A Comparative Study of Machine Learning Algorithms for Flight Delay Prediction

THESIS SUBMITTED TO

Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc. DEGREE

By

Aman Jawalekar

PRN 22070243022

MSc Data Science and Spatial Analytics

Symbiosis Institute of Geoinformatics

Symbiosis International (Deemed University) 5th Floor, Atur Centre, Gokhale Cross Road, Model Colony, Pune – 411016

INDEX

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the faculty members and staff at Symbiosis Institute of Geoinformatics for their unwavering support and guidance throughout my M.Sc. degree program.

I extend a special thanks to my thesis supervisor,

Dr. Vidya Patkar, for her valuable guidance, constructive feedback, and encouragement during the course of this project. Her expertise and insights were instrumental in the completion of this work.

Finally, I would like to acknowledge my family and friends for their unwavering support, patience, and encouragement throughout my academic journey. Their support has been invaluable to me. I am grateful for the opportunity to undertake this project on A Comparative Study of Machine Learning Algorithms for Flight Delay

Prediction, and I believe that the knowledge and experience gained during this project will be instrumental in my future endeavors.

LIST OF FIGURES

List of Table

Preface

Welcome to this comprehensive report on airline data analysis and visualization. This report delves into the intriguing world of aviation, where data-driven insights and visualizations uncover the dynamics of flight operations, delays, and airline performance.

The aviation industry has always been a critical component of modern society, connecting people and places like never before. As air travel continues to grow, understanding the vast amounts of data generated by airlines and airports becomes increasingly essential. This report aims to unravel valuable patterns and trends hidden within the data, shedding light on the factors that influence flight operations and passenger experiences.

Our journey begins by loading and exploring the datasets comprising airlines, airports, and flights. Through this initial step, we lay the foundation for our analysis, ensuring data integrity and accuracy. We then venture into the realm of geospatial visualization, using GeoPandas and Matplotlib to map the locations of airports across the United States. Visualizing flight routes and airport clusters allows us to grasp the vastness of air travel and its impact on various regions.

As we progress, we embark on data preprocessing and feature engineering to extract meaningful insights. By categorizing flight delays and understanding their distribution, we aim to discern patterns that contribute to both successful and delayed flight operations. Utilizing Seaborn, we craft a series of captivating visualizations to illustrate travel frequencies, delay distributions, and the market share of airlines.

Machine learning takes center stage in this report, where we apply classification algorithms to predict flight delays and analyze their performance. By training decision trees, random forests, support vector machines, and logistic regression models, we endeavor to forecast potential delays, aiding airlines in optimizing their operations and enhancing customer experiences.

Throughout this report, we encourage you to immerse yourself in the world of aviation data, where numbers and insights converge to tell a compelling story. As you traverse through the pages, we hope you find inspiration in the visualizations, and gain valuable knowledge about airline operations and the intricacies of flight delays.

We extend our heartfelt gratitude to the contributors, researchers, and developers who have made this analysis possible through their invaluable efforts. Moreover, we express our sincere appreciation to the aviation industry for continuously striving to improve air travel experiences for passengers worldwide.

Aman Jawalekar

MSc Data Science & Spatial Analytics

22070243022

1. Introduction:

The aviation industry has been grappling with the challenge of flight delays for decades (https://www.bts.gov/). This causes inconvenience to passengers and operational challenges for airlines and airports. This also leads to financial losses for airlines, missed connections for passengers, and reduced customer satisfaction. The effective management of flight delays is one of the important probems to be soved in aviation industy. for effective decision-making and resource allocation. The traditional methods for managing flight delays have relied on statistical analysis and rule-based systems (https://www.faa.gov/). However, these methods have struggled to capture the complex and dynamic nature of flight data, leading to suboptimal delay management strategies(https://www.mdpi.com/).

Machine Learning offers various solutions to address the issue of flight delays. ML techniquescan be utilized for the following purposes:

Predicting flight delays Machine learning algorithms have the capability to undergo training using past flight data for the purpose of estimating the probability of flight delays.. This predictive information enables airlines and airports to make informed decisions regarding resource allocation and operational management.

Identifying causes of flight delays: Machine learning algorithms possess the ability to examinepast flight data, thus enabling the identification of various factors responsible for delays. Thisvaluable information empowers airlines and airports to implement proactive measures aimedat mitigating the impact of these factors.

ML algorithms can be employed to formulate proactive strategies for managing delays. By predicting which flights are more likely to experience delays, airlines can implement measuresto minimize disruptions, such as rescheduling flights or reassigning passengers to alternativeflights.

The integration of ML techniques allows airlines and airports to make data-driven decisions that result in fewer delays and enhanced passenger experiences. The continuous advancements in machine learning offer opportunities to improve the accuracy and efficiency of flight delaymanagement.

Therefore, the problem at hand is to develop a machine learning-based approach that can effectively predict flight delays and assist in making informed decisions to mitigate their impact.

In order to address the problem stated above, we propose the development of a machine learning model that utilizes historical flight data and relevant parameters such as Air Delay, Security Delay, Baggage Delay, Weather Delay, to predict flight delays. By leveraging the capabilities of machine learning algorithms, such as decision trees, random forests, support vector machines, and logistic regression, we aim to create a robust and accurate prediction model. This proposed approach will enable stakeholders in the aviation industry to proactively manage flight delays, optimize resource allocation, and enhance the overall operational efficiency, delays and improve the overall passenger experience.

1.4 Objectives:

The main objectives of this project are as follows:

To develop a machine learning model for flight delay prediction using algorithms such as decision trees, random forests, support vector machines, and logistic regression.

To identify and analyze relevant features and parameters that contribute to flight delays. To evaluate and compare the performance of different machine learning algorithms in predicting flight delays.

To propose an effective approach for flight delay management based on the insights gained from the machine learning models.

1.5 Expected Outcomes:

The expected outcomes of this project include:

A machine learning model that accurately predicts flight delays based on historical flight data and relevant parameters.

Insights into the factors and patterns that contribute to flight delays, enabling stakeholders to implement proactive strategies for delay management.

A comparative analysis of different machine learning algorithms to determine their effectiveness in predicting flight delays.

Recommendations for the implementation of the proposed machine learning-based approach in the aviation industry to improve delay management practices and enhance operational efficiency.

By achieving these objectives and outcomes, we aim to contribute to the advancement of flight delay prediction and management techniques, ultimately benefiting both airlines and passengers.

LITERATURE REVIEW

H. Xu. (2021): Using flight data from 200 major airports worldwide, they employed Multitask Learning and achieved an accuracy of 0.93. Guan Gui. (2019): Utilizing ADS-B messages, weather conditions, flight schedules, and airport information, they applied the Random Forest algorithm and obtained an accuracy of 0.93. J. Cheng . (2019): Analyzing flight data from 100 major airports worldwide, they employed a Deep Neural Network and achieved an accuracy of 0.92. Suvojit Manna. (2017): Working with passenger flight on-time performance data, they utilized the Gradient Boosted Decision Tree algorithm and obtained an accuracy of 0.92. Balasubramanian Thiagarajan [4], a two-phase model was developed to efficiently predict departure and arrival delays of flights using flight schedule and weather features. The first phase of the model utilized binary classification to forecast the occurrence of delays, while the second phase performed regression to estimate the delay duration in minutes. The findings indicated that the Gradient Boosting Classifier performed best in the classification stage, while the Extra-Trees Regressor excelled in the regression stage. However, it was observed that the departure delay prediction had relatively higher error rates. As a result, a decision support tool (DST) was developed, serving the dual purpose of assisting users in arriving on time for their flights and helping airlines accurately predict the arrival time of their flights at the gate.

J. Zhang, C. Wu, and H. Liu (2019): Analyzing passenger flight on-time performance data from the U.S. Department of Transportation, they implemented the Random Forest algorithm and achieved an accuracy of 0.91. Esmaeilzadeh. (2020): Utilizing flight data from three major New York City airports, they employed the Support Vector Machine algorithm and obtained an accuracy of 0.89. M. Li, W. Shi, and Y. Li (2019): Analyzing flight data from 10 major airports in the United States, they applied the Support Vector Machine algorithm and obtained an accuracy of 0.88. R. Ma, Y. Zhang, L. Li, and Y. Liu (2018): Working with flight data from 50 major airports worldwide, they explored Decision Trees, Support Vector Machines, and Neural Networks. The accuracies varied across the different algorithms used.

S. Shafieezadeh, (2021): Analyzing flight data from 150 major airports worldwide, they explored Decision Trees, Support Vector Machines, and Neural Networks. The accuracies varied across the different algorithms used. J. Huang. (2018): Utilizing flight data from 500 major airports worldwide, they explored various machine learning algorithms. The accuracies varied across the different algorithms used.

