****

**Project Name:**

*Flight price Predictions*

**Author:**

*Laxmikant S. Walzade*

* **Problem Definition.**

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on - 1. Time of purchase patterns (making sure last-minute purchases are expensive) 2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases) So, you have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights. Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest

available ticket on a given flight gets more and less expensive over time. This usually happens as

an attempt to maximize revenue based on -

1. Time of purchase patterns (making sure last-minute purchases are expensive)

2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order

to reduce sales and hold back inventory for those expensive last-minute expensive

purchases)

So, you have to work on a project where you collect data of flight fares with other features and

work to make a model to predict fares of flights.

Below is the snapshot of a dataset: -



In this Dataset Income is the Label. We are building a model to Forecast the target variable price. There are complete 10 features which are: -

* airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price.

**Data Analysis.**

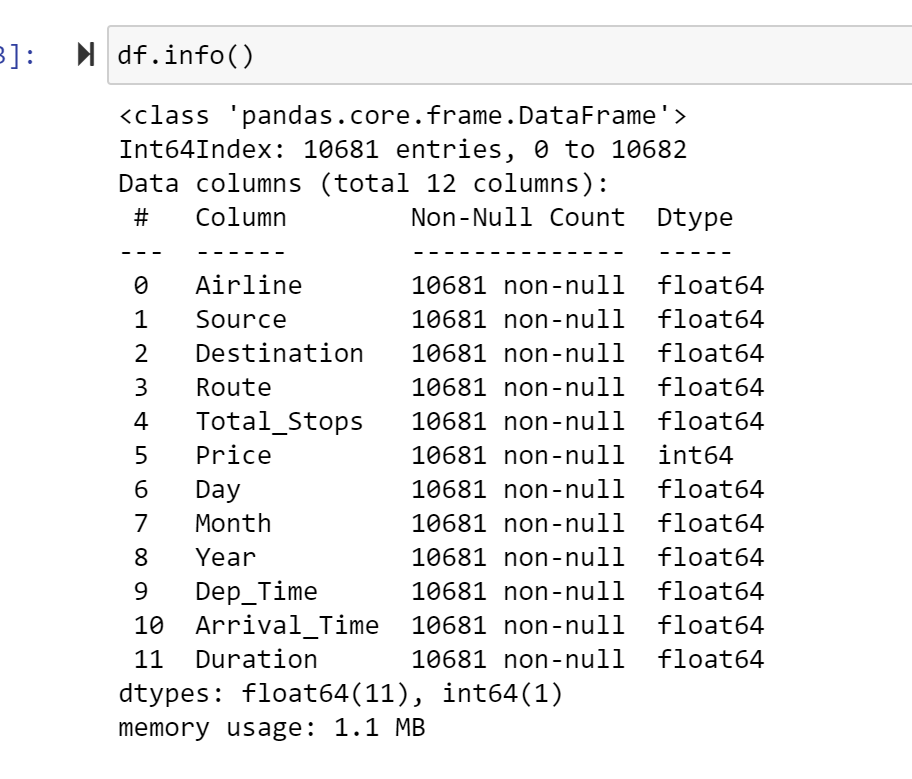
* We have total 10 columns including the label i.e Income column.
* Airline name-flight name
* date of journey-at what day the passenger travelled
* source- from where the passenger is go to board
* destination- destination of passenger
* route-how the flight is going to route
* departure time-take of time
* arrival time-landing time
* duration-time taken to land on destination
* total stops-how many stops had the flight taken
* price- cost of passenger

Each Features have some influence in predicting the price of flight . It is necessary to know which features have greater impact on price and which does not have co-relation with label.

