Shape

Description automatically generated with medium confidence



**Essay / Assignment Title:** **Predictive Analytics and Machine Learning Project Report**

**Programme title:MSc of Data Analytics**

**Name:Turan Can Gün - Q1055616**

**Year:2024 (2023 October Intake)**

**CONTENTS**

[INTRODUCTION 3](#_Toc1321210064)

[CHAPTER ONE Problem Formulation 4](#_Toc1416873324)

[CHAPTER TWO Data Collection and Preparation 5](#_Toc1201065261)

[2.1 Libraries 6](#_Toc224658235)

[2.2 Overview of the Dataset 7](#_Toc1303257412)

[CHAPTER TWO Exploratory Data Analysis 8](#_Toc821285476)

[2.1 Distribution of Source City 9](#_Toc1289313937)

[2.2 Distribution of Destination City 9](#_Toc948321514)

[2.3 Distribution of Class 10](#_Toc2106376117)

[2.3 Distribution of Airlines 11](#_Toc1665963864)

[2.3 Distribution of Stops 12](#_Toc1974879964)

[2.4 Distribution Departure – Arrival Time 12](#_Toc2007915999)

[CHAPTER THREE Model Selection and Implementation 13](#_Toc356446618)

[3.1 Model Selection 14](#_Toc1375590271)

[3.2 Implementation 14](#_Toc1093570797)

[3.2.1 Encoding and Scaling 14](#_Toc2016169599)

[3.2.2 Correlation Matrix 15](#_Toc297503952)

[3.2.3 Splitting Data Into Training and Test Part 16](#_Toc1201531914)

[3.2.4 Linear Regression 17](#_Toc383237417)

[3.2.5 Decision Tree Regressor 17](#_Toc1436398754)

[3.2.6 Random Forest Regression 17](#_Toc1444018035)

[3.2.7 K-Nearest Neighbor 18](#_Toc2092463367)

[CHAPTER FOUR Model Evaluation 19](#_Toc120319990)

[CHAPTER FIVE Conclusion and Recommendations 21](#_Toc1738928376)

[CHAPTER SIX Project Reflection 23](#_Toc1765045194)

[BIBLIOGRAPHY 24](#_Toc131853910)

[APPENDIX (if necessary) 25](#_Toc151787579)

**Statement of compliance with academic ethics and the avoidance of plagiarism**

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters): TURAN CAN GÜN

.......................................................................................................................................

Date: 08/02/2024

# INTRODUCTION

Predictive analytics and machine learning are really important for using information to make smart decisions. They help organizations look at past information to figure out what might happen in the future and what problems could come up. Machine learning algorithms help with this by getting better at understanding patterns over time. These tools help businesses to be ready for challenges, make things work better, and make smarter choices using all the information available. This can give them an advantage in the fast-paced world of business today.

During the COVID-19 duration most organizations, especially small and medium enterprises, have been struggling to survive based on tourism sector (Lee & Trimi, 2021). Guessing how much a plane ticket will cost is really important in the travel industry. It helps airlines and travel companies figure out how much to charge for tickets based on past information and other things that affect demand. This helps them offer prices that are competitive and fair. It also helps them change prices quickly if things in the market change. This makes customers happy and helps companies make more money. For customers, knowing how much a flight will cost helps them plan and save money. When companies can guess the right prices, it makes the travel industry work better for everyone.

# CHAPTER ONE Problem Formulation

Predictive analytics and machine learning help us figure out how much plane tickets will cost in the future. This is important because airfare can be really complicated and change a lot. By looking at past data considering things like how many people want to fly , what time of year it is.We can models that are really good at guessing how much tickets will cost.The tricky part is doing this fast enough so that airlines and travel websites can change prices in real-time.The goal is to help airlines make more money while also making sure that customers get a fair and competitive price.This problem involves making smart algorithms that are always learning and getting better at predicting prices, even when things are always changing in the travel industry.

Solving the problem of predicting prices of air flight ticket using predictive analytics and machine learning is of great importance in the travel industry business landscape for several reasons.

First, accurate price forecasting allows airlines and travel agents to implement dynamic pricing strategies to optimize revenue by adjusting ticket costs to changing demand and market conditions.This improves the financial performance of companies in the highly competitive aviation industry.

Next, it is important to resolve this issue to improve customer satisfaction.Transparent and predictable pricing allows travelers to plan and book flights at the best time, saving money and increasing trust in travel companies.Reduce frustration that can result from unexpected price changes and contribute to a better customer experience.

In summary, solving the challenge of predicting airfare using advanced analytics can help maximize revenue, increase customer satisfaction, and improve competitive advantage in the travel industry's dynamic and fast-changing business environment.

# CHAPTER TWO Data Collection and Preparation

## 2.1 Libraries



Figure 2.1 Order of Used Libraries

A number of crucial libraries for machine learning, data analysis, and visualization are imported by the Python script. The pandas library is used for analysis and manipulation of data, especially with its robust DataFrame format. Numerical operations on arrays and matrices are supported by numpy.

Matplotlib.pyplot is used to build visualizations, while the seaborn library is used to enrich them with aesthetically beautiful statistical visuals. The scikit-learn package, which contains preprocessing tools (StandardScaler, MinMaxScaler), clustering algorithms (KMeans, AgglomerativeClustering, DBSCAN), and ensemble learning techniques (RandomForestRegressor, AdaBoostRegressor, GradientBoostingRegressor), facilitates machine learning tasks. The scipy.cluster.hierarchy module facilitates the construction of dendrograms and hierarchical clustering.The operating system module manages interfaces with the operating system.