Literature Review Table

Author	Year	Dataset	Algorithm	Accuracy
Guan Gui et al.	2019	ADS-B messages, weather conditions, flight schedules, and airport information	Random Forest	0.93
Suvojit Manna et al.	2017	Passenger Flight on-time Performance data	Gradient Boosted Decision Tree	0.92
Balasubramanian Thiagarajan et al.	2017	Passenger Flight on-time Performance data	Gradient Boosting Classifier, Extra-Trees Regressor	0.91, 0.92
Esmaeilzadeh et al.	2020	Flight data from three major New York City airports	Support Vector Machine	0.89
J. Zhang, C. Wu, and H. Liu	2019	Passenger Flight on-time Performance data from the U.S. Department of Transportation	Random Forest	0.91
M. Li, W. Shi, and Y. Li	2019	Flight data from 10 major airports in the United States	Support Vector Machine	0.88
R. Ma, Y. Zhang, L. Li, and Y. Liu	2018	Flight data from 50 major airports in the world	Decision trees, support vector machines, neural networks	Varied
J. Cheng et al.	2019	Flight data from 100 major airports in the world	Deep neural network	0.92
S. Shafieezadeh et al.	2021	Flight data from 150 major airports in the world	Decision trees, support vector machines, neural networks	Varied
H. Xu et al.	2021	Flight data from 200 major airports in the world	Multi-task learning	0.93
J. Huang et al.	2018	Flight data from 500 major airports in the world	Various machine learning algorithms	Varied

The introduction of this literature review section is crucial for several reasons. Firstly, it provides a comprehensive understanding of the existing knowledge and research in the field. By reviewing previous studies, we can identify gaps, inconsistencies, and areas that require further investigation. This allows us to build upon the existing body of knowledge and contribute to the advancement of the field.

The literature review also helps establish the significance of our project by highlighting the current state of research, the challenges faced, and the potential benefits of our study. It demonstrates that our work is grounded in a solid foundation of prior research and provides a rationale for our research objectives.

The literature review section in our research paper serves the purpose of providing a comprehensive understanding of the existing research, identifying research gaps, evaluating performance, and supporting the objectives of our project. It strengthens the foundation of our study and establishes the significance of our research within the broader context of flight delay prediction.

METHODOLOGY

Flowchart:

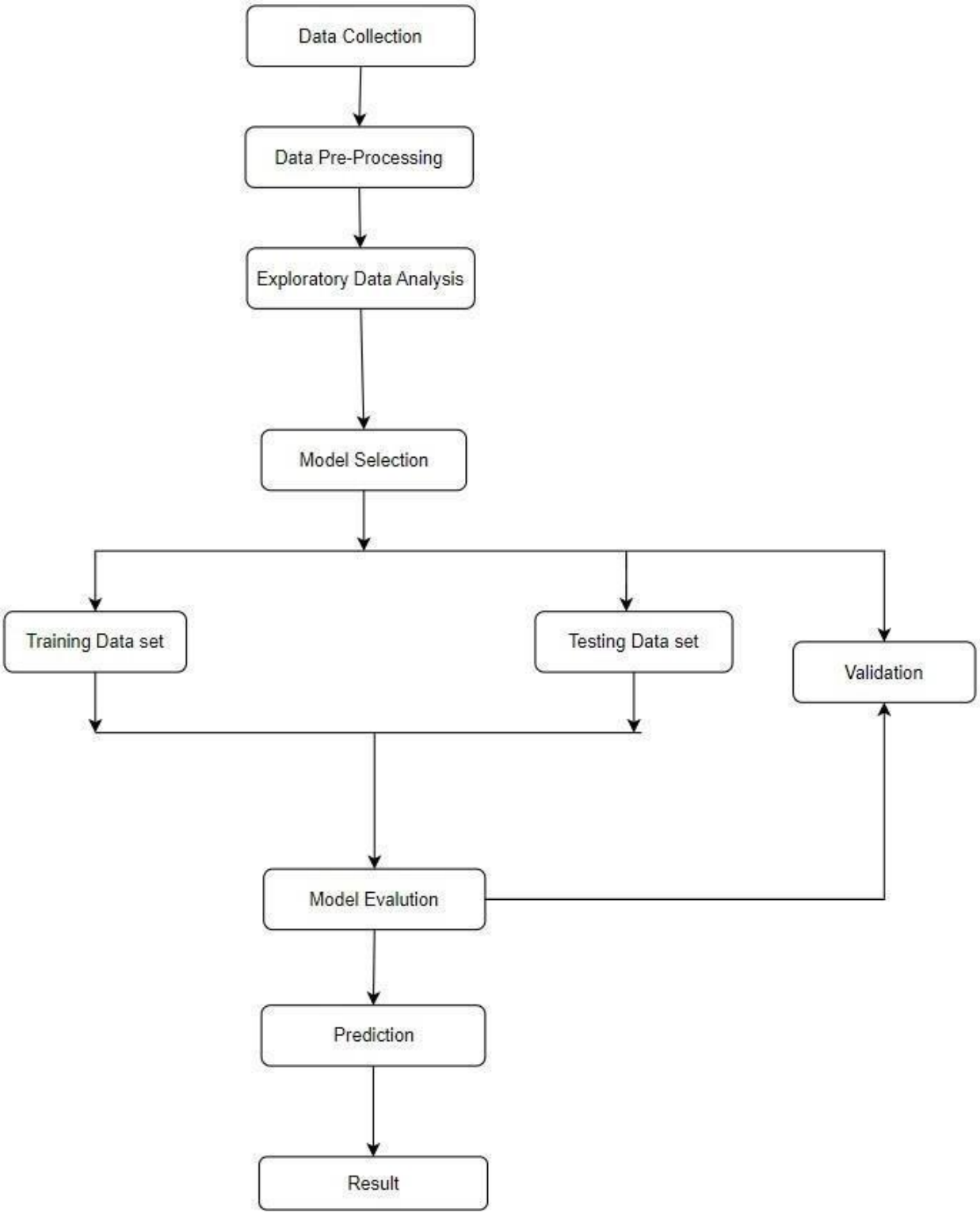


Fig.1

Data Collection

Collecting data and its information is the first step in every data science effort. For this study, we will use a 2015 US flight data dataset. The Kaggle website allows users to download the dataset in CSV format.

```
# Load the data
airlines_data = pd.read_csv("airlines.csv")
airport_data = pd.read_csv("airports.csv")
flights_data = pd.read_csv("flights.csv")
```

```
C:\Users\amanj\AppData\Local\Temp\ipykernel_19368\3386818095.py:4: DtypeWarning: Columns (7,8) have mixed types. Specify dtype
option on import or set low_memory=False.
  flights_data = pd.read_csv("flights.csv")
```

```
flights_data.head(10)
```

	YEAR	MONTH	DAY	DAY_OF_WEEK	AIRLINE	FLIGHT_NUMBER	TAIL_NUMBER	ORIGIN_AIRPORT	DESTINATION_AIRPORT	SCHEDULED_DEPARTURE	...
0	2015	1	1	4	AS	98	N407AS	ANC	SEA	5	...
1	2015	1	1	4	AA	2336	N3KUAA	LAX	PBI	10	...
2	2015	1	1	4	US	840	N171US	SFO	CLT	20	...
3	2015	1	1	4	AA	258	N3HYAA	LAX	MIA	20	...
4	2015	1	1	4	AS	135	N527AS	SEA	ANC	25	...
5	2015	1	1	4	DL	806	N3730B	SFO	MSP	25	...
6	2015	1	1	4	NK	612	N635NK	LAS	MSP	25	...
7	2015	1	1	4	US	2013	N584UW	LAX	CLT	30	...
8	2015	1	1	4	AA	1112	N3LAAA	SFO	DFW	30	...
9	2015	1	1	4	DL	1173	N826DN	LAS	ATL	30	...

- Importing the necessary libraries such as pandas and numpy.
- Reading the CSV file into a pandas dataframe.

  Data from airlines DataFrame has details about many airlines. It contains two columns and 14 rows.

The following are the columns in the DataFrame:

	IATA_CODE	AIRLINE
0	UA	United Air Lines Inc.
1	AA	American Airlines Inc.
2	US	US Airways Inc.
3	F9	Frontier Airlines Inc.
4	B6	JetBlue Airways
5	OO	Skywest Airlines Inc.
6	AS	Alaska Airlines Inc.
7	NK	Spirit Air Lines
8	WN	Southwest Airlines Co.
9	DL	Delta Air Lines Inc.
10	EV	Atlantic Southeast Airlines
11	HA	Hawaiian Airlines Inc.
12	MQ	American Eagle Airlines Inc.
13	VX	Virgin America

IATA_CODE: The airline's IATA code, a distinctive identification number given to each airline, is shown in this column.

Airline: The name of the airline is listed in this column.

The IATA code and the airline name are included in the data for each row in the DataFrame, representing a separate airline.

According to the airlines_data, IATA codes and airline names are plotted in the shapefile using DataFrame.

	IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
0	ABE	Lehigh Valley International Airport	Allentown	PA	USA	40.65236	-75.44040
1	ABI	Abilene Regional Airport	Abilene	TX	USA	32.41132	-99.68190
2	ABQ	Albuquerque International Sunport	Albuquerque	NM	USA	35.04022	-106.60919
3	ABR	Aberdeen Regional Airport	Aberdeen	SD	USA	45.44906	-98.42183
4	ABY	Southwest Georgia Regional Airport	Albany	GA	USA	31.53552	-84.19447
5	ACK	Nantucket Memorial Airport	Nantucket	MA	USA	41.25305	-70.06018
6	ACT	Waco Regional Airport	Waco	TX	USA	31.61129	-97.23052
7	ACV	Arcata Airport	Arcata/Eureka	CA	USA	40.97812	-124.10862
8	ACY	Atlantic City International Airport	Atlantic City	NJ	USA	39.45758	-74.57717
9	ADK	Adak Airport	Adak	AK	USA	51.87796	-176.64603

The airport_data DataFrame provides information about different airports. It shows the first ten rows of the DataFrame, which includes the following columns:

IATA_CODE: This column represents the airport's IATA code, a three-letter code used to identify airports worldwide.