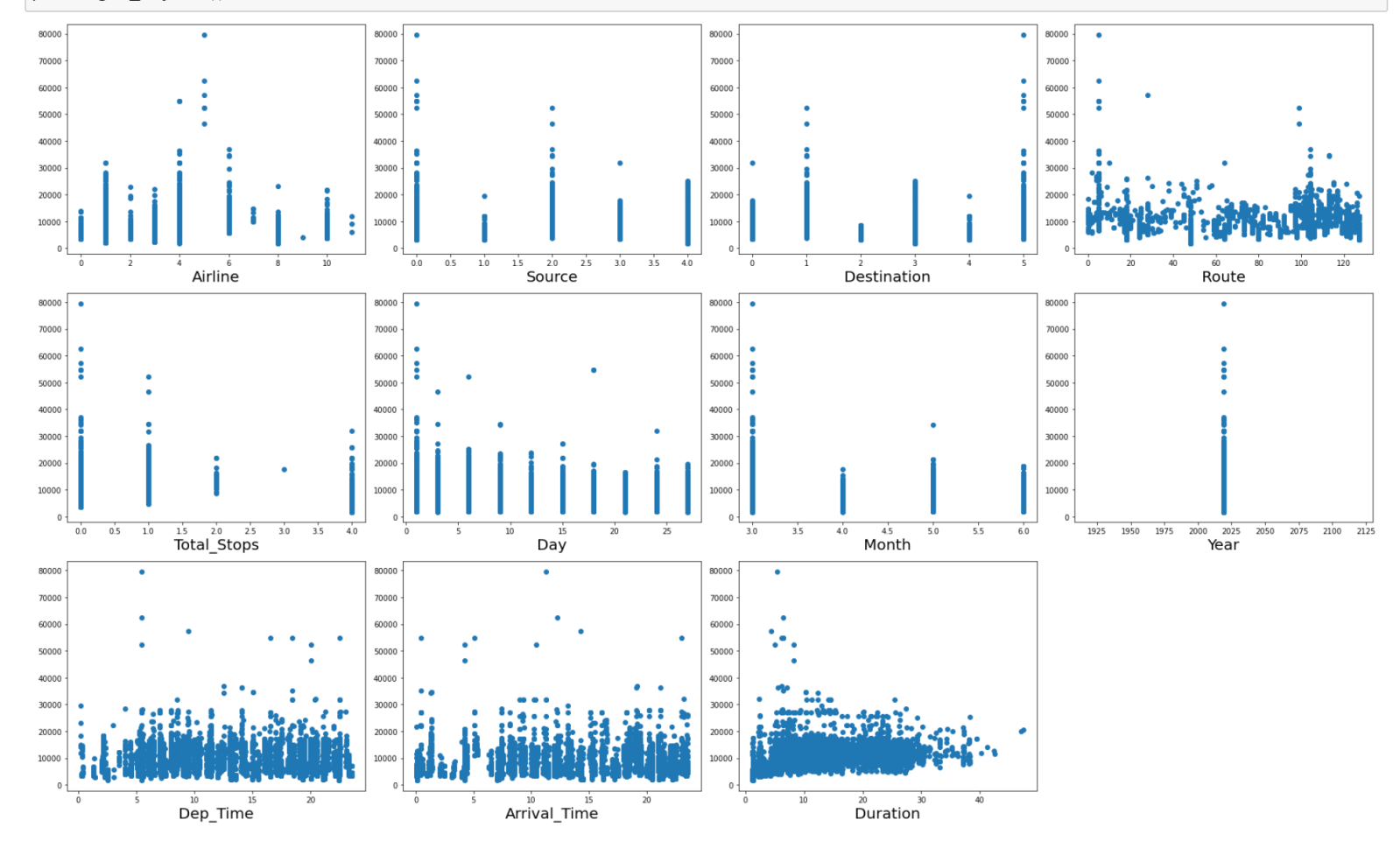
* **Pre-Processing Steps**

1. Identifying sources of the data
2. Analysing the information
3. Cleaning and handling the information
4. Selecting the most significant elements
5. Writing down findings and observations
6. Using various models to train the data
7. Selecting the best-fitted model for predictions
8. Predicting results for test information

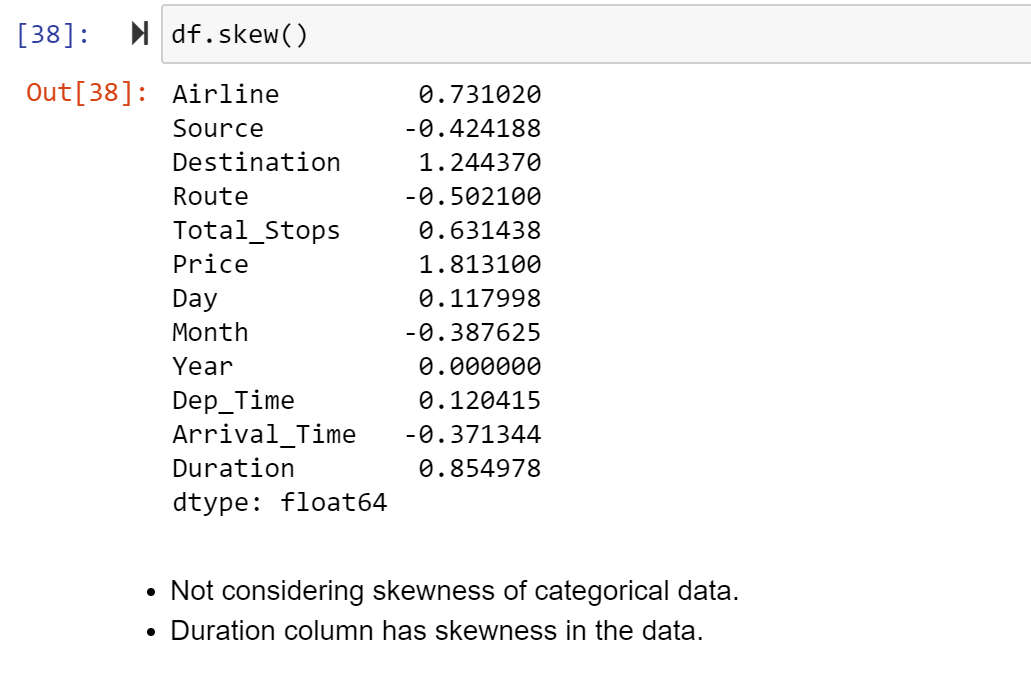
* **Pre-Processing Pipeline.**
* First let’s check the data type of the dataset.

