AIRLINES TICKET PRICE PREDICTION USING MACHINE LEARNING APPROACH

BY

Ashiqur Rahman Riaz
ID: 183-15-11943

Dewan Sakibur Rahman
ID: 183-15-11841
AND

Nazmus Sakib
ID: 183-15-11860

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Abdus Sattar
Assistant Professor and Coordinator M.Sc
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Md. Zahid Hasan
Associate Professor and Coordinator MIS
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
SEPTEMBER 2022

APPROVAL

This Project titled "**Airlines Ticket Price Prediction Using Machine Learning Approach**", submitted by Ashiqur Rahman Riaz, ID No: 183-15-11943, Dewan Sakibur Rahman, ID No: 183-15-11841 and Nazmus Sakib, ID No: 183-15-11860 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 12-September-2022.

BOARD OF EXAMINERS

Chairman

Dr. Sheak Rashed Haider Noori
Associate Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology

Internal Examiner

Md. Sabab Zulfiker
Senior Lectuer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Abdus Sattar, Assistant Professor and Coordinator M.Sc, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma

Supervised by:

Abdus Sattar

Assistant Professor and Coordinator M.Sc

Department of CSE

Daffodil International University

Submitted by:

Ashiqur Rahman Riaz

Ashiqur Rahman Riaz

ID: -183-15-11943

Department of CSE

Daffodil International University

Dewan Sakibur Rahman

Dewan Sakibur Rahman

ID: -183-15-11841

Department of CSE

Daffodil International University

Nazmus Sakib

Nazmus Sakib

ID: -183-15-11860

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes it possible for us to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Mr. Abdus Sattar**, **Assistant Professor and Coordinator M.Sc**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Prof. Dr. Touhid Bhuiyan, Professor, and Head,** Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The price of airline tickets is the most unstable thing nowadays. It changes abruptly during the morning and evening time. The passengers are always looking to get the tickets at the lowest price, on the other hand the sellers (Airlines) are trying to earn a huge revenue. We can see that the prices change within a short time because of some factors for which the prices are affected. There are some factors like purchasing time, fuel price, flight distance etc. The prices of the airfare depends on these factors. The passengers are not allowed to access the previous data of the flight prices to predict the best price for them but the airlines have all the information about that. In this research, we tried to find out a best model for predicting the airfare by which the passenger can get the best predicted price to travel. We have used the Random Forest regression algorithm, Decision Tree algorithm and Linear Regression Algorithm to predict the price of airline tickets. For applying the ML algorithms, we have extracted the best features from the collected data and after finishing all of the tasks we got the prediction accuracy 90.47% in Random Forest Regression, 79.20% in Decision Tree and 72.77% in Linear Regression. After all, we got the best model which is Random forest Regression Algorithm to predict the airfare price. By using this system, the customers will get a better prediction that can help them buy tickets at a lower price.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE NO. **PAGE NO.**

CHAPTER 1

INTRODUCTION

1.1 Introduction

The airline industry is one of the most profitable industries in the world. Nowadays their business strategy is more efficient and they use more complexities for airline ticket pricing. We can't even imagine that for the same flight the ticker price is changing dynamically.

Airfare price prediction is an important issue which needs to be observed. In this era, the people are traveling through the air to save time and money also. But the sellers (airlines) are trying to keep their profit as high as possible. They don't care about the buyers (passengers) which is an unbearable problem for the buyers. To set the price of airfare the airlines use some mathematical models and sophisticated rules which are based on a complex structure [1] [2]. By doing that, they are growing their revenue. The airlines are applying high complexity of the pricing models. For that reason, the customers are suffering and they can’t buy the tickets at lower cost [3].

There are two types of Machine learning algorithms. They are Supervised and unsupervised algorithms. We have used a most popular ML algorithm which is a supervised learning technique named Random Forest algorithm. And we also used Decision Tree and Linear Regression Algorithms to compare their results with the Random Forest Algorithm. Random Forest Algorithm is a strong modeling technique and it works with a collection of many trees, for that it is more powerful than a single decision tree. It is a mixture of tree predictors.

Decision Tree Algorithm is a supervised learning algorithm which can be used for both classification and regression problems [12]. Linear Regression is a ML algorithm based on supervised learning. It performs a regression task. It is mostly used for finding out the relationship between variables and forecasting.

People are suffering. For that reason many researchers are searching for a best model by which the passengers can get the best price of airfare to travel and save money. From that motive we are trying to find the best model for predicting the best price of airfare. We are using these algorithms to find out the best model to predict the airfare prices among them. The best model will be so helpful for the people.

1.2 Motivation

People always like to travel. Nowadays people are traveling by airplane. Ordering and purchasing online for airline tickets is now very popular. The airlines always set the prices of tickets based on the unsold seats and market demand.

Think you are going on a tour by airplane. After takeoff you started a conversation with the other passenger who was sitting next to you near the window. While you were talking about the price of the ticket, you found that you paid 50% more for the ticket with the same services both of you get. What will you do then?

Who wants to buy the same ticket with the same services by paying 50% higher price?! If they had some idea about the changing behavior of the airfare price, they could save some money. This is so frustrating for the customers and also an unbearable problem for them.
We are trying to do something by which the customers can know the lowest and best prices of airline tickets and the best time to purchase the tickets from the airlines.

1.3 Problem Definition

Nowadays, people are getting interested in traveling. Journey by airplane is now getting very popular. And purchasing airline tickets is now extremely popular. In the meantime, the airlines try to collect a huge amount of revenue by increasing the price of airline tickets. Customers usually want to buy cheaper tickets for their travel, on the other hand, the airlines seek to keep their overall revenue as high as possible to maximize their profit. They use some prediction system to know the demand for tickets and when and where the customers want to travel. They developed their own prediction system to set their ticket price. They use some management theories and sophisticated mathematical models to know the real-time airlines' ticket prices which are based on unsold tickets and customer demand. For that reason, we are trying to create a prediction model by using a regression model with the proper approach of Machine learning.

1.4 Research Objectives

The airlines' ticket prices are not stable all the time. Sometimes the prices can be higher or lower than before. Flight tickets prices are set based on unsold seats and recent market demand. It is the commercial secret of airlines. For that reason, it is difficult for travel agencies and customers to estimate how the price of airline tickets will change in the future. Forecasting the price of flight tickets can be so helpful for the travel agencies and customers to decide the best time to buy or purchase the tickets from the airlines.

We will use some algorithms to predict the prices of airline tickets. From the results of these algorithms we can find out the best predicting model to predict the airfare prices with a better accuracy. Predicting the airline tickets prices by which the customers can be sure about the best time to purchase the tickets. The satisfaction of the customer is that they can purchase their desired tickets at reasonable prices at the best time.

1.5 Research Layout

Chapter 1: In this section, we will discuss introduction, motivation, Problem Definition, Research Objectives.

Chapter 2: In this section, we will discuss the literature review.

Chapter 3: Will discuss the Research Methodology.

Chapter 4: Training and Testing of the model will be discussed here.

Chapter 5: Result, Comparison and Discussion will be discussed here.

Chapter 6: It will describe the conclusion and future work of this research.

CHAPTER 2

BACKGROUND

2.1 Introduction

We are living in a modern era. People are getting interested in traveling and they are now traveling by airplane. But a problem arises which is the prices of the airfare. The prices of airfare are not stable. It changes very fast even if you can’t buy a ticket in the evening with the same price which you bought in the morning. For that reason the researchers are trying to build a system that will predict the prices of airfare for the customers which will save a huge amount of customer’s money.

In the literature review section we will be focusing on some of the previous work.