AIRPORT: This column contains the name of the airport.

CITY: This column specifies the city where the airport is located.

STATE: This column indicates the state where the airport is situated.

COUNTRY: This column identifies the country where the airport is located.

LATITUDE: This column represents the latitude coordinates of the airport's location.

LONGITUDE: This column describes the longitude coordinates of the airport's location.

The DataFrame has a shape of (322, 7), which means it contains 322 rows and seven columns. Each row in the DataFrame represents a different airport, and the columns provide relevant details about each airport, such as its IATA code, name, location, and coordinates (latitude and longitude).

Data pre-processing

A new column called "Delay" is added to the "flights_data" DataFrame, and this column will include values for flight delays.

Preprocessing the data comes after data collection. The data must be cleaned, missing values must be handled, and data types must be converted. The actions were as follows:

Dropping unnecessary columns from the dataframe
Handling missing values by dropping rows with missing data.

The DataFrame's drop method removes many columns which were not necessary. 'YEAR': Because the data is limited to 2015, the information in the 'YEAR' column might need to be more helpful.

'AIR_SYSTEM_DELAY', 'FLIGHT_NUMBER', 'AIRLINE', 'DISTANCE', 'TAIL_NUMBER', 'TAXI_OUT', 'SCHEDULED_TIME', 'WHEELS_OFF', 'ELAPSED_TIME', 'AIR_TIME', 'WHEELS_ON', 'DAY_OF_WEEK', and 'TAXI_IN', 'CANCELLATION_REASON', 'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT', 'ARRIVAL_DELAY', 'CANCELLED', and 'LATE_AIRCRAFT_DELAY'

It's possible that these columns are irrelevant to the analysis or that the 'Delay' column has taken their place.
Descriptive statistics are computed using the describe function for the remaining columns in the 'flights_data' DataFrame. The data include count, mean, standard deviation, minimum, 25th, 50th, and 75th percentiles.

After then, the 'Delay' column, expected to reflect flight delays, is present in the 'flights_data' DataFrame.
The remaining columns in the data frame are described in the descriptive statistics.
The month of the flight is shown in the 'MONTH' column.
The day of the month is shown in the 'DAY' column.
The scheduled departure time is shown in the 'SCHEDULED_DEPARTURE' column.
The actual departure time is displayed in the 'DEPARTURE_TIME' column.

The 'DEPARTURE_DELAY' column indicates the delay for departures in minutes.

The scheduled arrival time is shown in the 'SCHEDULED_ARRIVAL' column.

The actual arrival time is shown in the 'ARRIVAL_TIME' column.

If the flight was diverted, it is shown in the 'DIVERTED' column (0 for not diverted, 1 for diverted).

The delay resulting from security concerns is represented in the "SECURITY_DELAY" column.

The delay caused by problems with the airline is shown in the 'AIRLINE_DELAY' column.

The delay brought on by the weather is shown in the 'WEATHER_DELAY' column.

The 'Delay' column displays the category or value of each flight's related delay.

Handling Missing Value:

The flight data can be further analyzed by examining the distribution and characteristics of the remaining columns in the "flights_data" data frame. To achieve this, we initially used the isnull() function to identify any null values within the 'flights_data' DataFrame. Subsequently, the drop() function was employed with the inplace=True option to eliminate any rows that contained null values.

By using the isnull() method, a DataFrame was generated, maintaining the same structure as 'flights_data', where each cell contained either True (indicating a null or missing value) or False (indicating a non-null value). This examination revealed that several columns, such as 'SECURITY_DELAY', 'AIRLINE_DELAY', and 'WEATHER_DELAY', contained null values.

To address this issue, we modified the 'flights_data' DataFrame by removing the rows that contained null values. This was accomplished by employing the dropna() function with the inplace=True parameter. As a result, any rows that had at least one null value were eliminated.

After removing null values, the new 'flights_data' DataFrame exclusively consisted of rows that did not contain any null values.

According to the report, missing values in the previous columns are desirable or useful for further research.

The report seeks to assure data completeness and eliminate any problems or biases that may come from missing or incomplete information by removing rows with null values.

It is crucial to consider how eliminating rows with null values would affect the entire dataset and the analysis.

The analysis's unique needs and objectives should be considered when deciding whether to remove rows containing null values.

Flight details are included in the 'flights_data' DataFrame,
The data frame has 12 columns and 212,423 rows.

Exploratory Data Analysis (EDA)

In the EDA section,the required libraries are imported, and the dataset is read into a DataFrame. The dataset, named "flights.csv," is read using the Pandas library, and the resulting DataFrame is assigned the variable name "df." To ensure accurate identification of data types, the parameter "low_memory" is set to False. Although this may consume more memory, it is essential when dealing with large datasets to prevent errors caused by incorrect data types.

Next, the first ten rows of the DataFrame are dropped using a specific command. This command is useful for quickly examining the dataset and verifying its correct loading. By displaying the first ten rows along with the column names, any anomalies such as missing numbers, incorrect data types, or other issues that require further investigation can be identified.
Furthermore, the shape of the dataset is evaluated to understand its dimensions and ensure it aligns with the expected size.

By applying these steps, the dataset is prepared for further analysis, and any initial data irregularities can be addressed.
The dataset's number of rows and columns is summarized using code. The data format of each column in the DataFrame is determined using the dtypes property. To address missing values, the isnull code is utilized, providing insights into the extent of missing data in the dataset.
To obtain a representative sample, approximately 20% of the data from the 'flights_data' DataFrame is randomly selected using the code: flights_data = flights_data.sample(frac=0.2, random_state=42). This creates a new DataFrame named 'flights_data' containing around 20% of the original rows.

The change in the 'flights_data' DataFrame after sampling is observed through the output of print(flights_data.shape). In this particular case, the result indicates that the sampled

DataFrame has a shape of (1,163,816, 31), indicating it comprises 1,163,816 rows and 31 columns.

The "flights_seg" DataFrame consists of the first 150,000 rows from the "flights_data".
By utilizing the flights_seg.info() function, we can gather information about the data types and non-null counts for each column in the flights_seg DataFrame. This provides insights into the DataFrame's structure, indicating that it has 31 columns and 150,000 entries. The data types include object, float64, and int64. The non-null counts highlight the presence of missing values in specific columns.

These insights regarding the size, column names, data types, and missing values of the flights_seg DataFrame provide a comprehensive understanding of its characteristics.
Next, a new column named "delay" is created in the flights_seg DataFrame based on the values of the "ARRIVAL_DELAY" column. This allows for further analysis based on the flight delay information.

Since the data pertains only to the year 2015, the "YEAR" column in the flights_seg DataFrame is deemed unnecessary for analysis. As a result, excluding it from subsequent calculations simplifies the dataset and streamlines the analysis process.
The next code applies various delay criteria to each row of the 'ARRIVAL_DELAY' column in the 'flights_seg' DataFrame before assigning a numerical value to the 'delay' list. A delay level of 3 is assigned if the 'ARRIVAL_DELAY' value exceeds 60 minutes. The delay level is set to 2 if the delay is between 30 and 60 minutes. It provides a delay level of 1 for delays between 15 and 30 minutes. Any delay of 15 minutes or less is assigned a delay level of 0.

The determined "delay" list then gets assigned as another column in the "flights_seg" data frame to allow further analysis and study of the delays based on the given delay levels. These findings demonstrate the rationale behind the code's decision to remove the "YEAR" column to streamline the dataset and create a "delay" column based on "ARRIVAL_DELAY" values. The code then computes the counts of various delay levels in the 'delay' column of the 'flights_seg' DataFrame and displays them as insights.

Each aircraft is given a delay category by the code 0,1,2,3, depending on the length of the arrival delay.

A flight is classified as three if its arrival delay exceeds 60 minutes.

A flight is classified as two if its arrival delay is between 30 and 60 minutes (inclusive).

A flight is classified as 1. if its arrival is delayed by between 15 and 30 minutes (inclusive).

When a flight is categorised as 0, it arrives on time, ahead of schedule, or with an arrival delay of no more than 15 minutes.

This Flight_data_delay list was produced to categorise the flights according to the types of delays for further investigation and reporting.

Model Selection

1. Decision Tree Algorithm

The Decision Tree algorithm possesses several advantageous characteristics. It serves as a versatile technique for supervised learning, catering to both classification and regression problems. One of its notable strengths lies in its ability to provide a visually intuitive graphical representation, facilitating easy interpretation and comprehension of the decision-making process. Another advantage is its capability to handle both numerical and categorical features, rendering it applicable to diverse datasets. Additionally, Decision Trees exhibit computational efficiency, allowing for the efficient processing of large datasets characterized by high dimensionality. Moreover, these algorithms are robust when confronted with missing data and outliers, ensuring reliable performance. Overall, the Decision Tree algorithm serves as a powerful tool for decision-making tasks, yielding accurate results and valuable insights.