Functions from sklearn.model\_selection and sklearn.metrics are used for model evaluation and selection.While sklearn.tree offers decision tree-based models, sklearn.linear\_model provides logistic regression and linear models. The script can perform a wide range of data analysis and machine learning tasks, such as clustering, regression, and classification, thanks to the complete integration of these libraries.

## 2.2 Overview of the Dataset

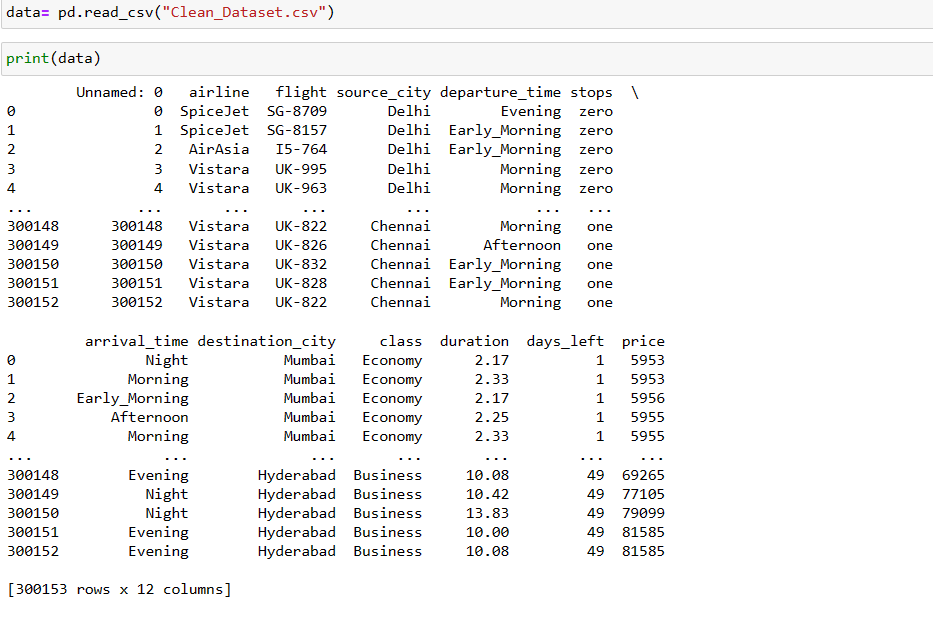


Figure 2.2 Data Overview

The data is provided by Kaggle and the dataset includes over 30.000 flights across various features.

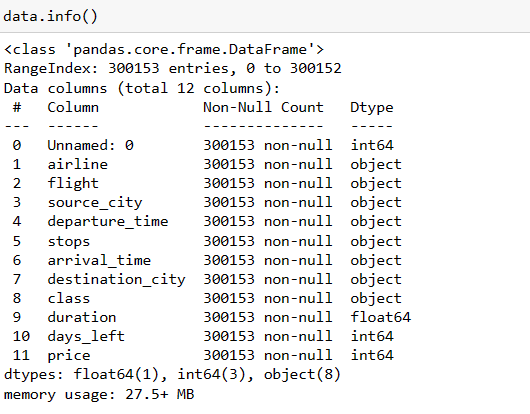


Figure 2.3 Features and Data Types

When review features and their data types except from duration , days left and price other all labels are classififed as an object.Other three variables are classified as float and integer respectively as seems on the figure at the above.

However if you look at the column there is a critical point which is related to ‘ Unnamed = 0 ’ label because it doesn’t have any menaingful situation as a result of this situation we have to remove it from the dataset by following code ;

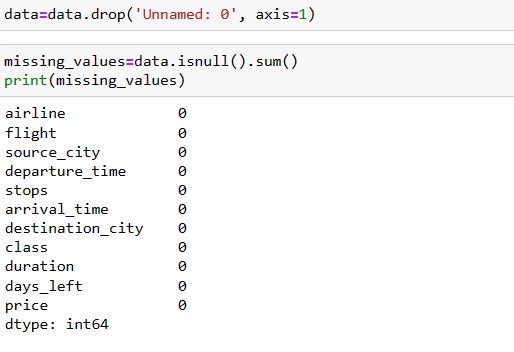


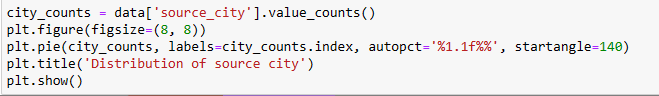
Figure 2.4 Review of Missing Values

After drop ‘ Unnamed = 0 ‘ column we need to evaluete is there any existence of missing values by using isnull().sum() structure luckily the datase doesn’t demonstrate any missing towards to the labels also from missing value lists where confirms that removing ‘ Unnamed = 0 ‘ column is completed succesfully.

# CHAPTER TWO Exploratory Data Analysis

On our dataset which has various labels so revealing visualization across these labels is crucial to understand inferences then the report will continue with this stage.