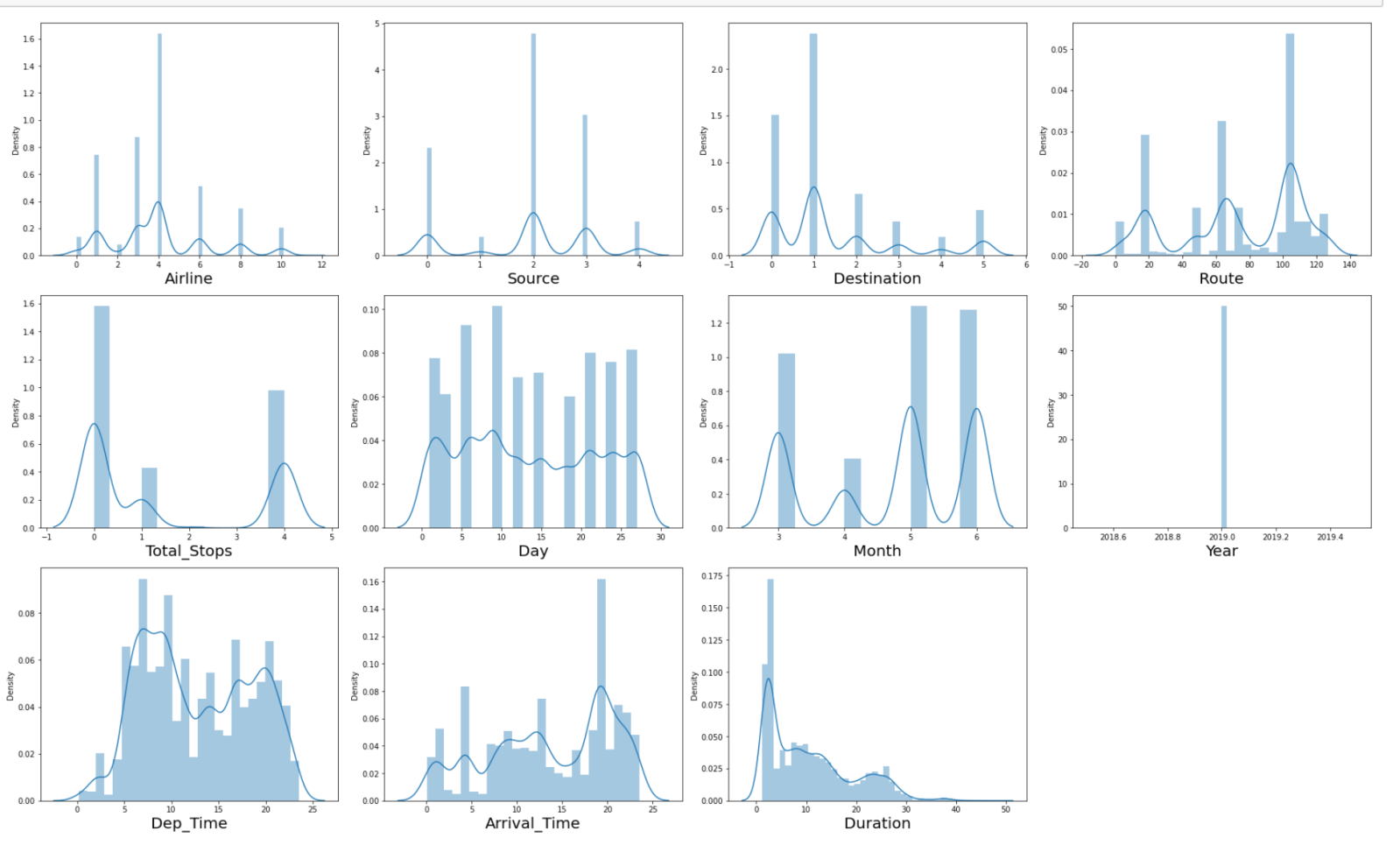


All are float values only price is an integer.

