagentforoptimizingairlineticketpurchasing"inproceedingsofthe2013internationalconference on autonomous agents and multi-agent systems.
10. https://towardsdatascience.com/machine-learning-basics-decisiontree-regression-1d73ea003fdaarticleondecisiontreeregression.
11. www.keboola.com/blog/random-forest-

regressionarticleonrandomforest.