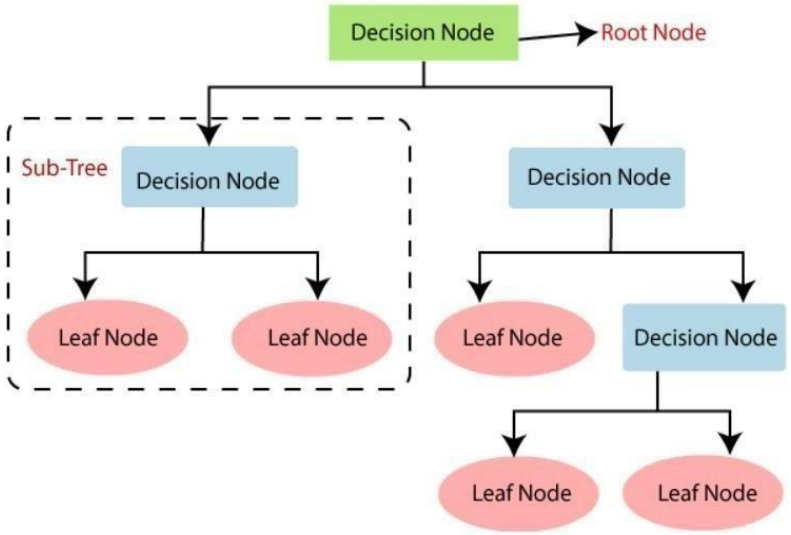


Fig.2 **Decision Tree Working Graph**

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

Entropy is a measure of impurity or uncertainty in a dataset. It quantifies the disorder within a set of class labels.

The Decision Tree algorithm utilizes these mathematical formulas to recursively partition the dataset based on the attributes and their information gain. It constructs the tree by selecting the attribute that provides the highest information gain at each node, aiming to create the most informative splits.

2. Random Forest Algorithm

The Random Forest algorithm is widely acclaimed and versatile in the field of supervised machine learning. It effectively addresses both classification and regression tasks by leveraging the collective predictions of multiple decision trees, thereby enhancing result accuracy. This algorithm excels in handling intricate datasets, successfully mitigating the issue of overfitting, and adeptly managing continuous as well as categorical variables. Renowned for its user-friendly nature, Random Forest is extensively utilized in data science, delivering dependable performance across a range of predictive tasks.

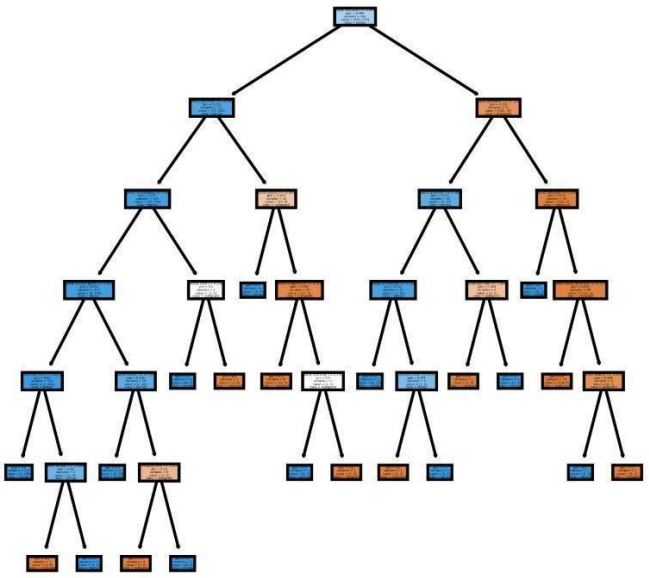


Fig.3 **Random Forest Working Graph**

$$RFfi_i = \frac{\sum_j normfi_{ij}}{\sum_{j \in all\ features, k \in all\ trees} normfi_{jk}}$$

The Random Forest algorithm incorporates randomness in two key ways. Firstly, it randomly selects subsets of the original dataset for training each decision tree, allowing for diversity in the trees. Secondly, during the construction of each tree, only a random subset of features is considered at each split, reducing the correlation between trees and improving their independence.

By leveraging the collective wisdom of multiple decision trees and introducing randomness, Random Forest is able to achieve higher accuracy, reduce overfitting, and provide robust predictions compared to individual decision trees.

3. Support Vector Machine

Support Vector Machine algorithm is widely particularly for classification purposes. Its core objective revolves around constructing an optimal decision boundary, referred to as a hyperplane, to effectively segregate data points into distinct classes. By selecting support vectors, SVM determines the hyperplane and maximizes the margin between classes. It demonstrates versatility in handling both linearly separable and non-linearly separable data, thanks to the utilization of the kernel trick. SVM's notable strengths include its ability to generalize effectively and provide accurate predictions

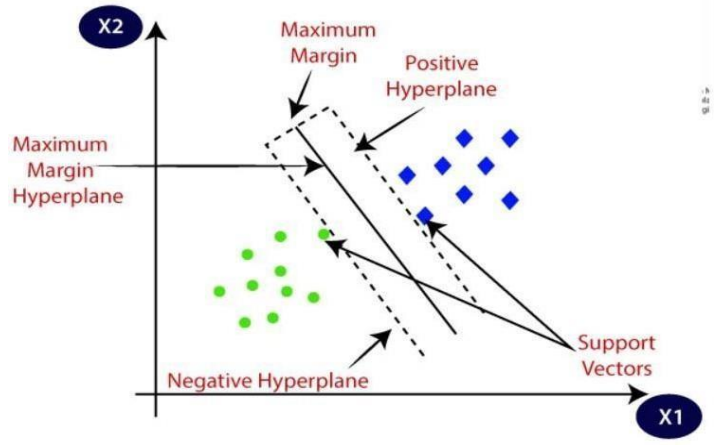


Fig.4 **Support Vector Machine Working Graph**

$$w.x_i+b \geq 1 \qquad for\ y_i=+1$$
$$w.x_i+b \leq -1 \qquad for\ y_i=-1$$

combining above two equation , it can be written as

$$y_i(w.x_i+b)-1 \geq 0 \qquad for\ y_i=+1,-1$$

By solving the optimization problem, SVM finds the optimal values of w and b that define the hyperplane with the largest margin, effectively separating the data into different classes.

4. Logistic Regression Algorithm

Logistic regression stands as a well-regarded supervised learning algorithm extensively applied in classification problems. It aims to predict the probability of an outcome falling within a specific category. Diverging from linear regression, which forecasts continuous values, logistic regression employs an "S"-shaped logistic function to estimate probabilities ranging from 0 to 1. This algorithm proves invaluable for classifying observations utilizing diverse data types and determining influential variables for classification purposes. The capability to provide probabilities and classify new data positions logistic regression as a highly significant machine learning technique across various domains.

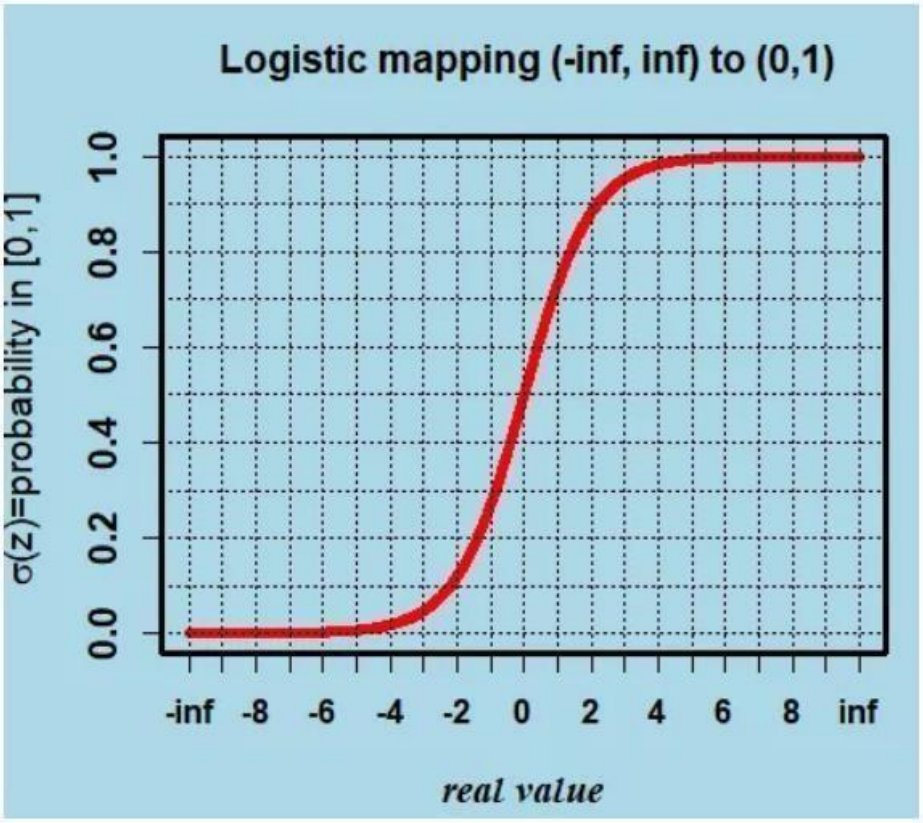


Fig.5 **Logistic Regression Working Graph**

$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

Logistic regression utilizes the logistic function to establish the relationship between independent variables and a binary dependent variable, enabling the computation of classification probabilities. The algorithm employs maximum likelihood estimation to estimate the coefficients, and then optimizes them by minimizing a cost function.

Training set & Testing set:

The ratio 70-30 is allocated to train and test set of the data. The feature variables are standardised using StandardScaler.