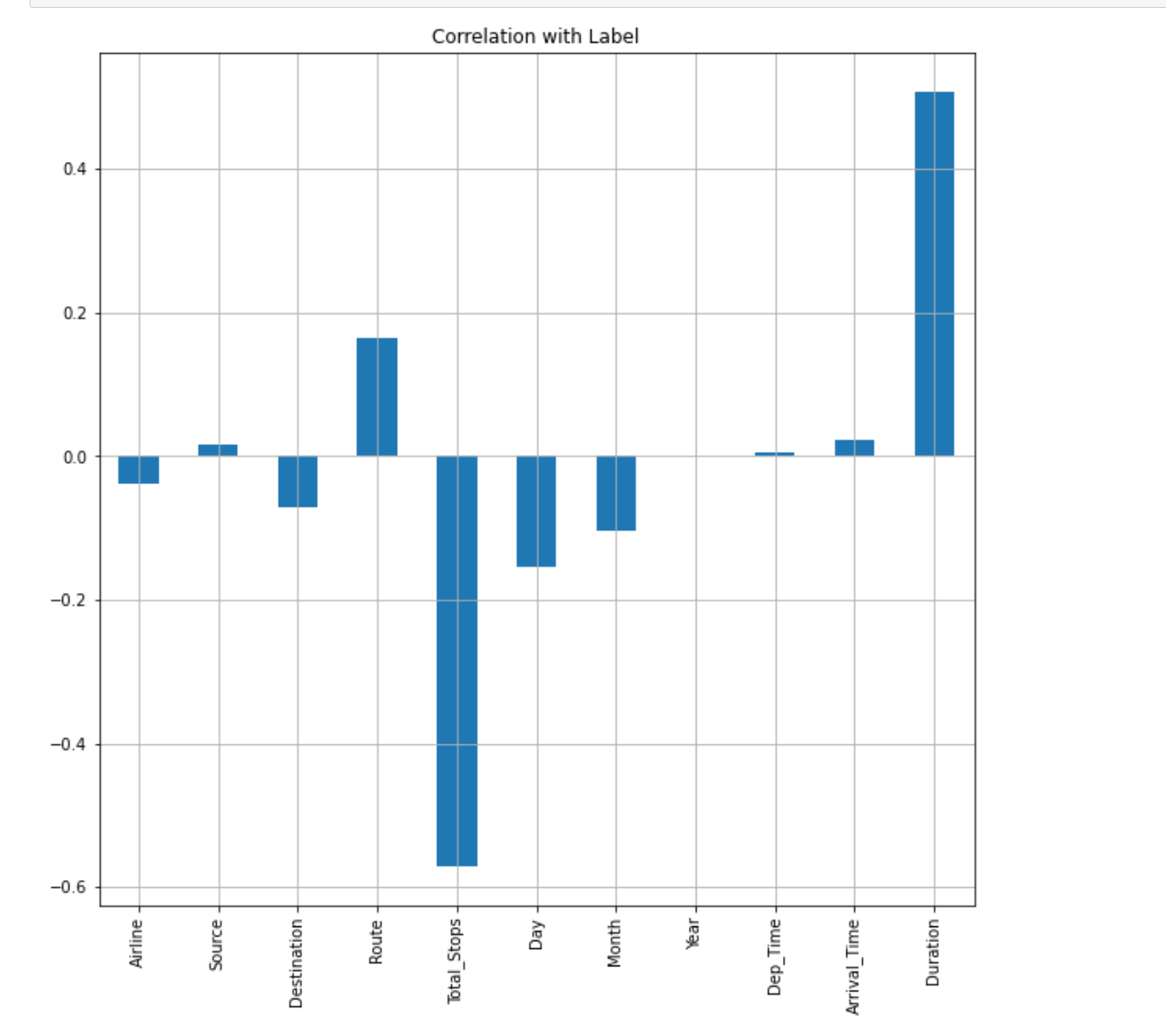


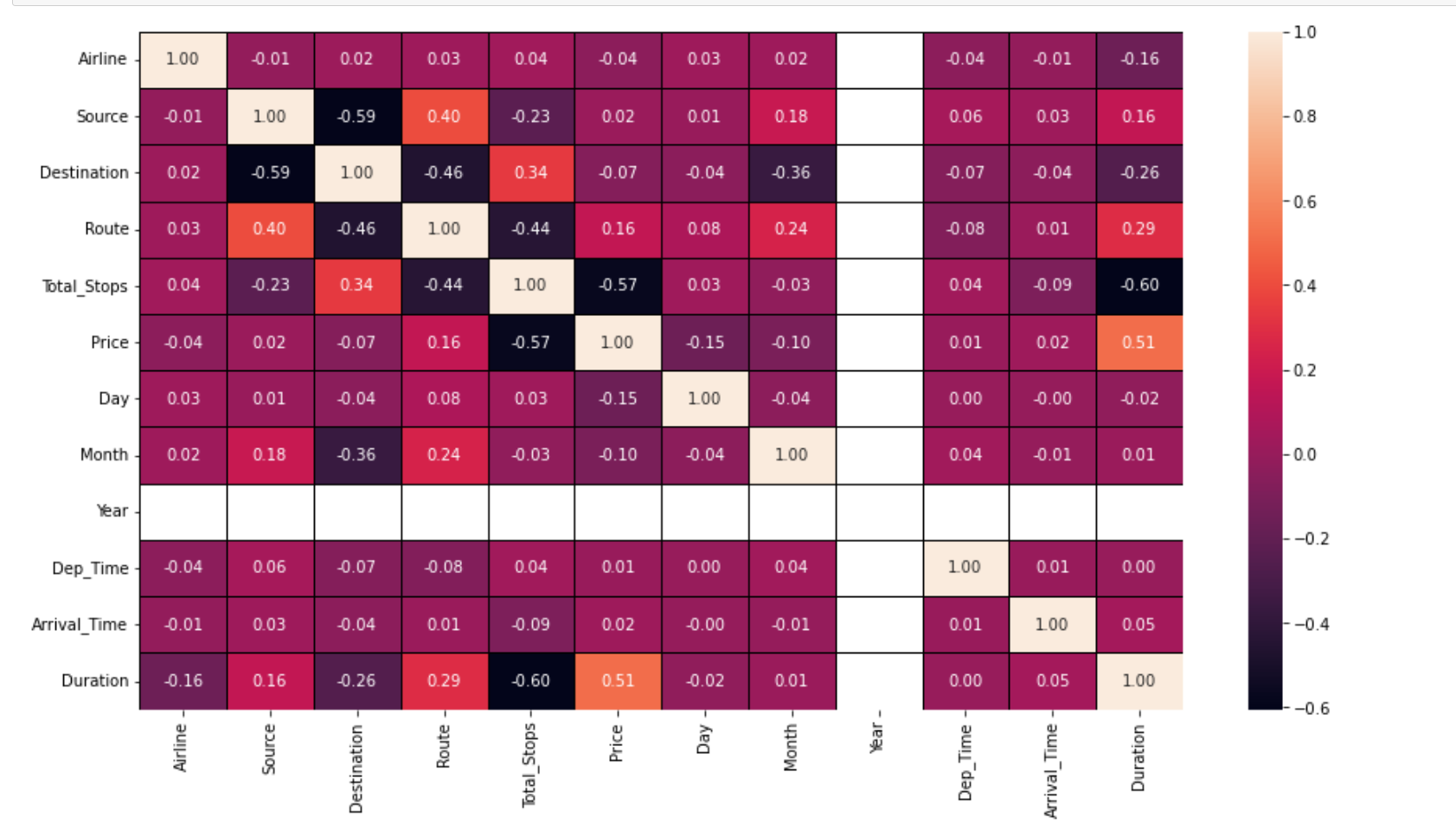
Airline, Source, Destination, Total\_stops, Day, Month has a good co-relation with Label.