## 2.1 Distribution of Source City



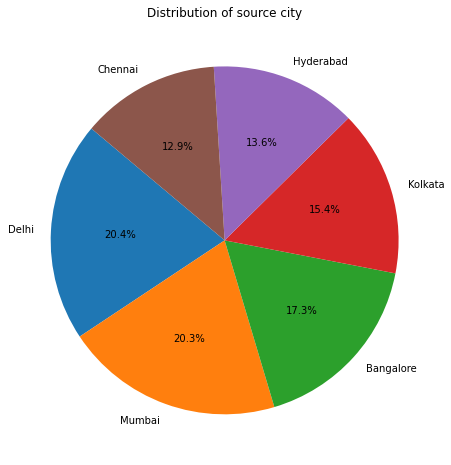


Figure 2.1 Distribution of Source City

The dataset entails 6 cities and they are distributed like Delhi has highest proportion then Mumbai is second this continues Bangalore , Kolkata , Hyderabad and Chennai respectively.To visualize this information pie chart is utilized.

## 2.2 Distribution of Destination City

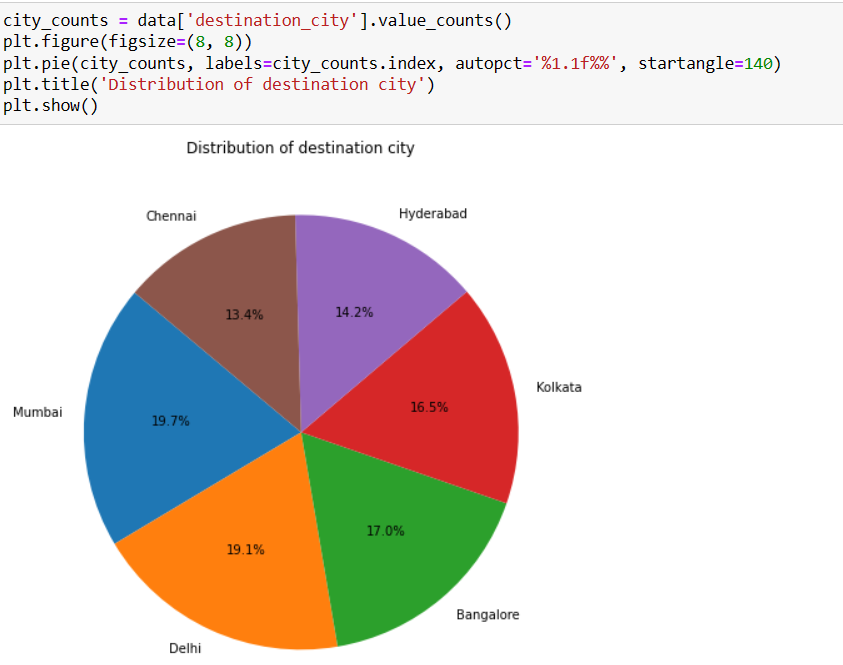


Figure 2.2 Distribution of Destination City

When destination city is visualized ranking of cities by percentage is that Mumbai , Delhi , Bangalore , Kolkata , Hyderabad and Chennai which means that highest demand belongs to Mumbai and Delhi cities.

## 2.3 Distribution of Class

If we look at class information of passengers the pie chart indicates majority of tickets were purchased with economy – class as mentioned the figure at the below.

At the same time this proportion might be useful to explore socio – economic status of passengers.