2.2 Related works

Nowadays Airfare price prediction is very important for us because the prices are unpredictable. There is a huge difference in the price between the morning and evening tickets. It changes abruptly. Since the factors like fuel price, purchasing time, flight distance have been involved in the pricing of airfare, the prices are changing dynamically and it is being so unbearable for the customers. These factors are the main reason why the prices fluctuate and also the reason why the prediction of the air ticket is very challenging. In [4], the researchers proposed to build a model using Naive Bayes, Softness regression, Support Vector Machines (SVMs) and Linear Regression (LR). The authors got the best training error result was 22.9% using LR model and the result of SVM regression model was not satisfying.

In [5], the authors wanted to get the best fit model from the four LR models. They compared the four LR models for this research. From this prediction system the passengers will get unbiased information about the ticket prices whether to buy the ticket or wait for the lower or affordable cost. For predicting the lowest airfare prices, the researchers suggested using linear quantile mixed models, which they called the “real bargains". But there are limitations to these models. It works only for economy class tickets.

The researchers of [6] used a multi-strategy data mining technique which is called HAMLET. They used crawled data from the web for their research. They assumed that, by using that data mining technique, this prediction can save the cost for the customers. But there is a problem: the key variable is missing here like the number of seats [17].

In this paper [7], to predict the air ticket sales revenue the authors applied Genetic Algorithm and Artificial Neural Network (ANN). They used the Taiwan stock market weighted index, monthly unemployment rate and international oil price as input features etc. To improve the performance of ANN the Genetic Algorithm Selects the finest input features. They got their models Mean Absolute Percentage Error 9.11% which was good.

In the recent year we can see that, for improving the airfare price prediction more advanced ML models have been created and used. In this paper [8] for predicting the price of airfare, the authors applied eight Machine Learning Algorithms such as SVM, LR and ANNs and so on. And they compared those models' performance and they got the best model in their comparison which was Regression tree. In [9], to get more accuracy in prediction, the researchers proposed Deep Regress or Stacking.

The authors of this paper [13], applied the Partial Least Square Regression model for optimizing the airfare prices. They got the accuracy of their research 75.3%.

In this paper [14], the researcher applied three models Logistic Regression (69.9%), Support Vector Machine (69.4%) and Ripple down Rule Learner (74.5%). By these predictions he described that, in the future, the price of the airfare will drop. In [5], for predicting the price of airfare with acceptable performance the researcher applied a Linear Quantile Mixed Regression model. The authors of [4], applied four models to find out the best model for predicting the prices of airfare. To predict the airfare prices and find out the best model, they studied the performance of the Naive Bayes (73.06%), Support Vector Machine (80.6%), Softmax Regression (76.84%) and Linear Regression (77.06%).

Table 2.2.1: Some results of previous works.

Authors	Algorithms	Accuracy
A Framework for Airfare Price Prediction: A Machine Learning Approach [3].	Random Forest Regression	86.9%
Airfare prices prediction using machine learning techniques [8].	Bagging Regression Tree	87.91%
A regression model for predicting optimal purchase timing for airline tickets [13].	Partial least Square	75.3%
Predicting Airfare Prices [14].	Logistic Regression, SVM	69.9%, 69.4%
Prediction of airline ticket price [4].	Linear Regression, Naïve Bayes, Softmax Regression, SVM.	77.06%, 73.06%,76.84%, 80.6%.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this research, we used three algorithms Random Forest Regression, Decision Tree Algorithm and Linear Regression to find out the best one which can predict the best price for the passengers. We got the best model to predict the airfare which is Random Forest Regression with the best accuracy 90.47%.

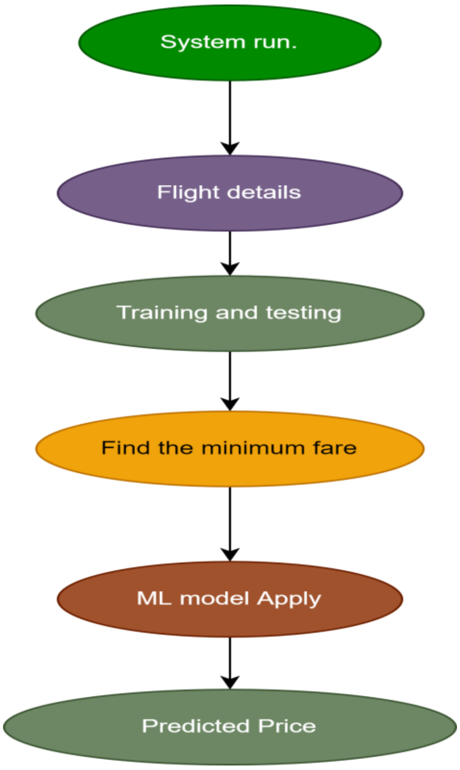


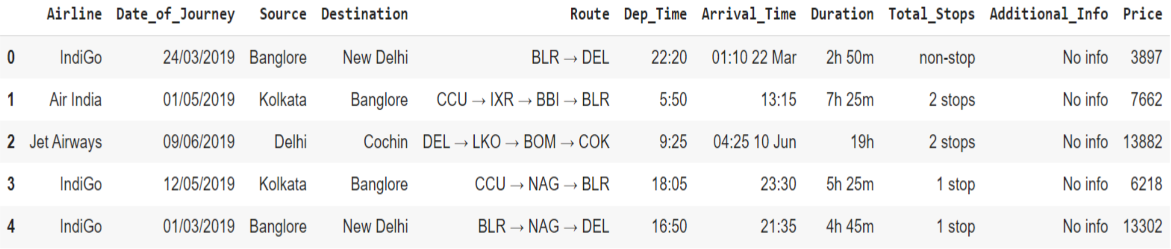
Figure 3.1.1: Flow diagram of models.

3.2 Experiment Data Set

We collected the dataset from kaggle. This dataset consists of 10000 records. We wanted to collect data from our country but the collecting process is not so easy. For that reason we used this public dataset from kaggle. We have imported some libraries such as pandas, numpy, Matplotlib, seaborn etc. After that we imported our dataset to run and we got the train data shape (10000,11). We have train data info such as

Airline, Date of Journey, Source, Destination, Route, Departure time, Arrival time, Duration, Total stops, Additional Info and price segment in our data set.

Here is a blink of our collected dataset.



	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	01/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	5:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	09/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	9:25	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

Figure 3.2.1 Collected Dataset

3.3 Data Pre-Processing

To get better accuracy results from the model, clean data is needed and it will be great quality work. Data preprocessing is an important part. First we need to clean the data set. We checked null values and we removed NAN values from the stops column and got two types of data such as object and integer. After that, we described the object features types and also for numerical. We counted the total number of flights and checked the price variation of each airline. We counted the number of flights for every source and destination. After that we checked the price variation of each stop and checked the time range. We converted the duration column Duration_in_min. We replaced the duration time by using the lambda function. Then we showed the plot between Duration_in_min and Price.