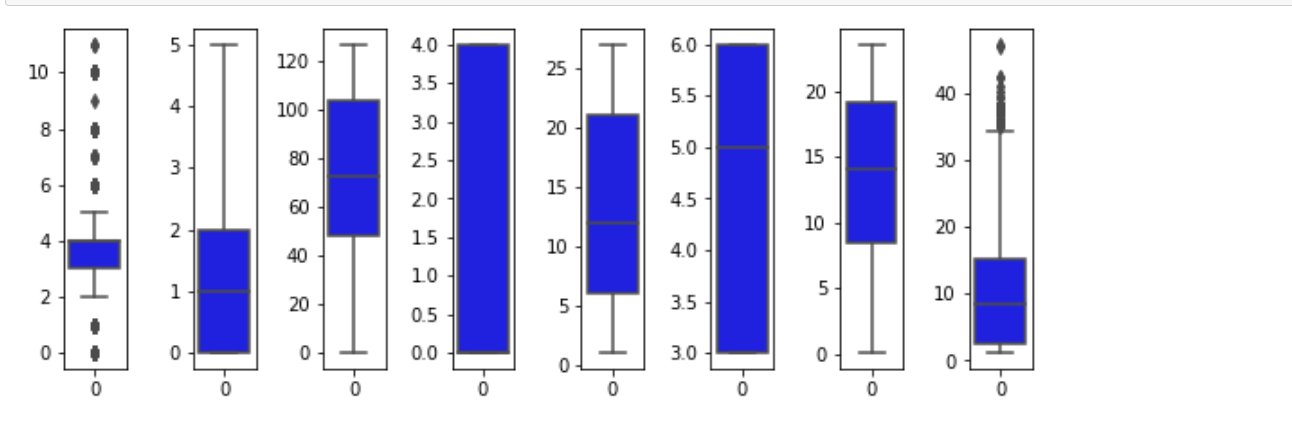
. Duration column has skewness and may have outliers in the data.

****

****

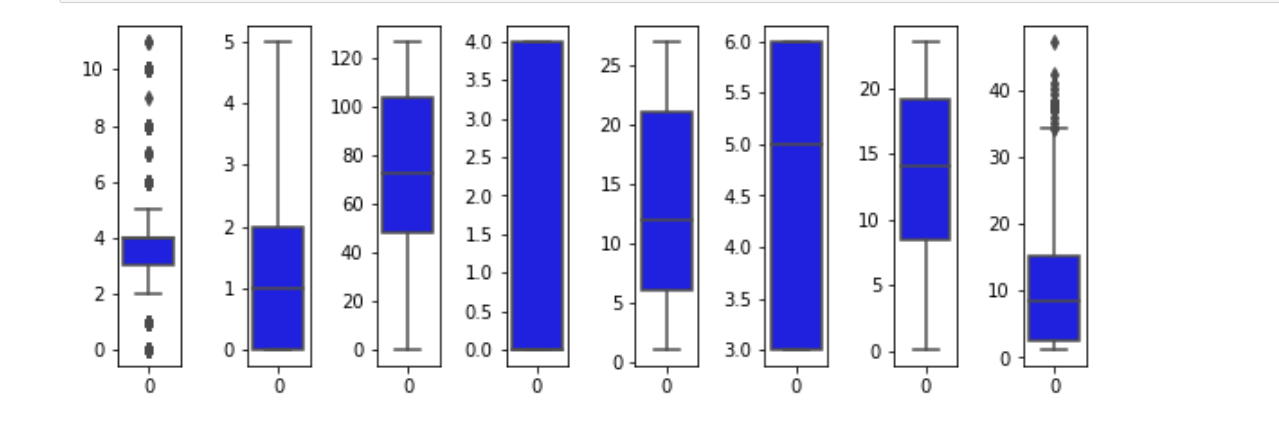
**'Year', 'Source' and 'Dep\_Time' since it has very low or no corelation with label**