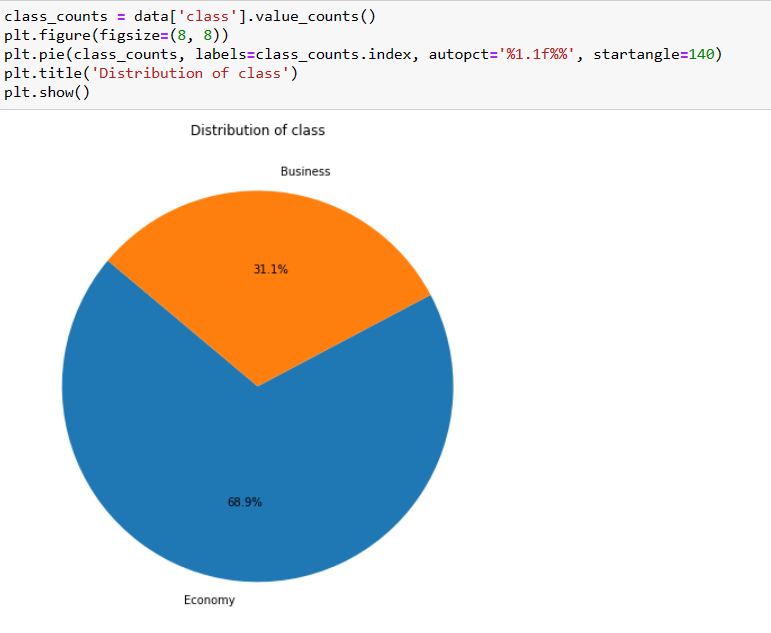


Figure 2.3 Distribution of Class

## 2.3 Distribution of Airlines

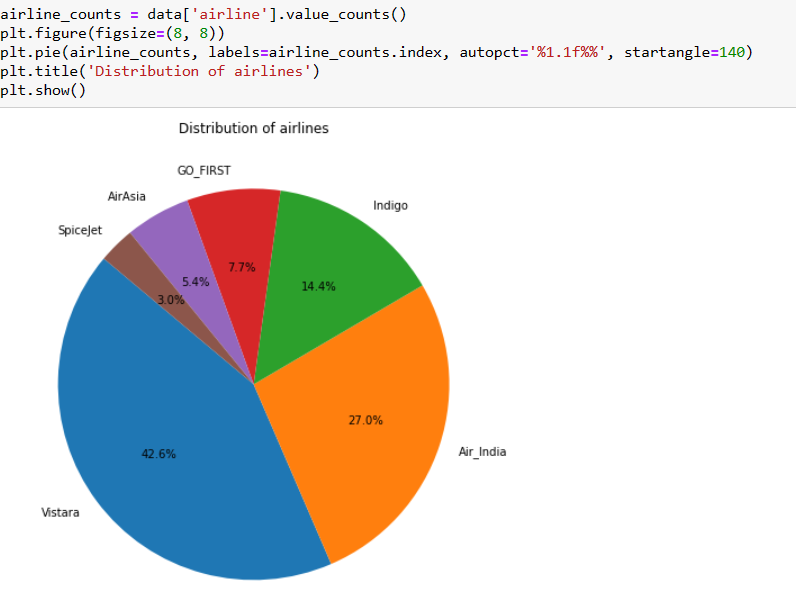


Figure 2.4 Distribution of Airlines

There are six different airlines commpanies and this dsitrubiton flight operations rate by airlines companies.Moreover highest number of flights belongs to Vistara Airlines by 42.6 percent then it is followed by Air India by 27 percent.Moreover this ranking continue with Indigo , Go First , Air Asia , SpiceJet by percentage of 14.4 , 7.7 , 5.4 , 3.0 respectively.

## 2.3 Distribution of Stops

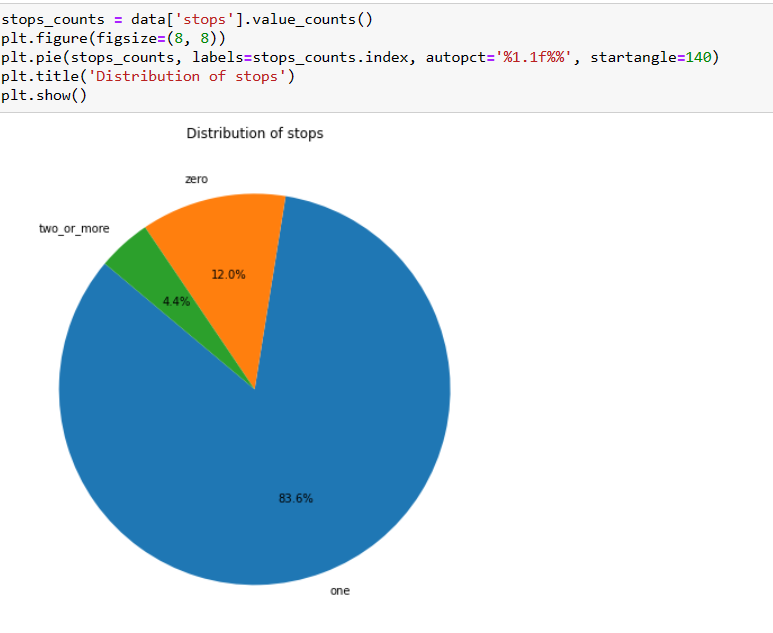


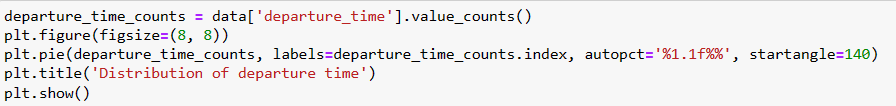
Figure 2.5 Distribution of Number of Stop

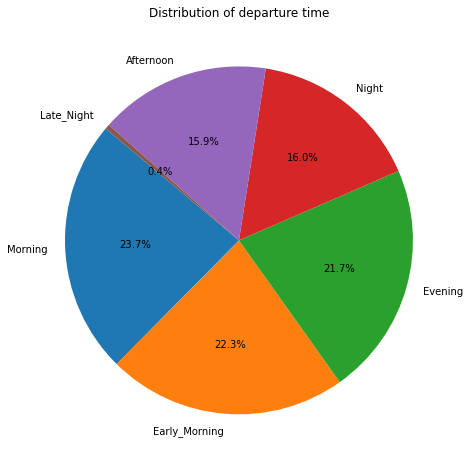
Distribution of stops declares that number of stops between source city and destination city.Firstly , majority flights include at least one stop by 83.6 percent only 12 percent represents no-stops and 4.4 percent of flights contain 2 stops.

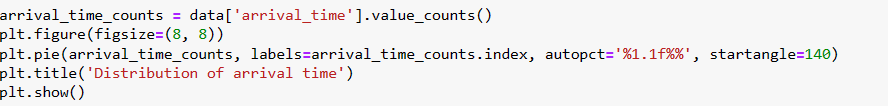
## 2.4 Distribution Departure – Arrival Time

For departure time most of flight operations are realized during early morning , morning and evening by totally 67.7 percent.Remain travels occurs night , afternoon and late night.

For arrival time planes landing to morning , evening and night part of a day by 77.5 percent others are at early morning , afternoon , and late night.