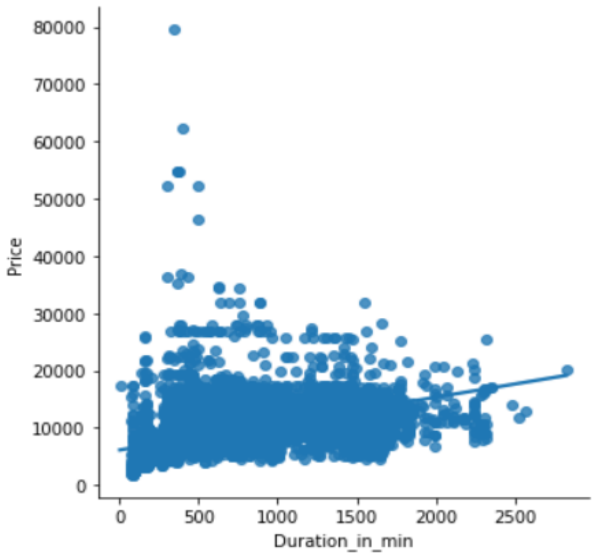
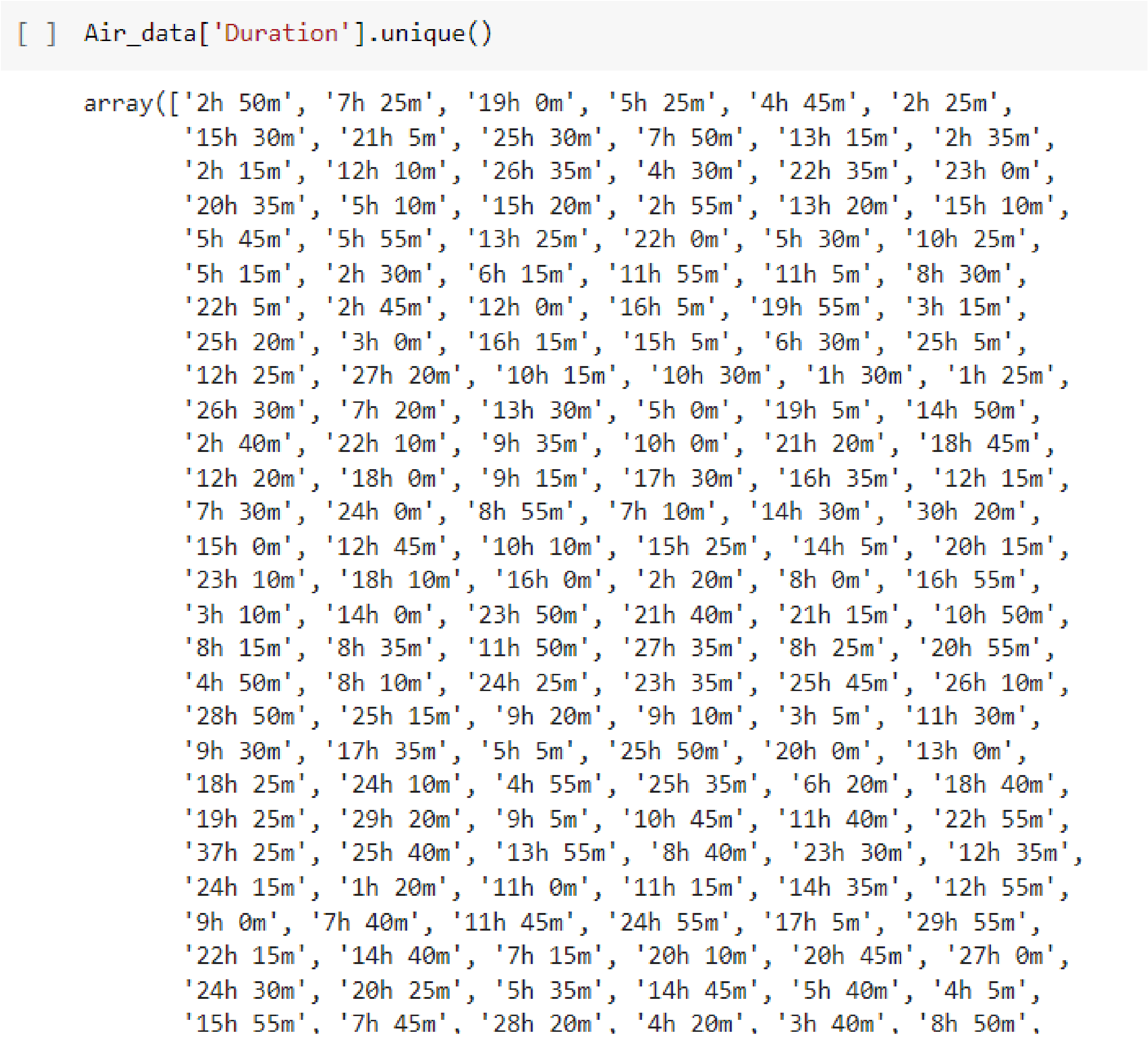


Figure 3.3.1: Plot between Duration_in_min and Price

We counted the unique Duration from the dataset.

Figure 3.3.2 is the representation of unique Duration.



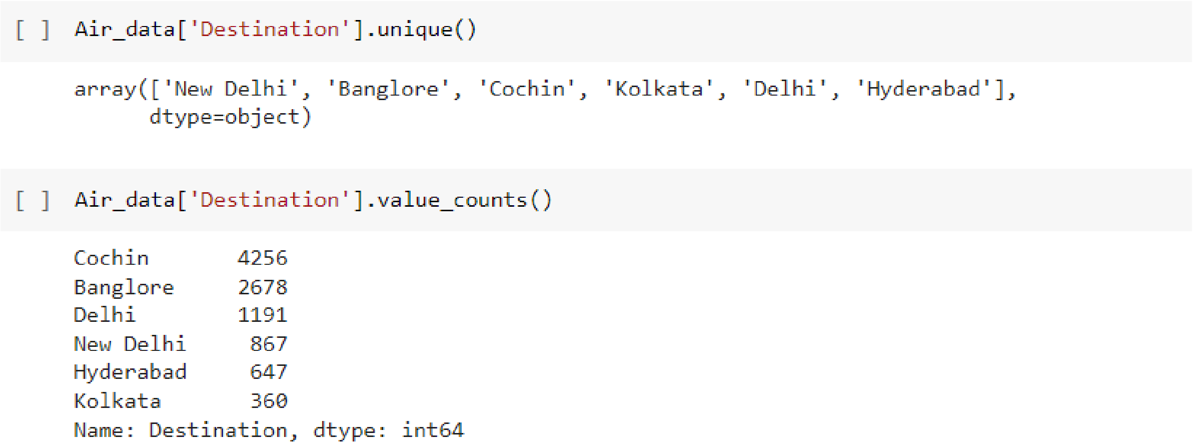
```
[ ]  Air_data['Duration'].unique()

     array(['2h 50m', '7h 25m', '19h 0m', '5h 25m', '4h 45m', '2h 25m',
            '15h 30m', '21h 5m', '25h 30m', '7h 50m', '13h 15m', '2h 35m',
            '2h 15m', '12h 10m', '26h 35m', '4h 30m', '22h 35m', '23h 0m',
            '20h 35m', '5h 10m', '15h 20m', '2h 55m', '13h 20m', '15h 10m',
            '5h 45m', '5h 55m', '13h 25m', '22h 0m', '5h 30m', '10h 25m',
            '5h 15m', '2h 30m', '6h 15m', '11h 55m', '11h 5m', '8h 30m',
            '22h 5m', '2h 45m', '12h 0m', '16h 5m', '19h 55m', '3h 15m',
            '25h 20m', '3h 0m', '16h 15m', '15h 5m', '6h 30m', '25h 5m',
            '12h 25m', '27h 20m', '10h 15m', '10h 30m', '1h 30m', '1h 25m',
            '26h 30m', '7h 20m', '13h 30m', '5h 0m', '19h 5m', '14h 50m',
            '2h 40m', '22h 10m', '9h 35m', '10h 0m', '21h 20m', '18h 45m',
            '12h 20m', '18h 0m', '9h 15m', '17h 30m', '16h 35m', '12h 15m',
            '7h 30m', '24h 0m', '8h 55m', '7h 10m', '14h 30m', '30h 20m',
            '15h 0m', '12h 45m', '10h 10m', '15h 25m', '14h 5m', '20h 15m',
            '23h 10m', '18h 10m', '16h 0m', '2h 20m', '8h 0m', '16h 55m',
            '3h 10m', '14h 0m', '23h 50m', '21h 40m', '21h 15m', '10h 50m',
            '8h 15m', '8h 35m', '11h 50m', '27h 35m', '8h 25m', '20h 55m',
            '4h 50m', '8h 10m', '24h 25m', '23h 35m', '25h 45m', '26h 10m',
            '28h 50m', '25h 15m', '9h 20m', '9h 10m', '3h 5m', '11h 30m',
            '9h 30m', '17h 35m', '5h 5m', '25h 50m', '20h 0m', '13h 0m',
            '18h 25m', '24h 10m', '4h 55m', '25h 35m', '6h 20m', '18h 40m',
            '19h 25m', '29h 20m', '9h 5m', '10h 45m', '11h 40m', '22h 55m',
            '37h 25m', '25h 40m', '13h 55m', '8h 40m', '23h 30m', '12h 35m',
            '24h 15m', '1h 20m', '11h 0m', '11h 15m', '14h 35m', '12h 55m',
            '9h 0m', '7h 40m', '11h 45m', '24h 55m', '17h 5m', '29h 55m',
            '22h 15m', '14h 40m', '7h 15m', '20h 10m', '20h 45m', '27h 0m',
            '24h 30m', '20h 25m', '5h 35m', '14h 45m', '5h 40m', '4h 5m',
            '15h 55m'. '7h 45m'. '28h 20m'. '4h 20m'. '3h 40m'. '8h 50m'.
```