**Multicollinearity problem does not exist in this database**

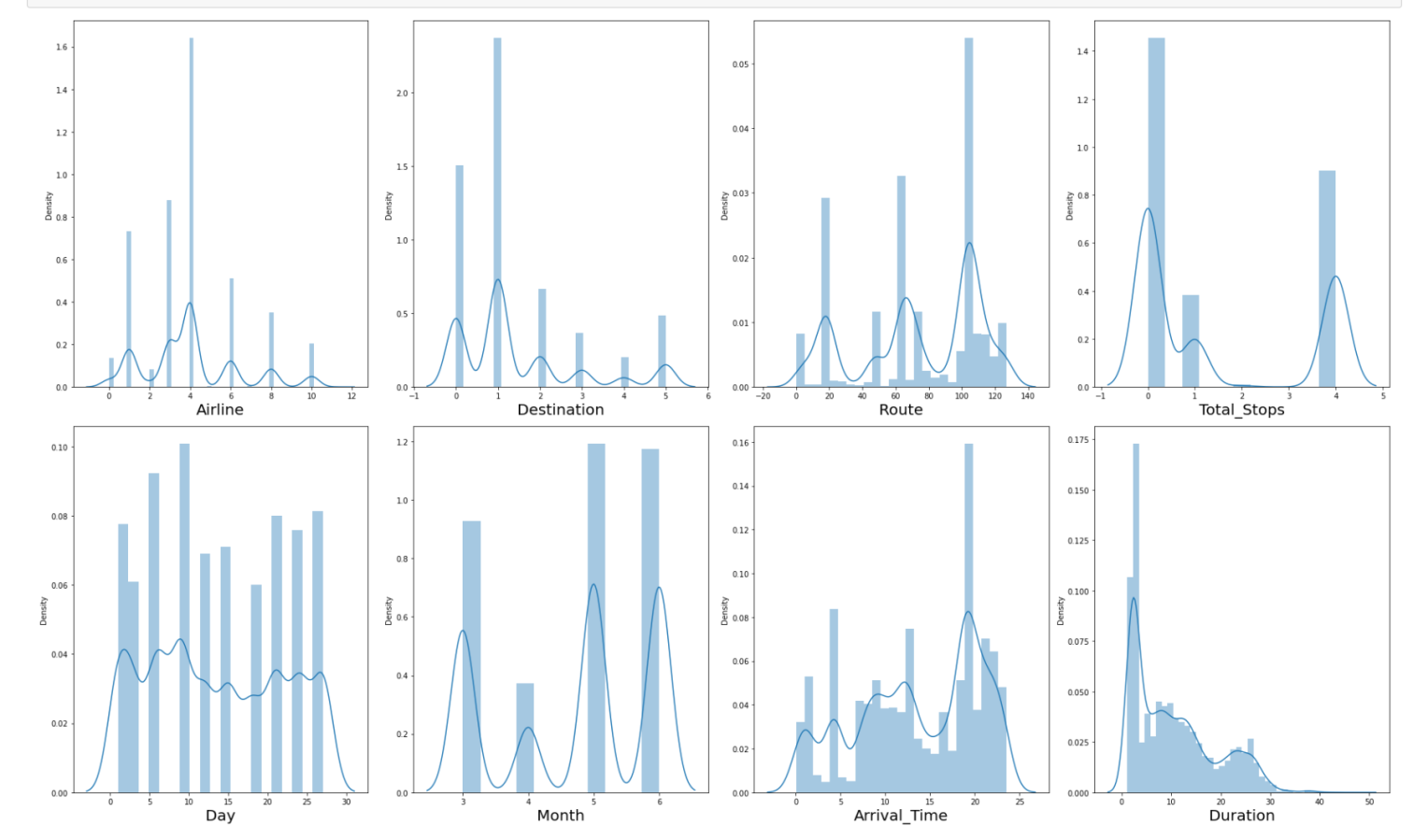
****

* **Outliers are present in Airline and Duration.**
* **Airline is not a continuous data hence can not remove outliers from Airline column**

**Removed Outliers that are not allowed above and below the specified limit**

****

**There is a skewness present in the data hence using PowerTransformer to remove it**

****

Removed skewness from the data

* **EDA Concluding Remark.**

1. The information was not organized and coordinated and subsequently cleaned the information utilizing different information cleaning and pre-handling techniques.
2. There are numerous anomalies present in the information consequently eliminating exceptions
3. There was a skewness in the information thus have eliminated the skewness from the information.
4. There was an irregularity in the information thus have utilized SMOTE strategy to balance the information.
5. Scaled the data utilizing Standard Scalar to make the information normalized to fabricate a model.

* **Hardware and Software Requirements and Tools Used**

1. Libraries and packages used

* import numpy as np - For Numpy work
* import pandas as pd - To work on DataFrame
* import seaborn as sns - Plotting Graphs
* import matplotlib.pyplot as plt - Plotting Graphs
* import pickle – To save the Model
* from sklearn.preprocessing import StandardScaler (To scale the train data), OrdinalEncoder(To encode object data to Integer), PowerTransformer (To remove skewness from dataset)
* from statsmodels.stats.outliers\_influence import variance\_inflation\_factor
* enc = OrdinalEncoder() = Assigned OrdinalEncoder to variable
* from sklearn.model\_selection import train\_test\_split, GridSearchCV, cross\_val\_score,(To split the data into train and test, Search the best parameters, to calculate cross validation score)
* from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix, roc\_curve, roc\_auc\_score - To calculate and analyse model metrics.
* from sklearn import metrics

Models which are used

* from sklearn.ensemble import RandomForestClassifier
* from sklearn.linear\_model import LogisticRegression
* from sklearn.ensemble import GradientBoostingClassifier
* from sklearn.tree import DecisionTreeClassifier
* from sklearn.neighbors import KNeighborsClassifier
* from sklearn.svm import SVC
* import warnings
* warnings.filterwarnings('ignore') - To ignore unwanted Warnings

1. Hardware used – 11th Gen Intel(R) Core (TM) i3-1115G4 @ 3.00GHz 3.00 GHz with 8.00 GB RAM and Windows 11
2. Software used – Anaconda and Jupyter Notebook to build the model.