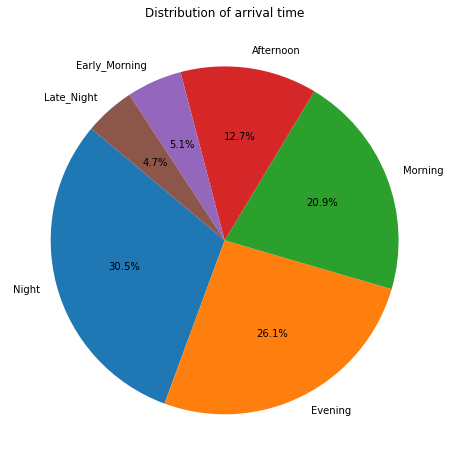


Figure 2.6 Distribution of Departure – Arrival Time

# CHAPTER THREE Model Selection and Implementation

## 3.1 Model Selection

Any datasset for projects might have either labels or not and this property is very critical to select suitable model.

If the dataset has labels on it, like telling you what something is, then there are two kinds of things you can do with it. One is called regression, where you try to figure out what a number might be. The other is called classification, where you try to figure out what category something belongs to.

Clustering is like a game where you have a bunch of toys, but they don't have any labels or categories. You want to figure out if there are any groups or patterns among the toys without someone telling you what those groups are.

According to this conditions we need select either classification or regression method since the dataset has labels.

## 3.2 Implementation

### 3.2.1 Encoding and Scaling

Target encoding, or likelihood encoding, includes supplanting each category of a categorical include with the mean or middle of the target variable for that category. This strategy captures the relationship more absolutely than one-hot encoding, particularly for categorical highlights with various categories. Moreover, target encoding can decrease the dimensionality of the include space.In encoding step which must address to target variable in this case it is price because aim of the project is predicting flight ticket prices.

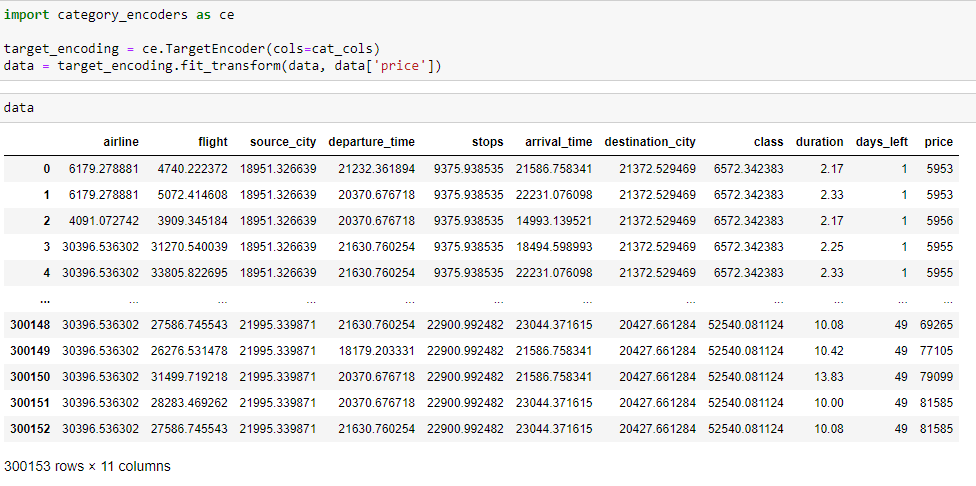


Figure 3.1 Target Encoding of The Dataset

Scaling the dataset.Scaling could be a preprocessing step in machine learning that points to standardize the extend or scale of the input highlights. The objective of scaling is to guarantee that each include features a comparative scale or extend, which can offer assistance a few machine learning models to merge speedier and progress their execution.After this point as a name scaled\_data will be used instead of data.

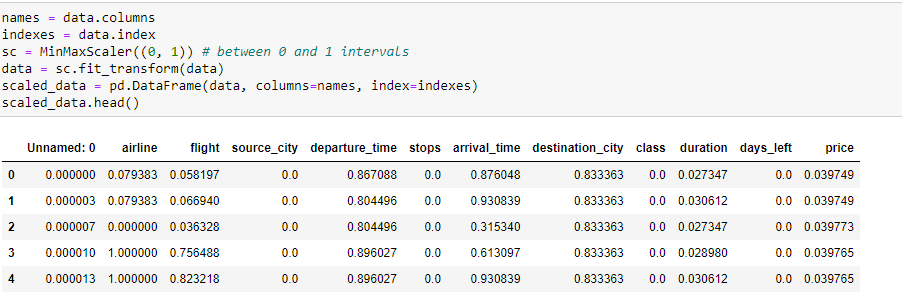


Figure 3.2 Scaling of The Dataset

### 3.2.2 Correlation Matrix

When we look at Figure 3.3 about correlation matrix which indicates there are strong correlaitons between flight (flight code) - airline and class-price. Of course flight-airline correlation isn’t important for this project but class-price is very critical for us.This case can be observed at the below.