Figure 3.3.2: Unique Durations.

After that, we found a number of unique destinations such as New Delhi, Bangalore, Cochin, Delhi and Hyderabad.

Figure 3.3.3 is the representation of unique destinations.



```
[ ]  Air_data['Destination'].unique()

     array(['New Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Delhi', 'Hyderabad'],
           dtype=object)

[ ]  Air_data['Destination'].value_counts()

     Cochin       4256
     Banglore     2678
     Delhi        1191
     New Delhi     867
     Hyderabad     647
     Kolkata       360
     Name: Destination, dtype: int64
```

Figure 3.3.3 : Unique destinations.

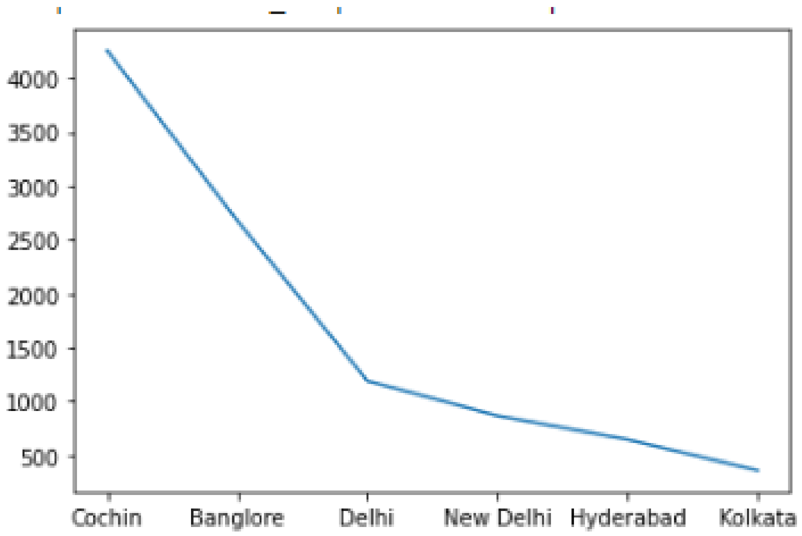


Figure 3.3.4: Graph of the unique destinations

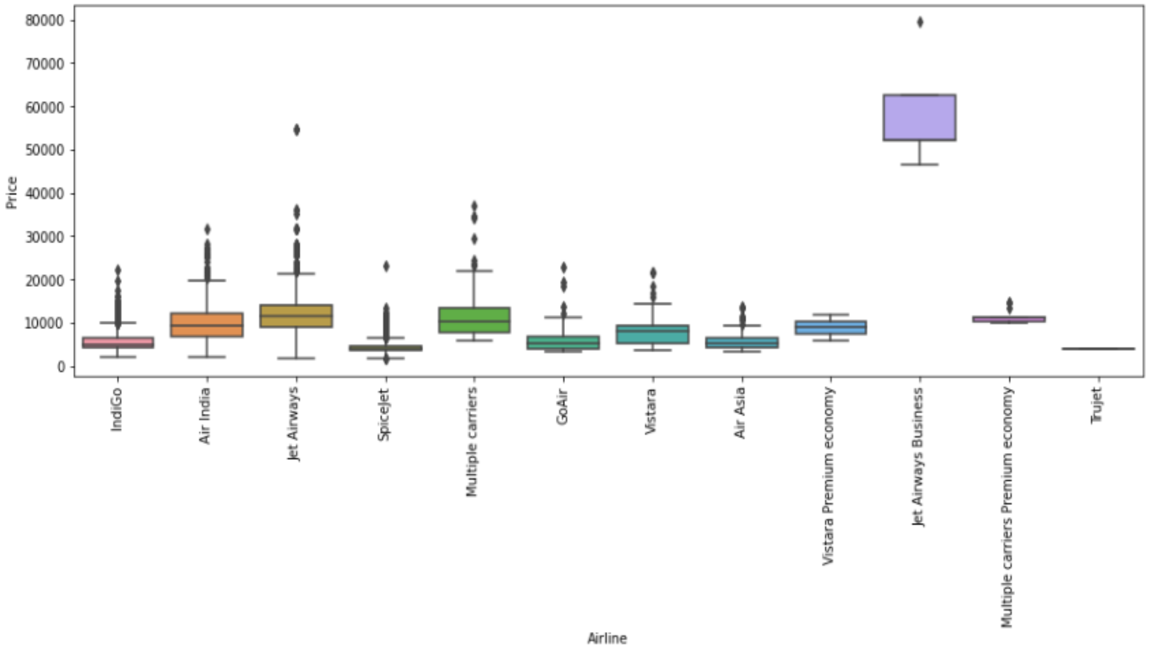
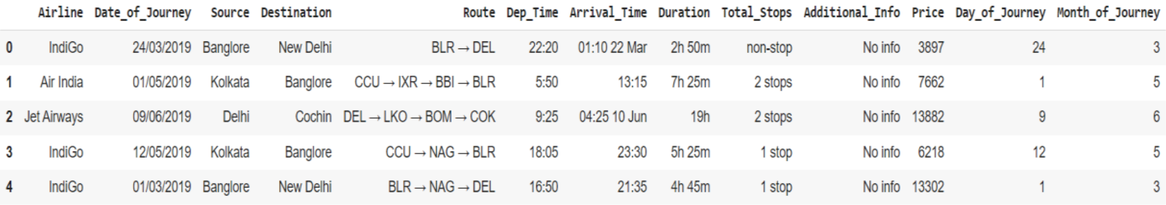


Figure 3.3.5: Airlines and their price differences.

3.4 Feature Extraction

We selected some features to extract such as Date of journey, Arrival time, Departure time. We can see that the year of the journey is 2019. For that, we just extracted the date of the journey column into two portions. We derived the min, max values of date of journey and extract it to journey date, journey month.

Figure 3.4.1 is presenting the dataset after extraction of the Date_of_Journey into Day_of_journey and Month_of_Journey.