* **Building Machine Learning Models.**

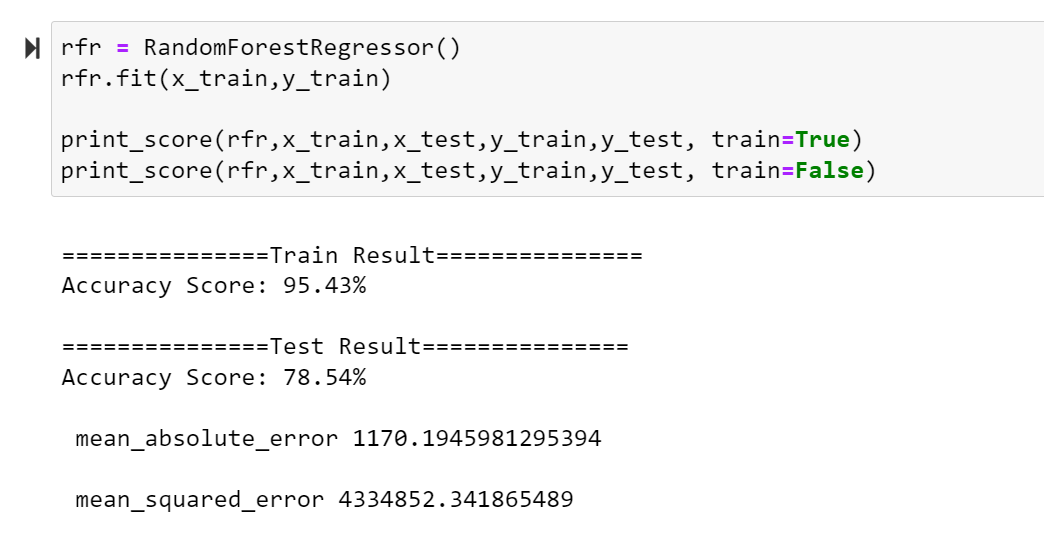
I have built 6 machine learning models to predict the label. Below are the machine learning models which are been used.

1. LogisticsRegression
2. RandomForestClassifier
3. DecisionTreeClassifier
4. GradientBoostingClassifier
5. Support Vector Classifier
6. KNeighborsClassifier