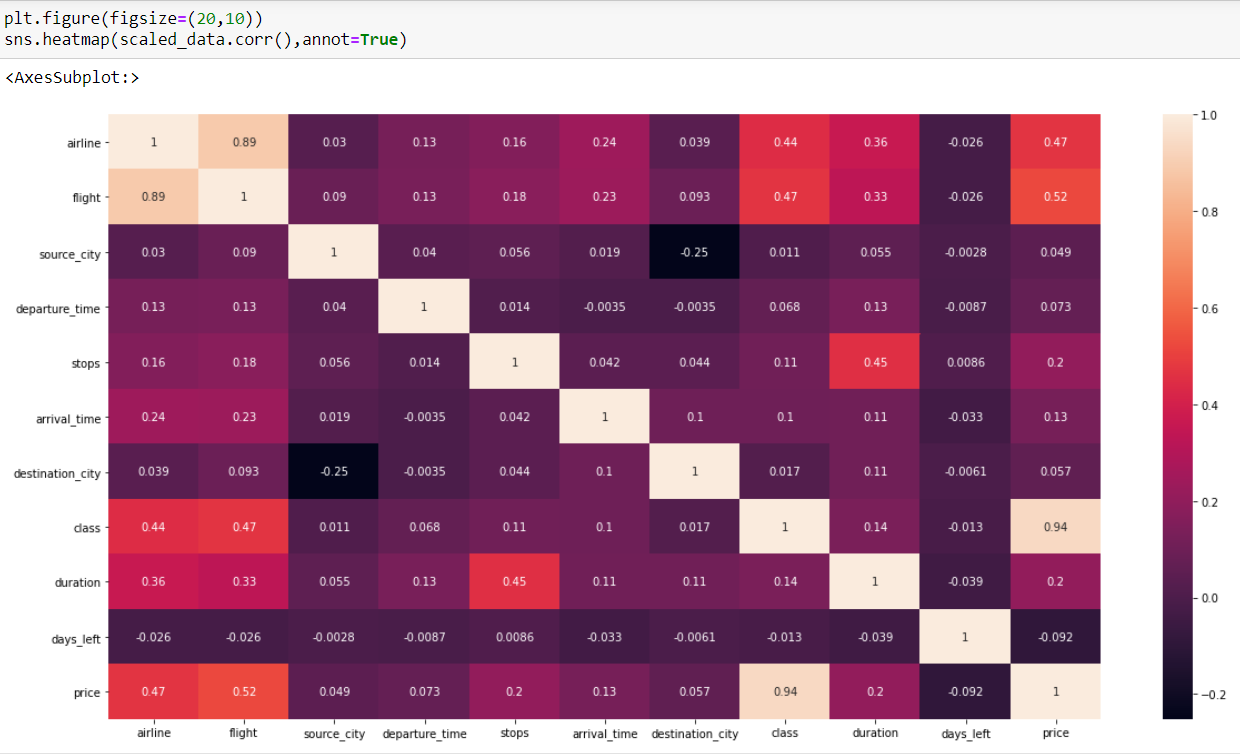


Figure 3.3 Correlation Matrix Between Features

### 3.2.3 Splitting Data Into Training and Test Part

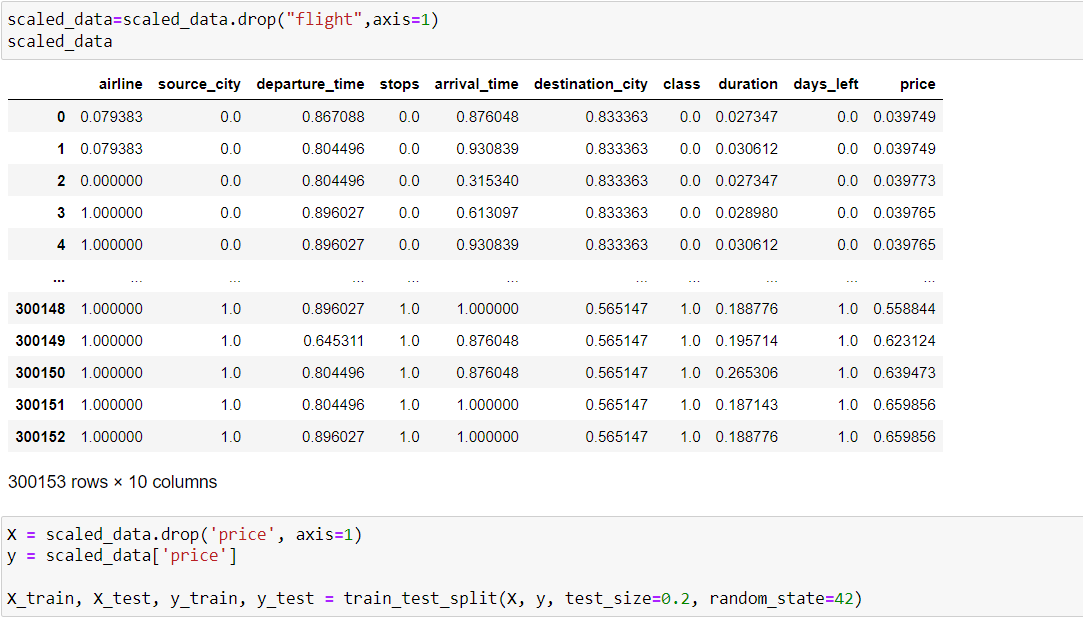


Figure 3.4 Splitting the Data Into Training and Test

Before splitting the data into training and test flight feature was removed from the dataset since it isn’t necessary as well as no meanningful impact.After that the data has been splitted by 20 percent as test size then remains are as training part.Furthermore our target variable is price label so we removed it from the data then it was transferred into another place to achieve prediction.Except from price other variables have been utilized to predict price feature by using different alogrithms.