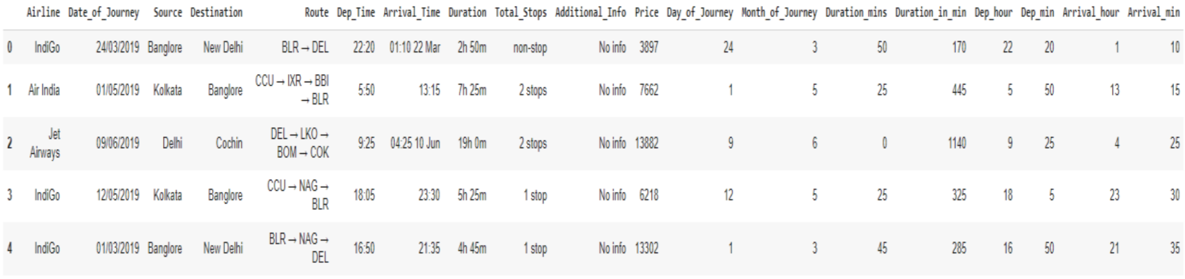


	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Day_of_Journey	Month_of_Journey
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897	24	3
1	Air India	01/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	5:50	13:15	7h 25m	2 stops	No info	7662	1	5
2	Jet Airways	09/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	9:25	04:25 10 Jun	19h	2 stops	No info	13882	9	6
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218	12	5
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302	1	3

Figure 3.4.1: After extracting the Date_of_Journey Column

After extracting the Date_of_Journey column we extracted the Arrival_time into Arrival_hour and Arrival_min. And also extracted Departure_time into Dep_hour and Dep_min.

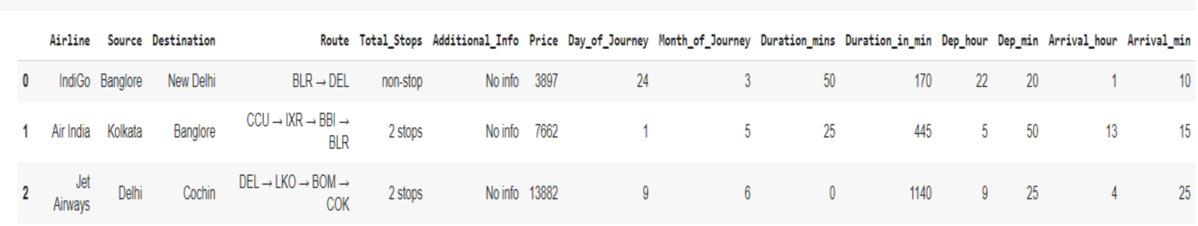
figure 3.4.2 is presenting the dataset after extracting the Arrival_time and Dep_time column into Arrival_hour, Arrival_min and Dep_hour, Dep_min.



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Day_of_Journey	Month_of_Journey	Duration_mins	Duration_in_min	Dep_hour	Dep_min	Arrival_hour	Arrival_min
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897	24	3	50	170	22	20	1	10
1	Air India	01/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	5:50	13:15	7h 25m	2 stops	No info	7662	1	5	25	445	5	50	13	15
2	Jet Airways	09/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	9:25	04:25 10 Jun	19h 0m	2 stops	No info	13882	9	6	0	1140	9	25	4	25
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218	12	5	25	325	18	5	23	30
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302	1	3	45	285	16	50	21	35

Figure 3.4.2: After extracting Arrival_time and Dep_time.

After preprocessing we got the unique duration and destination. Then we took the additional info from the preprocessed data, counted it and divided it with data length into 100 and converted it into a numpy array. After that, we dropped the route, additional info, Duration total mins and journey year from the preprocessed dataset.



	Airline	Source	Destination	Route	Total_Stops	Additional_Info	Price	Day_of_Journey	Month_of_Journey	Duration_mins	Duration_in_min	Dep_hour	Dep_min	Arrival_hour	Arrival_min
0	IndiGo	Banglore	New Delhi	BLR → DEL	non-stop	No info	3897	24	3	50	170	22	20	1	10
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	2 stops	No info	7662	1	5	25	445	5	50	13	15
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	2 stops	No info	13882	9	6	0	1140	9	25	4	25

Figure 3.4.3: After dropping the replicated columns.

3.5 Architecture of the Model

The model is designed to predict the airfare prices for the passengers by which the passengers can save the money.

The architecture diagram of our models is given below. It shows the processes of how the model has been built. We used some libraries such as numpy, pandas, matplotlib. We also imported seaborn. After collecting the dataset we performed data pre-processing including data cleaning and after that, we

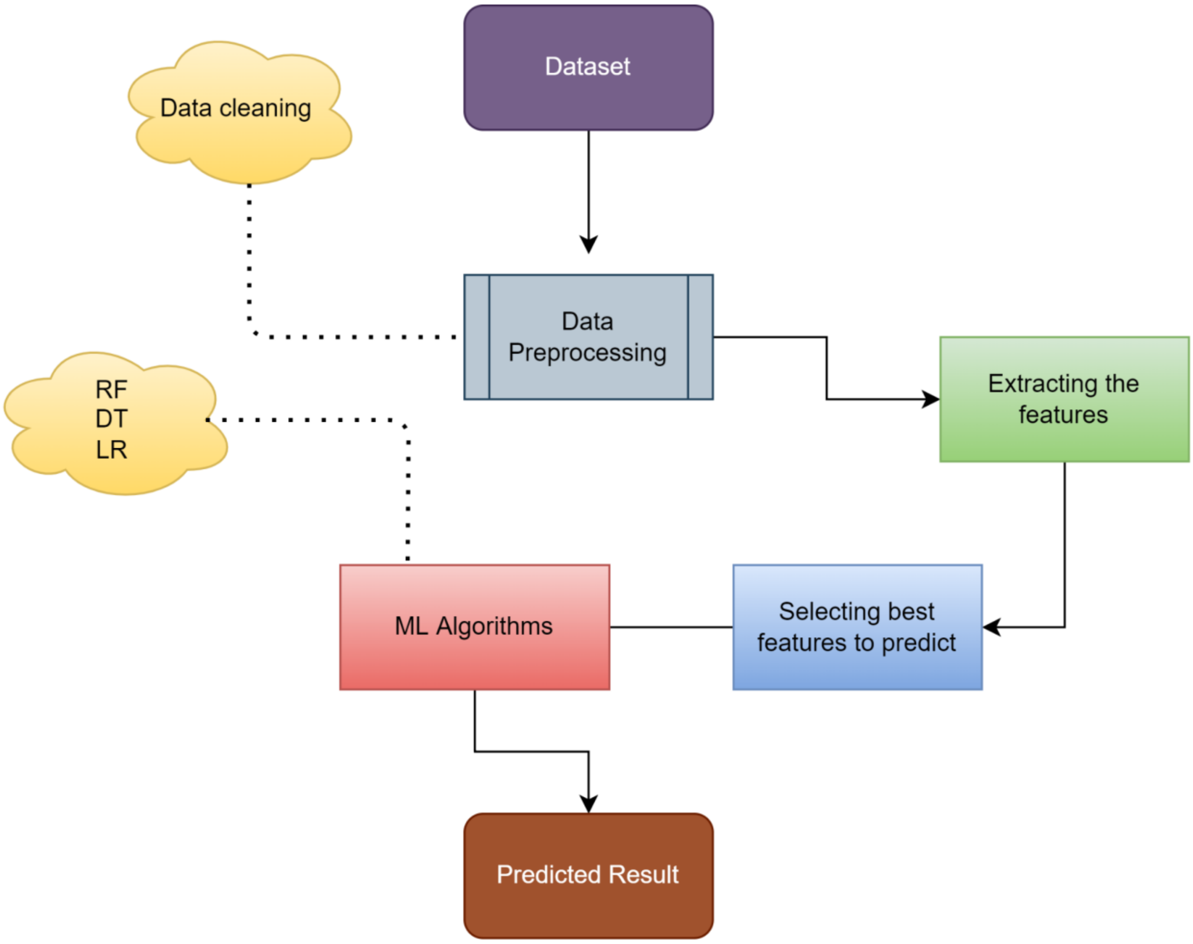


Figure 3.5.1: Architecture of the Model.