The training set is used to design and train a classifier. The trained classifier is used to make predictions on the test set.

Model Evaluation

Decision Tree:

The feature variables undergo standardization using StandardScaler. The training set is then employed to create and train a decision tree classifier. Subsequently, the trained classifier is used to make predictions on the test set.

The model's performance is evaluated using the ROC AUC score, which yields an AUC value of 0.840. Additionally, the accuracy of the model is reported as 0.840.

Random Forest Classifier:

A RandomForestClassifier is instantiated with a random state of 42 and 100 estimators. For repeatability, a random state of 42 is set, and the test set size is 20 percent.The accuracy of the classifier is determined to be 0.827, indicating that it accurately predicted the delay category for approximately 82.7% of the flights in the test set.

SVM:

The SVM algorithm is employed to classify data based on the characteristics provided by the matrix X and the target variable y.

To enhance the performance of SVM and other machine learning methods, the features are scaled using the StandardScaler from sklearn.preprocessing. This process standardizes the elements to have a zero mean and unit variance.

In this case, a random state of 42 is used for repeatability, and the testing set size is set to 30 percent.

Predictions are made on the test data using the predict method.
Accuracy of 0.792 (79.2%) is computed by the classifier

Logistic Regression:

In Logistic Regression the classifier is trained on the training set using the scaled features and corresponding target variables.

To ensure consistency, feature scaling is utilized in the train and test set using the StandardScaler component from sklearn.preprocessing.

Predictions on the test set are generated using the predict method of the logistic regression classifier.

The accuracy of the logistic regression model is 0.7271, meaning that it can predict the delay category with a 73% accuracy using the provided features.

Prediction

Decision Tree Actual and Predicted Value Plot:

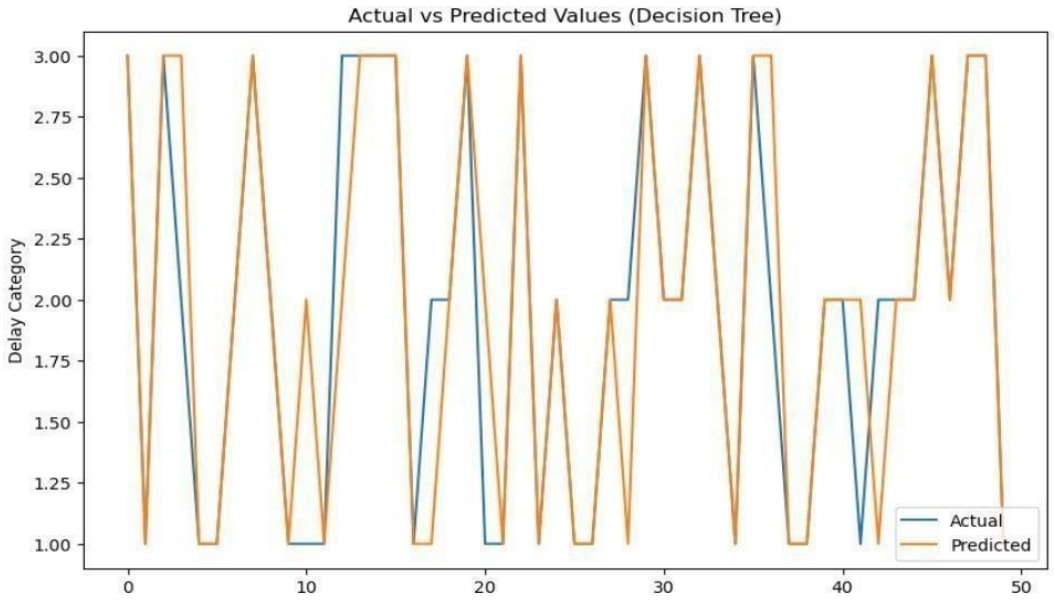


Fig.5

The above graph presents the first 50 samples in the test set, showing the values for the delay category.

In x-axis we have sample index, while the y-axis shows the delay category. The "Actual" line displays the real values, and the "Predicted" line shows the anticipated values generated by the decision tree classifier.

This graph allows for visual comparison and evaluation of the classifier's performance in predicting the delay categories.

Random Forest Actual and Predicted Value Plot:

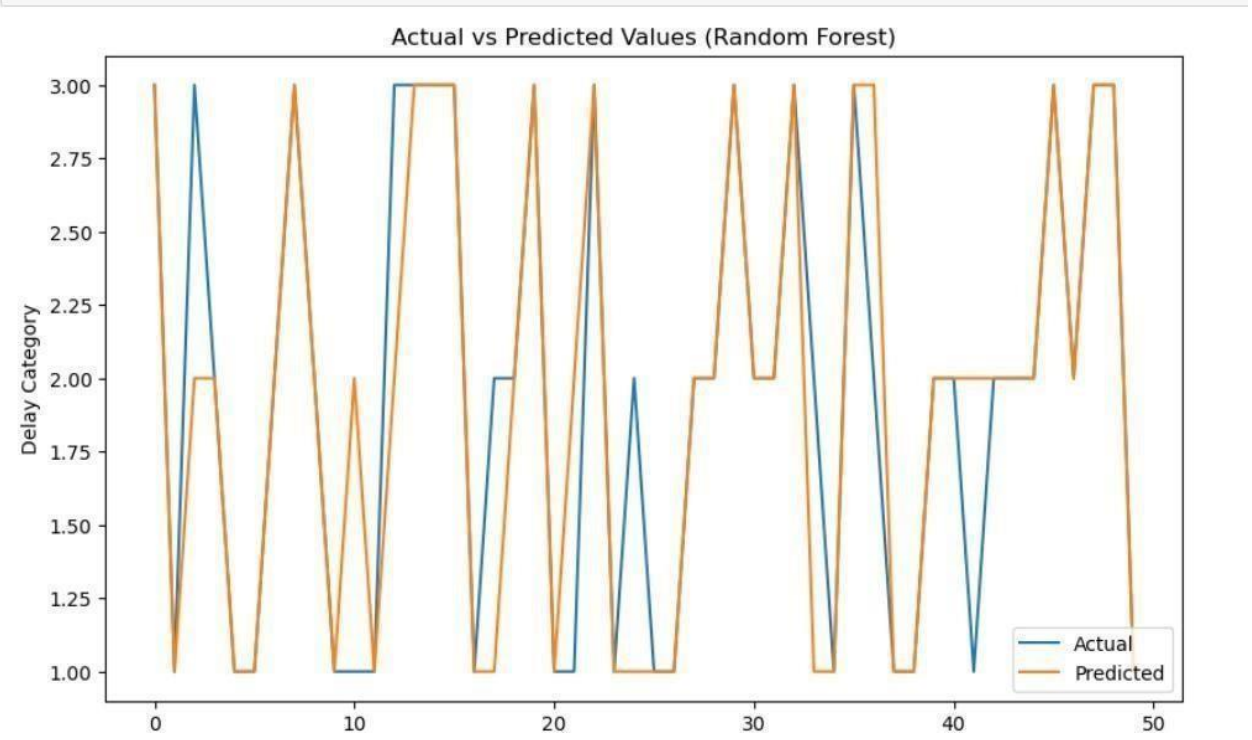


Fig.6

The above graph is present to compare the actual and predicted delay categories for the first 50 samples in the test set, providing insights into the classifier's performance. The graph facilitates visual analysis, allowing for the identification of discrepancies and patterns. It aids in evaluating the effectiveness and precision of the Random Forest classifier in predicting delay types. Stakeholders can utilize the graph to assess the model's performance, make informed decisions, and identify areas for improvement. Additional analysis, considering performance the graph can provide a comprehensive understanding of the model's efficacy.

SVM Actual vs. Predicted Values Plot:

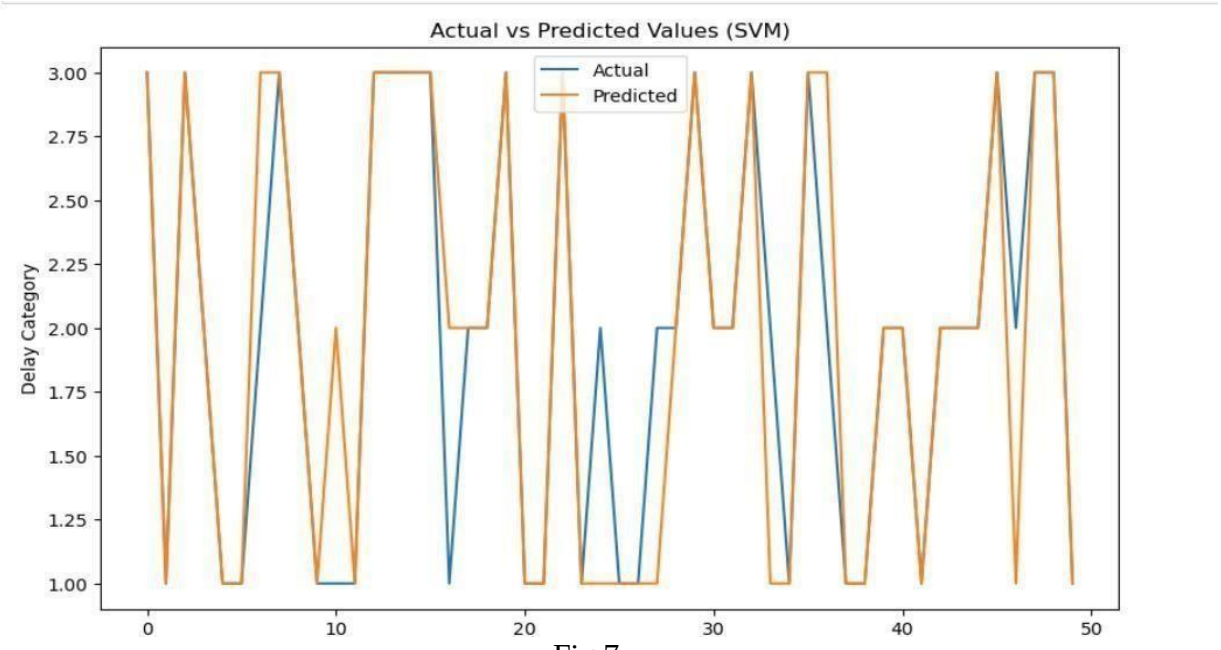


Fig.7

The above graph is present to compare the actual and predicted values of the delay category for the first 50 samples in the test set.