### 3.2.4 Linear Regression

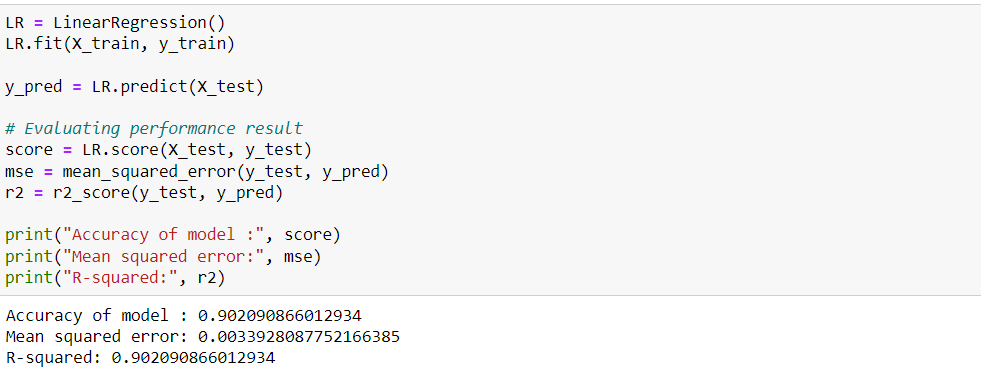


Figure 3.5 Linear Regression

Linear regression is a way to figure out how things are connected by using numbers. It helps us understand how one thing can change when another thing changes.

### 3.2.5 Decision Tree Regressor



Figure 3.6 Decision Tree Regressor

A Decision Tree Regressor is like a smart computer program that can help predict a number. It looks at different characteristics and splits the data into smaller groups based on the most important things it finds. Then, it creates a set of rules to help make predictions.

### 3.2.6 Random Forest Regression

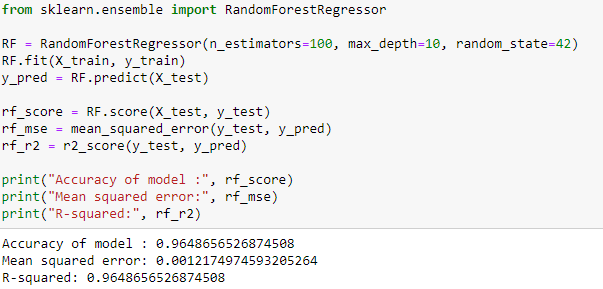


Figure 3.7 Random Forest Regressor

A Random Forest Regressor is a special algorithm that uses a bunch of decision trees to make predictions. It takes a lot of different factors into account and then combines the predictions from all the trees to give a more accurate and reliable answer than just using one tree.

### 3.2.7 K-Nearest Neighbor

K-Nearest Neighbors (KNN) is a special way of predicting things based on information we already have. It looks at the closest pieces of information we already know and uses them to make a guess about something new. It also uses a special way of measuring how similar or close things are to each other.Also K value was adjusted to 5.



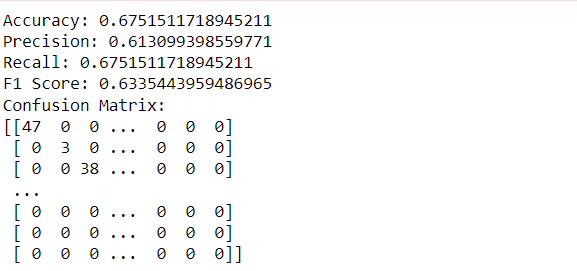


Figure 3.8 K-Nearest Neighbor

In the next chapter perofmrance scores will be evaulated.

# CHAPTER FOUR Model Evaluation

|  |  |  |  |
| --- | --- | --- | --- |
| Benchmarks / Models | Linear Regression | Decision Tree | Random Forest |
| Accuracy | 0.9020 | 0.9429 | 0.9648 |
| Mean Squared Error | 0.0033 | 0.0019 | 0.0012 |
| R-Squared | 0.9020 | 0.9429 | 0.9648 |

|  |  |
| --- | --- |
| Benchmarks / Model | K-Nearest Neighbor |
| Accuracy | 0.6751 |
| Precision | 0.6130 |
| Recall | 0.6751 |
| F-1 Score | 0.6335 |
| Confusion Matrix | [[47 0 0 ... 0 0 0]  [ 0 3 0 ... 0 0 0]  [ 0 0 38 ... 0 0 0]  ...  [ 0 0 0 ... 0 0 0]  [ 0 0 0 ... 0 0 0]  [ 0 0 0 ... 0 0 0]] |

Table 4.1 Comparision of Performance Results

Accuracy could be a degree of how regularly a show makes redress forecasts, calculated as the ratio of accurately anticipated occasions to the overall occasions in a dataset. It's a common metric for overall model performance, particularly when classes are adjusted.

Mean Squared Blunder (MSE) could be a metric utilized in relapse issues to degree the normal squared distinction between anticipated and genuine values. It penalizes larger errors more intensely and is calculated as the normal of the squared differences.

R-squared could be a measure in relapse examination that demonstrates the extent of the fluctuation within the subordinate variable that's unsurprising from the autonomous factors. A better R-squared esteem proposes distant better;a much better;a higher;a stronger;an improved" an improved fit of the demonstrate to the information.

Precision could be a metric that measures the exactness of the positive forecasts made by a show. It is calculated as the proportion of genuine positives to the whole of genuine positives and wrong positives, giving experiences into the model's capacity to dodge wrong positives.