3.6 Descriptions of the used Algorithms

We have used three Algorithms such as Random Forest Regression, Decision Tree, and Linear Regression to find the best model among them.
The description of those algorithms are given below.

Random Forest Algorithm:

Random Forest Regression uses a method for regression which is called ensemble learning method. It is a technique that combines predictions from multiple decision trees to make a more accurate prediction than a single model. It is a supervised Machine Learning Algorithm.

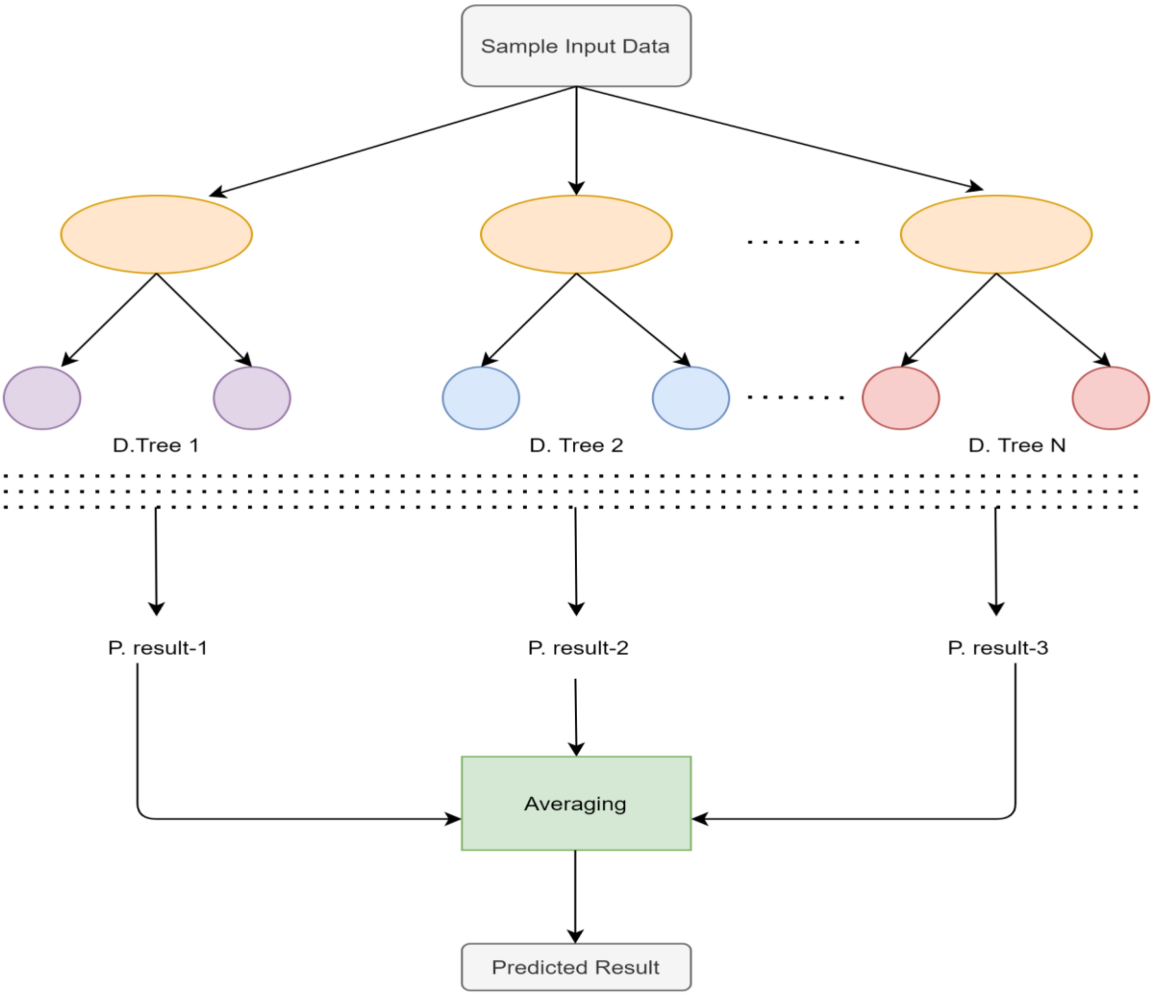


Figure 3.6.1: Random Forest Algorithm

We can see the diagram of the Random Forest Algorithm above. In this algorithm the trees are not connected amongst themselves. During the training time this algorithm works by making various decision trees and gathering the results of all decision trees and using the mean value for predicting the final result [10][11].

The step by step process of Random Forest Regression Algorithm,

1. We have to pick a random data x from the training dataset.
2. Corresponding to data point x, build a decision tree.
3. Select the number of N trees which will be built and repeat step 1 & 2.

4. For a new data point, make each one of N-trees predict the value of y for the data point in question and assign the new point to the average across all of the predicted y values.

Decision Tree Algorithm:

Decision Tree Algorithm is a supervised learning algorithm which can be used for both classification and regression problems. It is a tree-structured classifier, where the internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome [15].

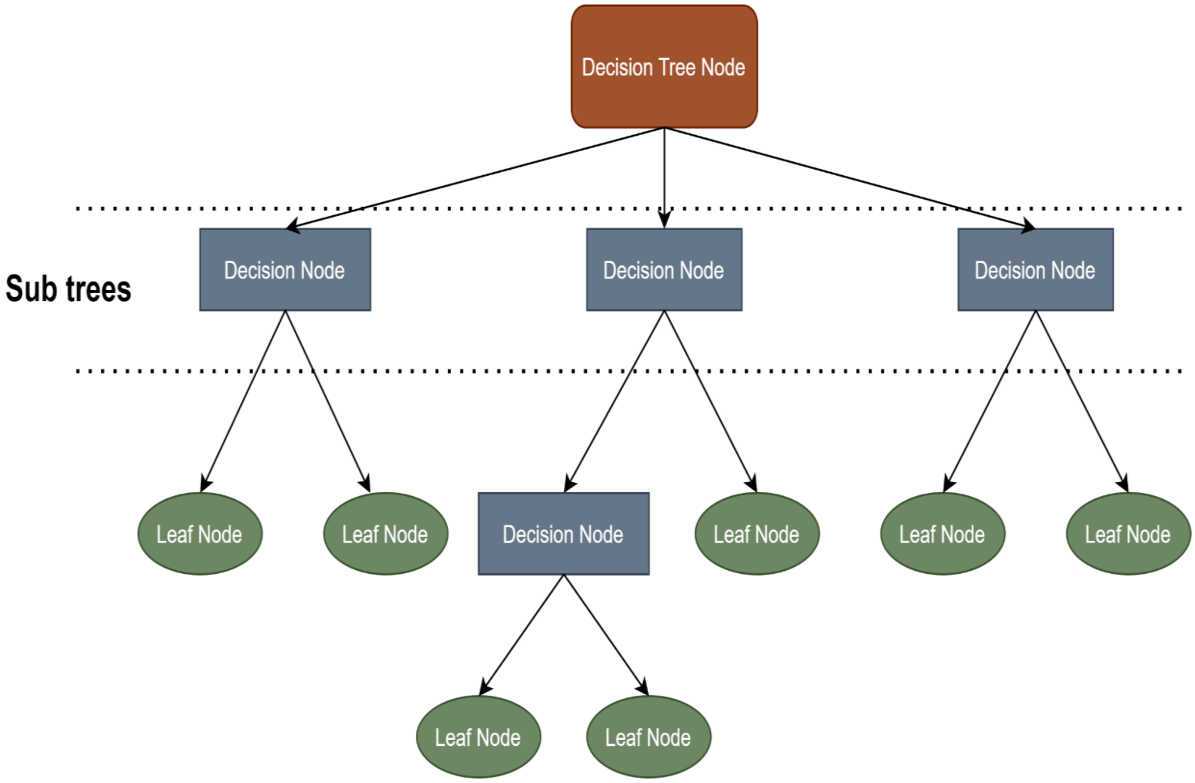


Figure- 3.6.2: Decision Tree Algorithm

The step by step process of Decision Tree Algorithm,

- ❖ Step-1: Start the tree with the root node, says X, which contains the complete dataset.

- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM)
- Step-3: Divide the X into subsets that contain possible values for the best attribute.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in the step-3. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.

Linear Regression Algorithm:

Linear Regression is a ML algorithm based on supervised learning. It performs a regression task. It is mostly used for finding out the relationship between variables and forecasting. Different types of regression models differ based on the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. Linear Regression Algorithm performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name of that algorithm is Linear Regression Algorithm [16].

CHAPTER 4

TRAINING AND TESTING OF THE MODEL

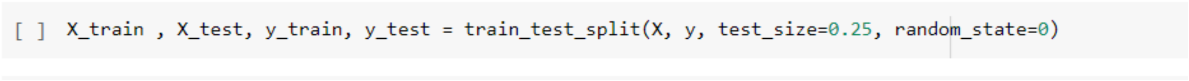
4.1 Training and testing

We splitted training data and test data. The total data is around 10000. To train the model we split the dataset into two two sets such as train and test set. 75% of our data is for the train set and 25% of the dataset for the test. After that, we put the features in x = all of the features except price.

y = price.

we put all of the features in x except price and put the price in y.

Then we fitted x and y into mutual info regression where the values are sorted by column importance in descending order. Now we imported train_test_split from sklearn model selection. Where test set size is 25% of the total training set size. We performed training with the train set and prediction with the test set.

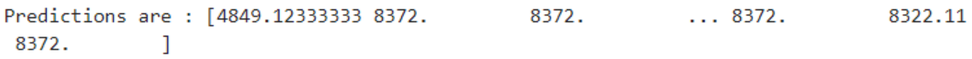


```
[ ]  X_train , X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=0)
```

Figure 4.1.1: Training of the model.

4.2 Implementation of our Model.

We gained the training score 92.2% and after a successful training we sent our test set into our prediction function where the used model is a Random Forest Regression. Then we opened the random forest pickle file and used this parameter corresponding to our test set. Now, at the end of our prediction function, from the y_train and y_test split our predicted price comes as an output. These show an array that indicates the predicted price of airfare.



```
Predictions are : [4849.12333333 8372.        8372.        ... 8372.        8322.11
 8372.        ]
```

Figure 4.2.1: Predicted prices

CHAPTER 5

RESULT COMPARISON AND ANALYSIS

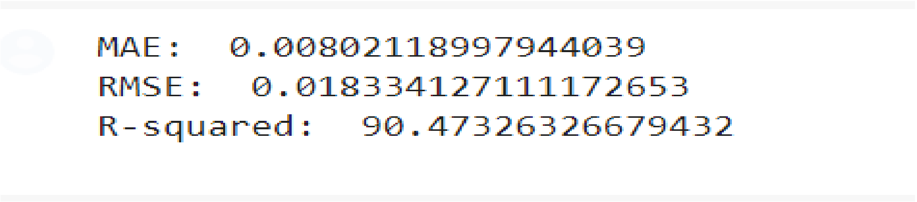
After importing all the important libraries, dataset, finishing the data preprocessing and dropping any columns for necessity, our ML models are ready.

Finally the Machine Learning models are ready, and x_train and y_train are fitted into these models. After a successful training x test set can be used to predict the y test set where y test set is the actual price of the airline ticket. We also applied Decision Tree Algorithm and Linear Regression Algorithm as well.

Here, in this case the prediction shape is 2500. Now we can use the model to predict the price of airfare. With the accurate training score there will be an accurate prediction. For that dataset we got better accuracy in Random Forest Regression 90.47%. The results of the Decision Tree Algorithm is 79.20% and for the Linear Regression Algorithm is 72.77% which are not so bad. We can see the differences of the results we got from that research.

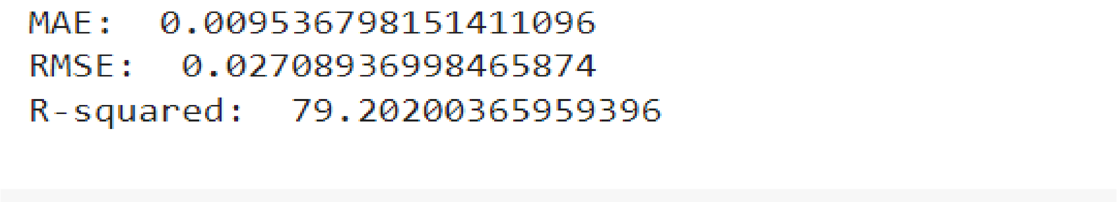
Table 5.1: Comparison of our research

Algorithm	Accuracy
Random Forest Regression	**90.47%**
Decision Tree	79.20%
Linear Regression	72.77%



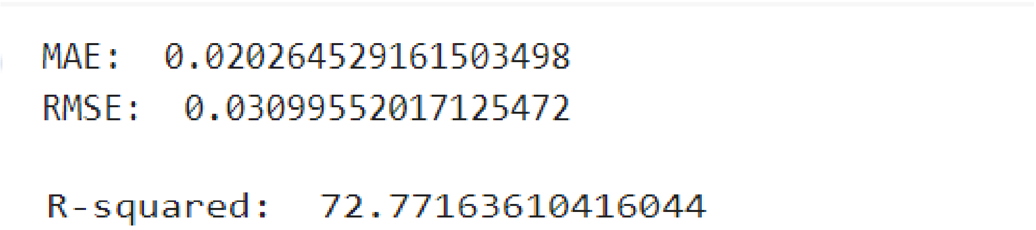
```
MAE:   0.00802118997944039
RMSE:   0.01833412711172653
R-squared:   90.47326326679432
```

Figure 5.1: Random Forest Regression result.



```
MAE:   0.009536798151411096
RMSE:   0.02708936998465874
R-squared:   79.20200365959396
```

Figure 5.2: Decision Tree Algorithm result.



```
MAE:   0.020264529161503498
RMSE:   0.03099552017125472

R-squared:   72.77163610416044
```

Figure 5.3: Linear Regression Algorithm result

Table 5.2 represents the comparison of our work with some previous works.

Table 5.2: Comparison with others work.

Authors	Algorithms	Accuracy
A Framework for Airfare Price Prediction: A Machine Learning Approach [3].	Random Forest Regression	86.9%
Airfare prices prediction using machine learning techniques [8].	Bagging Regression Tree	87.91%
A regression model for predicting optimal purchase timing for airline tickets [13].	Partial least Square	75.3%
Predicting Airfare Prices [14].	Logistic Regression, SVM	69.9%, 69.4%
Prediction of airline ticket price [4].	Linear Regression, Naïve Bayes, Softmax Regression, SVM.	77.06%, 73.06%,76.84%, 80.6%.
Airlines Ticket Price Prediction Using Machine Learning Approach.	**Random Forest Regression, Decision Tree, Linear Regression.**	**90.47%, 79.20%, 72.77%**

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

In this research, we developed a system that can predict the airfare price for the consumers. It will help them to save more money which is more important for them. There are a lot of machine learning algorithms that have been used in the previous research to predict the airfare price such as Support Vector Machine (SVM), K-Nearest Neighbors (KNN) etc. We applied three ML algorithms such as Random Forest Regression Algorithm, Decision Tree Algorithm, and Linear Regression Algorithm to compare their results and find out the best model to predict the airfare prices for the passengers. After a successful training we got the best model with high accuracy. Our best model accuracy for the Random Forest Regression is 90.47% which is huge. We found that the Random Forest gives better accuracy then the other Algorithms which is more important. It takes a short time to predict the price which is also helpful for the customer. The customers can make the decisions within a short time whether they need to buy tickets at low cost or not.