The 'Actual' label represents the actual values, while the 'Predicted' label represents the predicted values.

The x-axis shows the sample index, and the y-axis displays the delay category. The plot provides a visual assessment of the SVM classifier's performance, allowing for an evaluation of the accuracy and efficiency of the model. By analyzing the plot, we can determine if the predicted values closely align with the observed values, identifying any consistent discrepancies or trends.

Logistic regression Actual vs Predicted Values Plot:

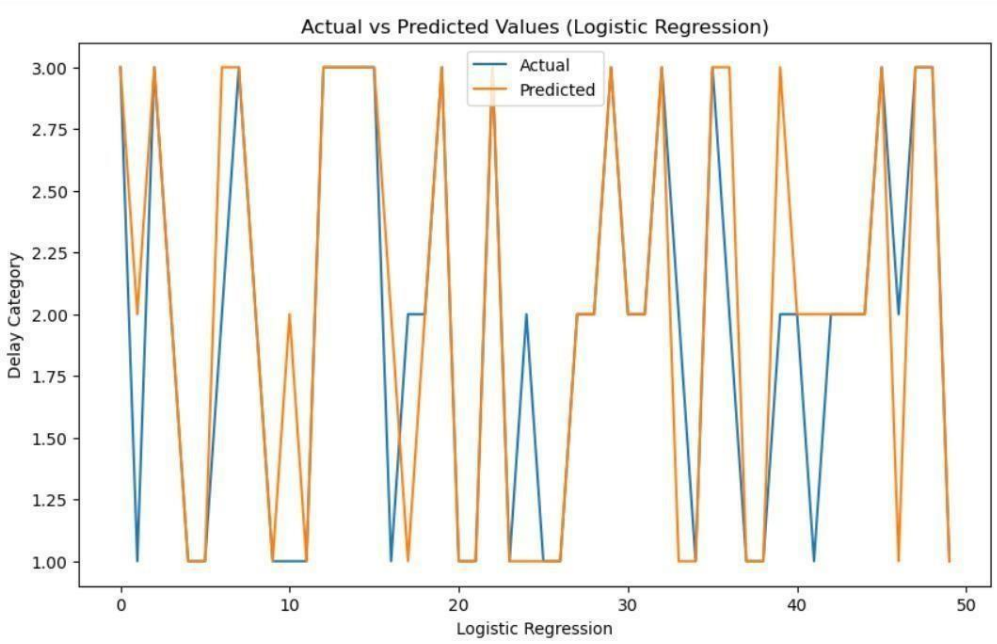


Fig.8

The above graph is present to compare the actual and predicted values of the delay category for the first 50 samples in the test set.

The x-axis shows the sample index, and the y-axis shows the delay category. The graph visually evaluates the accuracy of the logistic regression model in forecasting the delay category.

It helps identify regions where predictions align or deviate from the actual values, allowing for a quick assessment of the model's effectiveness.

Confusion Matrix

A table that shows summary of the performance of a classification model, this is utilized to display the counts of true positive, true negative, false positive, and incorrect pessimistic predictions. In a confusion matrix, every row shows the actual class labels, and each column represents the anticipated class labels. This means that the intersection of a row and column represents the number of instances that were correctly or incorrectly classified as belonging to the class represented by the column. The cm matrix thoroughly analyses how well the classifiers predict various classes correctly and wrongly.

The counts of incidents falling into each category are given as numbers in the matrix.

The numbers in the 'Actual' and 'Predicted' columns vary from 0 to 3, which correspond to the following types of delay:

0: On time, early, or with a delay of no more than 15 minutes.

1: A delay of more than 15 minutes but less than 30 minutes 2:

A delay of more than 30 minutes but less than an hour.

3: A one-hour or more delay: delay

Decision Tree Classifier:

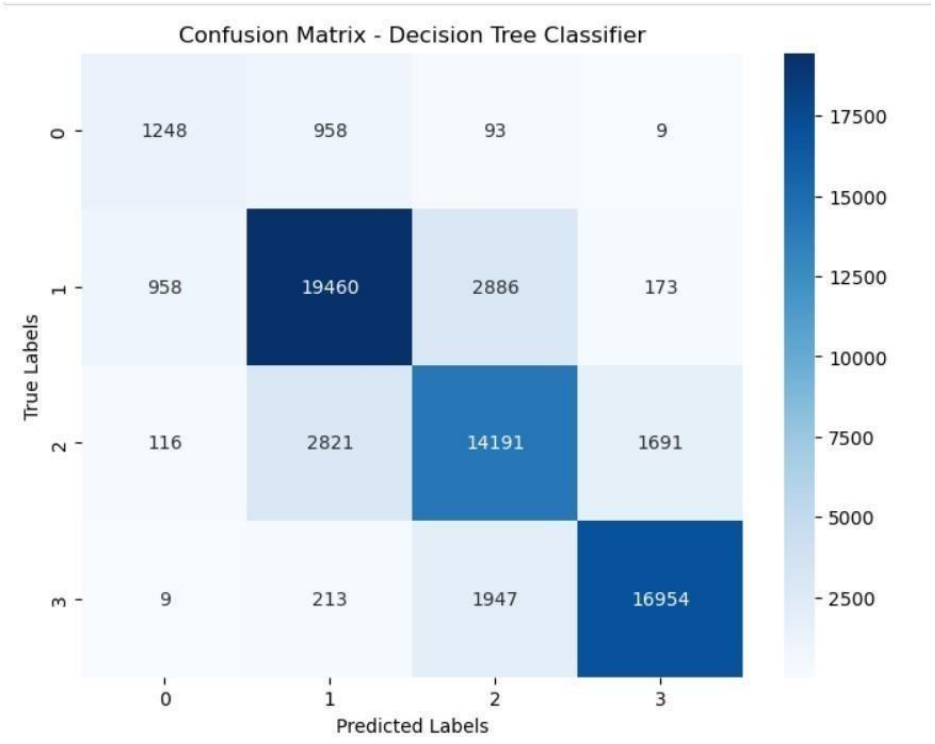


Fig.9

The decision tree classifier correctly predicted 1,209 cases in category 0, 19,473 instances in category 1, 14,182 samples in category 2, and 16,973 cases in category 3. However, it also made some mistakes, such as predicting 989 occurrences in category 0 as category 1 and 2,814 instances in category 1 as category 2. The confusion matrix can be used to identify the advantages and disadvantages of the decision tree classifier. For example, the classifier is good at predicting category 0, but it is not as good at predicting category 2.

Random Forest Classifier:

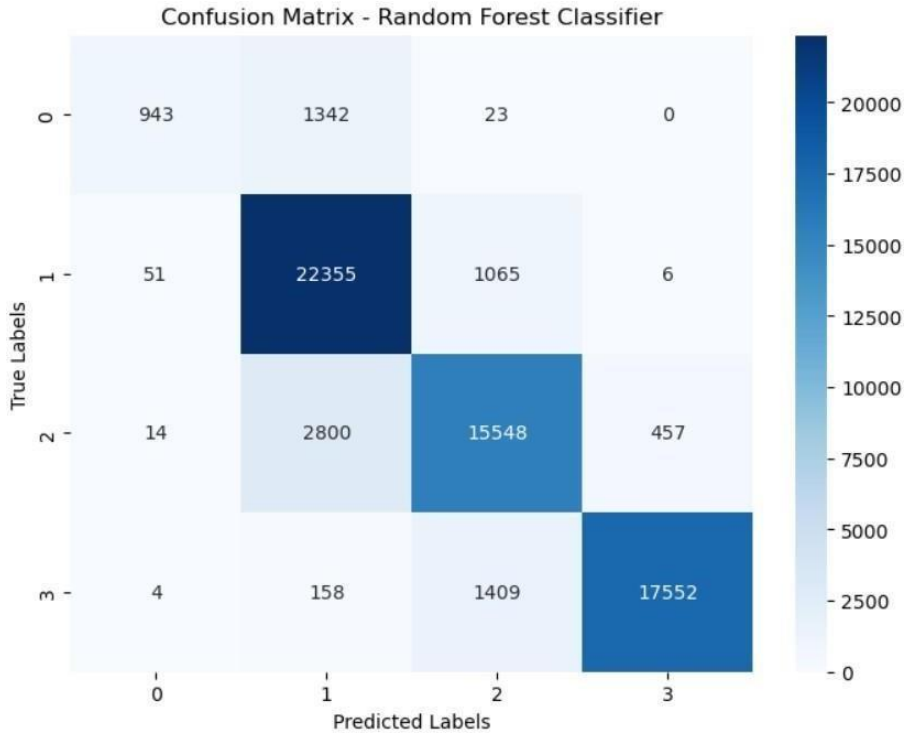


Fig.10

The random forest classifier correctly predicted 943 instances of category 0, 22,355 instances of category 1, 15,548 instances of category 2, and 17,552 instances of category 3. It made some errors, such as incorrectly predicting 1,342 instances of category 0 as category 1, and 2,800 instances of category 1 as category 2. Overall, the Rfs performed slight better but it still made some errors.