Recall could be a metric that measures the capacity of a demonstrate to capture all the significant occasions of a course. It is calculated as the ratio of genuine positives to the entirety of genuine positives and wrong negatives, giving experiences into the model's capacity to avoid wrong negatives.

Confusion Matrix could be a table that summarizes the execution of a classification algorithm. It appears the checks of genuine positive, genuine negative, untrue positive, and untrue negative forecasts, giving a point by point understanding of the model's qualities and shortcomings.

F-1 Score is the consonant cruel of exactness and review, giving a balance between the two measurements. It is valuable when there is an uneven course distribution and is calculated as 2 times the item of accuracy and recall separated by their sum.

When we review overall situation regarding to benchmark scores across different algorithms, among them highest accuracy belongs to random forest method because random forest capacity to handle non-linearity, its gathering learning approach, highlight significance investigation, and strength to exceptions make it a solid candidate for flight cost forecast compared to linear regression and decision tree.This positive case also valid for other benchmark types.

On the other hand K-Nearest Neighbor demonstrates poor performance relatively to other algorithms since KNN's restrictions in high-dimensional spaces, affectability to exceptions, versatility issues, need of interpretability, reliance on include scaling, challenges in dealing with meager datasets, and challenges with categorical highlights make it less reasonable for flight cost expectation compared to other machine learning methods that address these issues more viably.Also as a classification method for linear regression same problems can be said.

# CHAPTER FIVE Conclusion and Recommendations

The predictive analytics arrangement created for estimating airfare costs has illustrated commendable viability in tending to the challenges confronted by the travel industry, especially amid the riotous period of the COVID-19 widespread. The models, counting Linear Regression, Decision Tree Regressor, Random Forest and K-Nearest Neighbor, were carefully chosen and actualized to foresee ticket costs precisely.

The discoveries from the show assessment stage uncover that the Random Forest show shown the most noteworthy exactness, cruel squared blunder, and R-squared values among the tried calculations. This means its strong capability to handle the complex non-linear connections inalienable in airfare estimating. The Decision Tree Regressor too performed well, appearing a high degree of prescient exactness. In any case, the K-Nearest Neighbor show slacked behind, demonstrating impediments in its reasonableness for this particular forecast assignment.

Suggestions for tending to the recognized commerce issue incorporate a vital center on leveraging the qualities of the Random Forest Regression demonstrate. Its capacity to handle non-linearity, decipher include significance, and adjust to changing showcase conditions positions it as a capable device for optimizing income within the flying industry. Persistent observing and fine-tuning of the demonstrate parameters based on advancing travel designs and financial conditions will encourage upgrade its execution.

Furthermore, there's a got to investigate advance include designing and refinement to possibly move forward the prescient capabilities of the models. Understanding and consolidating extra components such as geopolitical occasions, financial pointers, and worldwide wellbeing emergencies might contribute to more exact forecasts. Collaboration with industry specialists to pick up domain-specific bits of knowledge and remaining overhauled on developing patterns will be urgent in improving the models' flexibility to the energetic nature of the travel segment.

In conclusion , The predictive analytics arrangement not as it were gives a apparatus for aircrafts and travel organizations to optimize estimating methodologies but moreover underscores the significance of ceaseless advancement and adjustment to outside components. By grasping a data-driven approach and actualizing the suggested techniques, businesses can explore the challenges of the travel industry, maximize income, and provide a more fulfilling encounter to clients.

# CHAPTER SIX Project Reflection

For this analysis the dataset is provided by Kaggle then required pre-process and visualizaiton were done at the model stage there were some problems because the dataset as its nature which isn’t appropratie for clustering as well as classification although classification seems right choice across labeled dataset but this dataset have multi dimensional shape for feature.As a result of this situation even K-Nearest Neighbor was tried still the accuracy is lower than regression methods.Also logistic regression is difficult to apply due to long time process as well as encoding problems.

There might be some improvement for instance if classification methods will be applied by incremental learning apporach might be useful since incremental learning, a self-adaptive algorithm uses the previously learned model information, then learns and accommodates new information from the newly arrived data providing a new model, which avoids the retraining. The incrementally learned knowledge helps to classify the unstructured data (Madhusudhanan ,2018). Another option might be dimensional reduction which aims to transform a dataset into lower level dimension this situation decreases reuqired computational power especially these two approaches are critical for logistic regression.

# BIBLIOGRAPHY

Bathwal ,S. (2022) Flight Price Prediction [online] available from < https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data > [ 30 January 2024]

Ahmad ,N. , Li ,S. , Hdia ,M. , Belav ,J. , Hussain ,W. (2023) ‘ Assessing the COVID-19 pandemic impact on tourism arrivals: The role of innovation to reshape the future work for sustainable development’. Journal of Innovation & Knowledge (8) 1-3

Madhusudhanan ,S. , Jaganathan ,S. , Jayashree ,L. (2018) ‘ Incremental Learning for Classification of Unstructured Data Using Extreme Learning Machine’. MDPI Journal (11) 1-2

Petinrin ,O. , Saeed ,F. , Salim ,N. , Toseef ,M. , Liu ,Z. , Muyide ,I. (2023) ‘ Dimension Reduction and Classifier-Based Feature Selection for Oversampled Gene Expression Data and Cancer Classification ’ MDPI Journal (11) 5-6

# APPENDIX (if necessary)

KNN : K-Nearest Neighbor