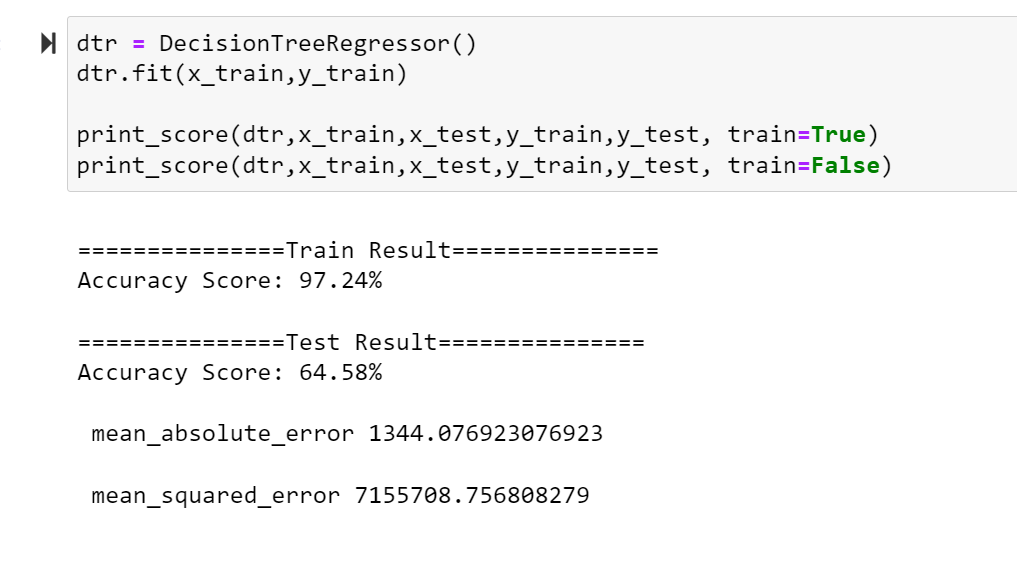
**1.LogisticsRegression**

****

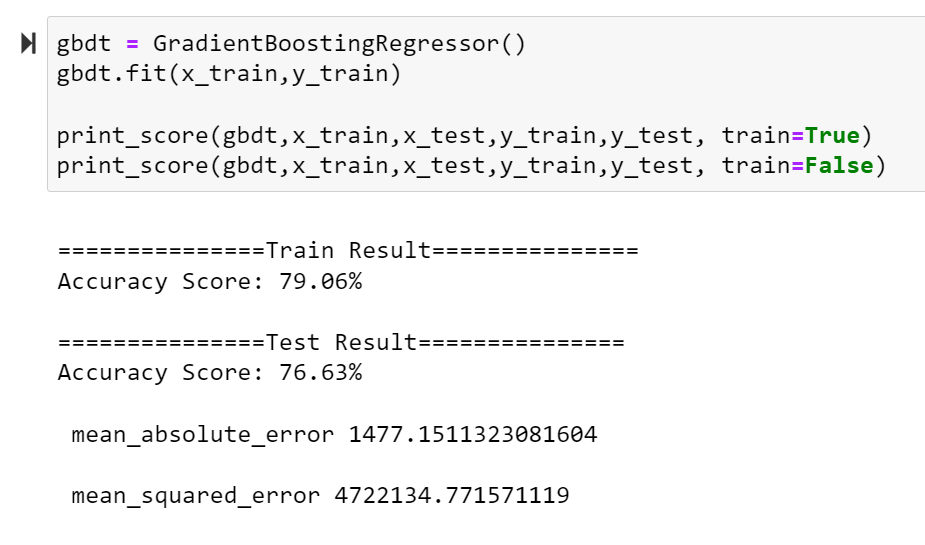
**2.RandomForestClassifier**

****

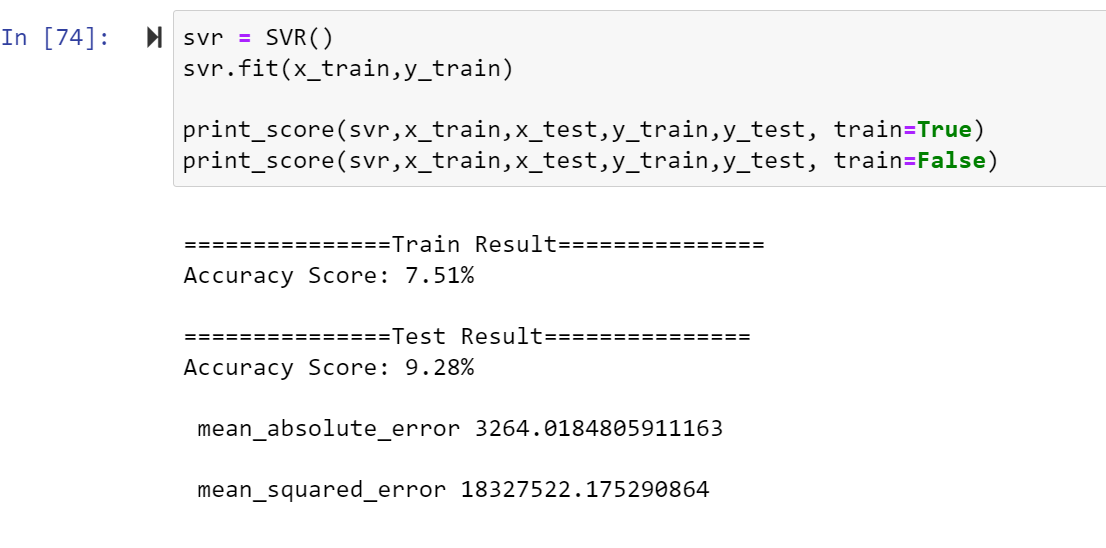
**3.DecisionTreeClassifier**

****

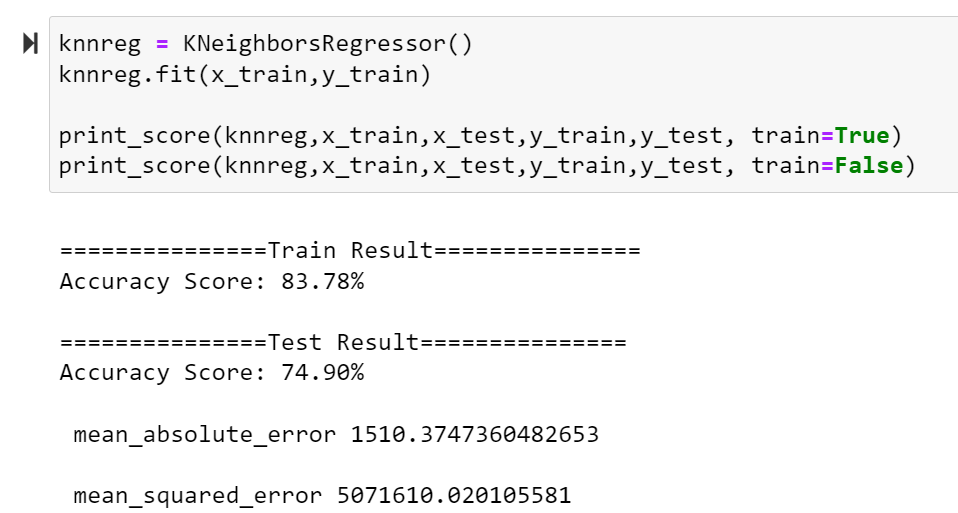
**4.GradientBoostingClassifier**

****

**5.Support Vector Classifier**

****

**6.KNeighborsClassifier**

****

**Findings**

* LinearRegression train accuracy score 43.15% and test accuracy score 44.15%
* Support Vector Regression train accuracy score 7.51% and test accuracy score 9.28%
* DecisionTreeRegressor train accuracy score 97.24% and test accuracy score 64.58%
* AdaBoostRegressor train accuracy score 39.60% and test accuracy score 32.17%
* GradientBoostingRegressor train accuracy score 79.06% and test accuracy score 76.63%
* RandomForestRegressor train accuracy score 95.43% and test accuracy score 78.54%
* KNeighborsRegressor train accuracy score 83.78% and test accuracy score 74.90%

# Conclusion

# *Choosing GradientBoostingRegressor model since the Accuracy score and test scores are greater and clsoe to each other.*[*¶*](http://localhost:8888/notebooks/MLModels-main/Flight_Price_Prediction.ipynb#Choosing-GradientBoostingRegressor-model-since-the-Accuracy-score-and-test-scores-are-greater-and-clsoe-to-each-other.)