6.2 Future Work

In the future, we will try to improve the accuracy of our prediction model. To improve our prediction models performance, we will try to collect more data with more features which will be added to our model, such as available seats, whether the day of departure is a holiday or not etc. It will make our system more perfect and performance of the prediction model will be better and the accuracy will be more efficient.

REFERENCES

[1]. B. Mantin and B. Koo, “Dynamic price dispersion in airline markets,” Transportation Research Part E: Logistics and Transportation Review, vol. 45, no. 6, pp. 1020–1029, 2009.

[2]. J. Stavins, “Price discrimination in the airline market: The effect of market concentration,” Review of Economics and Statistics, vol. 83, no. 1, pp. 200–202, 2001.

[3]. Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso, Steven Luis,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), 2019.

[4]. R. Ren, Y. Yang and S. Yuan, “Prediction of airline ticket price,” Technical Report, Stanford University, 2015.

[5]. T. Janssen, T. Dijkstra, S. Abbas, and A. C. van Riel, “A linearquantile mixed regression model for prediction of airlineticket prices,” Radboud University, 2014.

[6]. Oren Etzioni, Rattapoom Tuchinda, Craig A Knoblock, and Alexander Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 119–128. ACM, 2003

[7]. H.-C. Huang, “A hybrid neural network prediction model of air ticket sales,” Telkomnika Indonesian Journal of Electrical Engineering, vol. 11, no. 11, pp. 6413–6419, 2013.

[8]. K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, “Airfare prices prediction using machine learning techniques,” in the 25th IEEE European signal processing conference, 2017, pp. 1036–1039.

[9]. E. J. Santana, S. M. Mastelini, and S. Barbon Jr, “Deep regressor stacking for air ticket prices prediction,” in the XIII Brazilian symposium on information systems: information systems for participatory digital governance. Brazilian Computer Society (SBC), 2017, pp. 25–31

[10]. Academia <https://www.academia.edu/45565822/A_Machine_Learning_Approach_to_Predict_Price_of_Airlines_Tickets> sunday, 14th August, 2022, 11.48 pm.

[11]. levelup coding. Available at <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84> Last accessed on saturday, 3rd september, 2022, 11.46 pm.

[12]. javaTpoint. Available at <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>Last accessed on sunday, 4th september, 2022, 12.10 am.

[13]. W. Groves and M. Gini, “A regression model for predicting optimal purchase timing for airline tickets,” Technical Report 11-025, University of Minnesota, Minneapolis, 2011.

[14]. M. Papadakis, “Predicting Airfare Prices,” 2014.

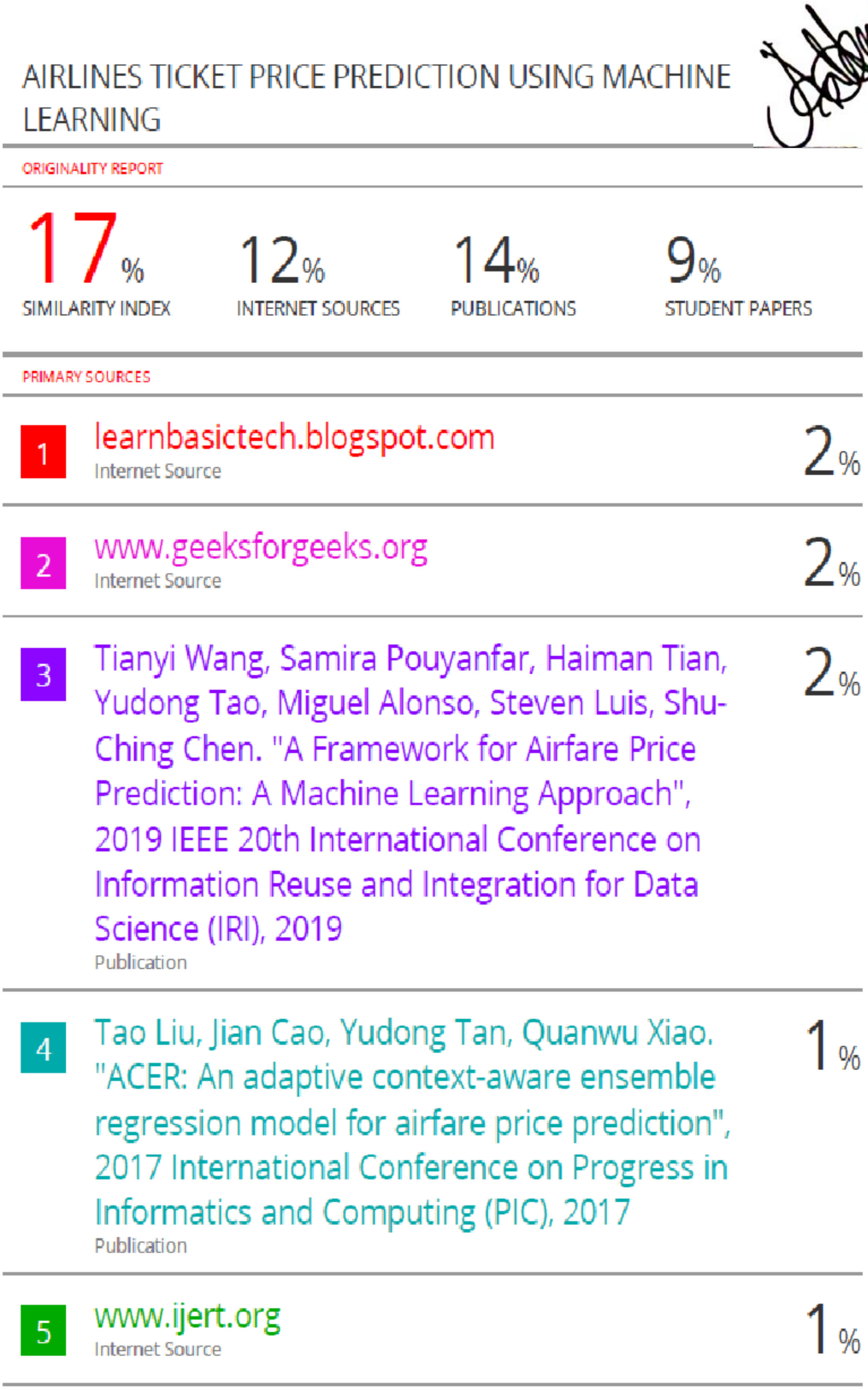
[15]. javaTpoint. Available at<https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm> Last accessed on sunday. 4th september, 2022, 8.20 pm.

[16]. GeeksforGeeks. Available at <https://www.geeksforgeeks.org/ml-linear-regression> Last accessed on sunday, 4th september, 2022, 8.39 pm.

[17]. 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), 2017

PLAGIARISM REPORT

21/09/22



AIRLINES TICKET PRICE PREDICTION USING MACHINE LEARNING

ORIGINALITY REPORT

17%	12%	14%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	learnbasictech.blogspot.com Internet Source	2%
2	www.geeksforgeeks.org Internet Source	2%
3	Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso, Steven Luis, 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), 2019 Publication	2%
4	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), 2017 Publication	1%
5	www.ijert.org Internet Source	1%