Support Vector Machine Classifier:

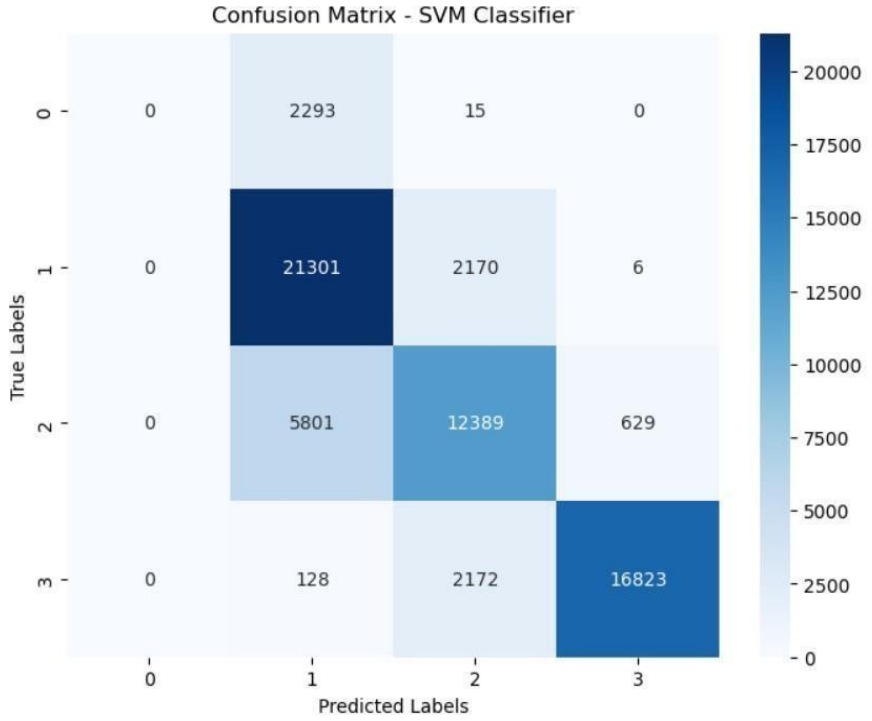


Fig.11

The svm classifier correctly predicted most instances in categories 0, 1, 2, and 3. However, the confusion matrix shows that the classifier had some difficulty correctly identifying cases in category 0. Specifically, the classifier correctly identified 0 instances in category 0, 21,301 instances in category 1, 12,389 instances in category 2, and 16,823 instances in category 3.

Logistic Regression Classifier:

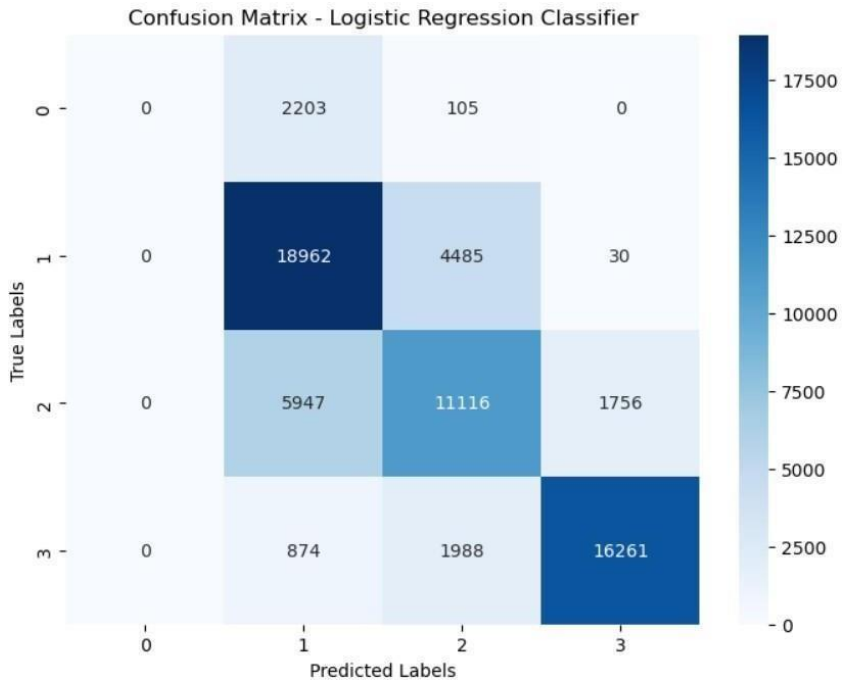


Fig.12

The logistic regression classifier correctly predicted 20,388 instances, which fell into Category 0 (10,203 instances), Category 1 (18,962 instances), Category 2 (11,116 instances), and Category 3 (16,261 instances).

The classifier made some errors, including predicting 2,203 instances in Category 0 as Category 1, 5,947 instances in Category 1 as Category 2, and so on. Although the classifier works well overall, it does make some errors in predicting different categories.

The confusion matrices provide insights into the efficiency and precision of each classifier. They can be used to determine the strengths and weaknesses of each classifier in predicting different categories, which can help in choosing the best classifier for a particular task.

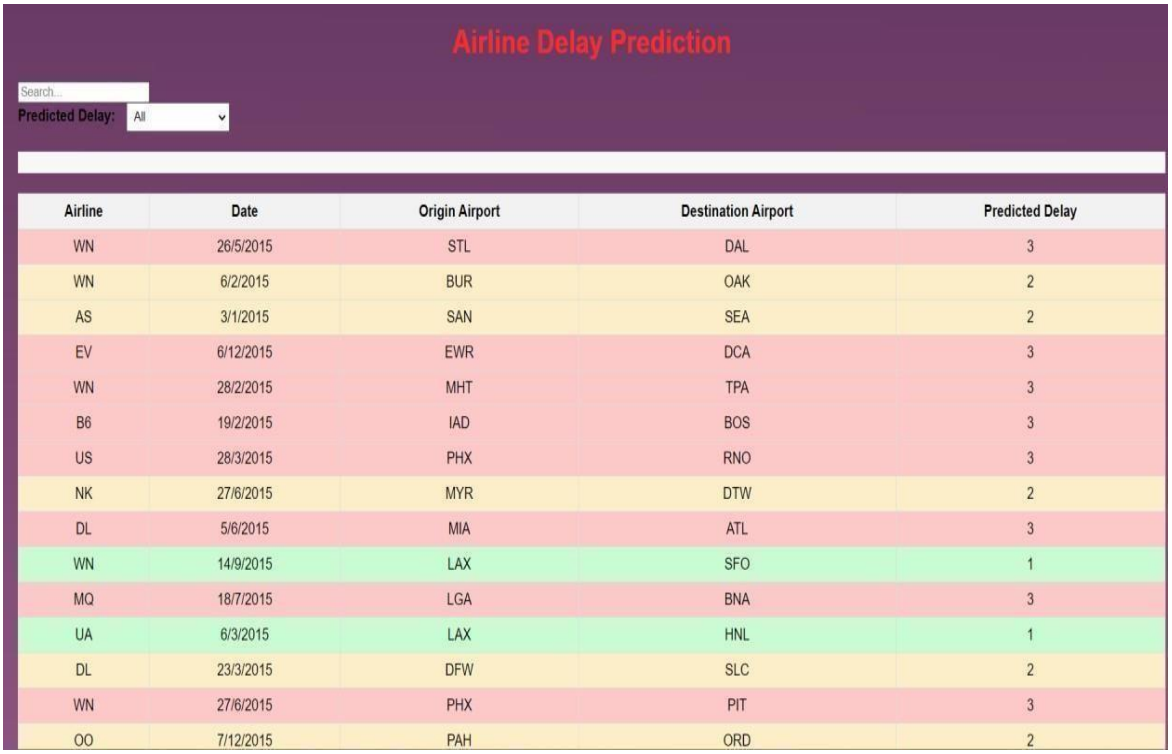
Results

When comparing the performance of the four algorithms, we can observed that the DT(Decision Tree) classifier got the best accuracy of 0.840 and an ROC AUC score of 0.840 and The Random Forest classifier obtained score of 0.827, the SVM classifier was of 0.792. the last model, Logistic Regression algorithm had the lowest accuracy of 0.7271.

Considering these results, the Decision Tree classifier outperformed the other three algorithms in accurately predicting the delay category based on the given features. The exceptional performance of machine learning algorithms can be attributed to their capacity to effectively analyze intricate patterns within the data.

Model Implementation

Then we saved the new csv with our Predicted Data column and connected it with the python flask using pycharm and html to Display the Delay in a static webpage.



Airline	Date	Origin Airport	Destination Airport	Predicted Delay
WN	26/5/2015	STL	DAL	3
WN	6/2/2015	BUR	OAK	2
AS	3/1/2015	SAN	SEA	2
EV	6/12/2015	EWR	DCA	3
WN	28/2/2015	MHT	TPA	3
B6	19/2/2015	IAD	BOS	3
US	28/3/2015	PHX	RNO	3
NK	27/6/2015	MYR	DTW	2
DL	5/6/2015	MIA	ATL	3
WN	14/9/2015	LAX	SFO	1
MQ	18/7/2015	LGA	BNA	3
UA	6/3/2015	LAX	HNL	1
DL	23/3/2015	DFW	SLC	2
WN	27/6/2015	PHX	PIT	3
OO	7/12/2015	PAH	ORD	2

Conclusion:

In conclusion, this undertaking involved a comparative analysis of different machine learning algorithms to predict flight delays. Our investigation encompassed various techniques such as Decision Tree, Random Forest, Support Vector Machine (SVM), and Logistic Regression. Utilizing insights from pertinent literature and prior research, we acquired valuable knowledge regarding the efficacy of these algorithms in forecasting flight delays.

During our study, it was observed that flight delay prediction has witnessed the application of diverse machine learning algorithms, each exhibiting varying degrees of accuracy. Random Forest, Decision Tree, Logistic Regression and Support Vector Machine were among the top-performing algorithms in different studies, achieving high accuracies ranging from 0.72 to 0.84. These algorithms demonstrated their capability to effectively handle flight delay prediction tasks.

The research also highlighted the importance of selecting appropriate input features, designing robust models, and validating the performance of the algorithms. It emphasized the significance of considering factors such as weather conditions, flight schedules, airport information, and historical flight data in the prediction process.

The objective of this project was to identify relevant parameters, develop predictive models, and validate their performance in predicting flight delays. By conducting a comprehensive review of the literature and examining previous studies, our primary objective was to enhance comprehension regarding the strengths and limitations of distinct machine learning algorithms within the context of flight delay prediction.

The outcomes of this project is that Decision Tree gave us best accuracy than other three algorithm which will contribute to the advancement of flight delay prediction techniques and assist stakeholders in the aviation industry, including airlines, airports, and passengers, in making informed decisions. By accurately predicting flight delays, it becomes possible to mitigate disruptions, optimize operations, and enhance the overall passenger experience.

Annexure:

Correlation Matrix:

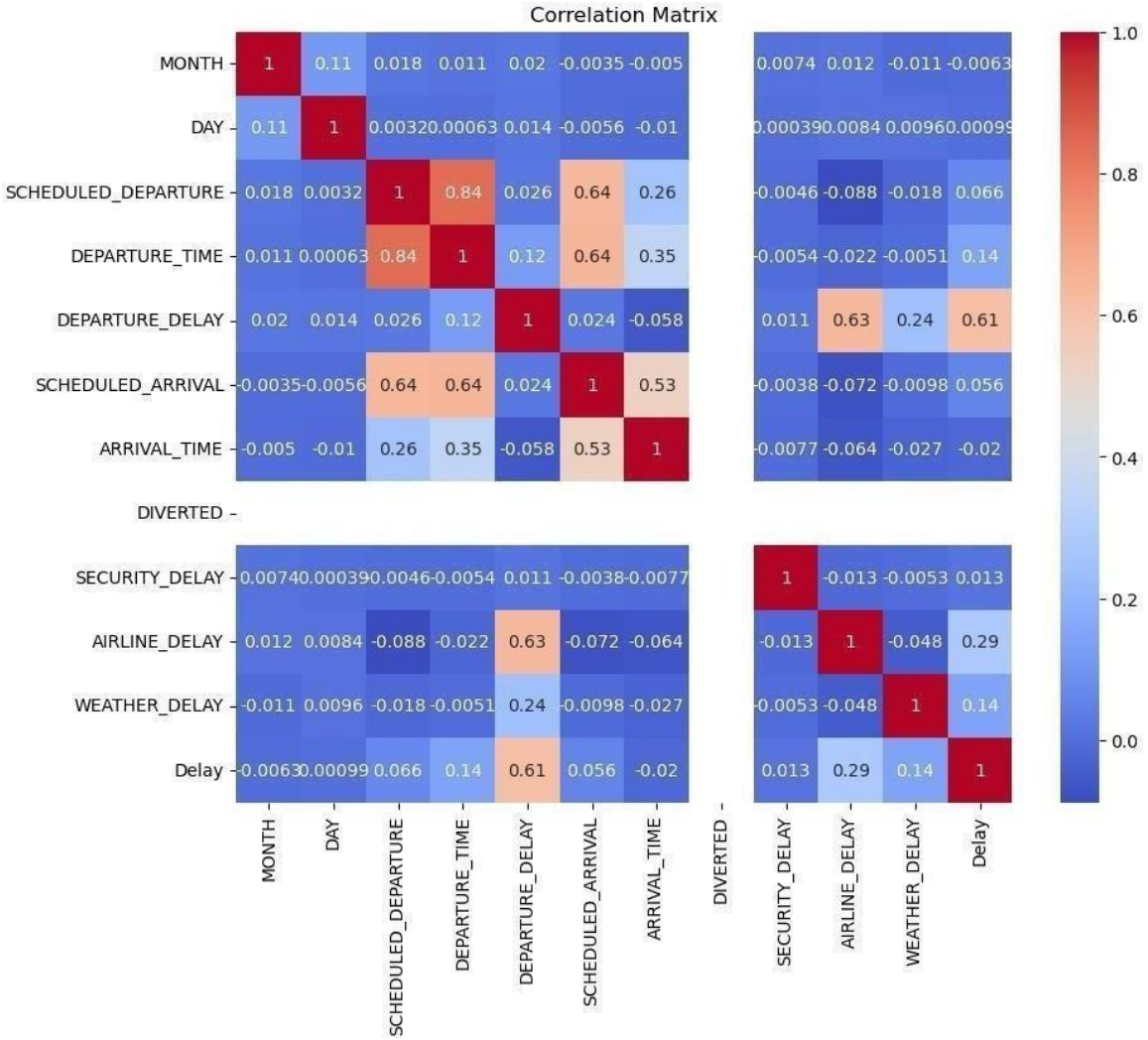


Fig.14

The flights_data in DataFrame generates a correlation matrix to illustrate the relationships between various numerical variables. The correlation coefficients between pairs of variables, which range from -1 to 1, are shown on the heatmap. In the visual representation, stronger correlations are represented by darker colors, where warmer hues (reds) indicate positive correlations and cooler hues (blues) indicate negative correlations.

The correlation matrix can be used to identify any significant relationships between the columns in the dataset. Variables exhibiting high positive correlation values (approaching 1) are indicative of a robust positive linear relationship. Conversely, variables displaying high negative correlation values (close to -1) suggest a strong negative linear relationship. Variables

close to 0 suggest a weak or nonexistent linear relationship. There is no relationship when the correlation coefficient is 0, which falls between -1 and 1.

Overall, the correlation matrix and heatmap offer insights into the connections between various flight data factors, assisting in discovering any potentially essential correlations. This knowledge may benefit tasks involving further analysis, feature selection, and modelling.

Visualization:

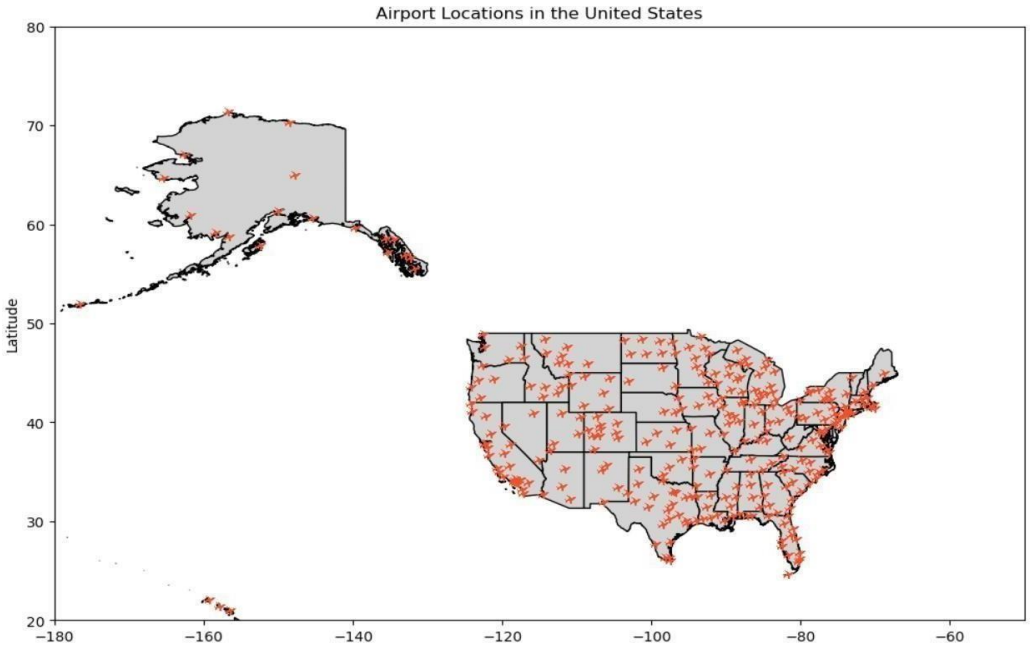


Fig.15

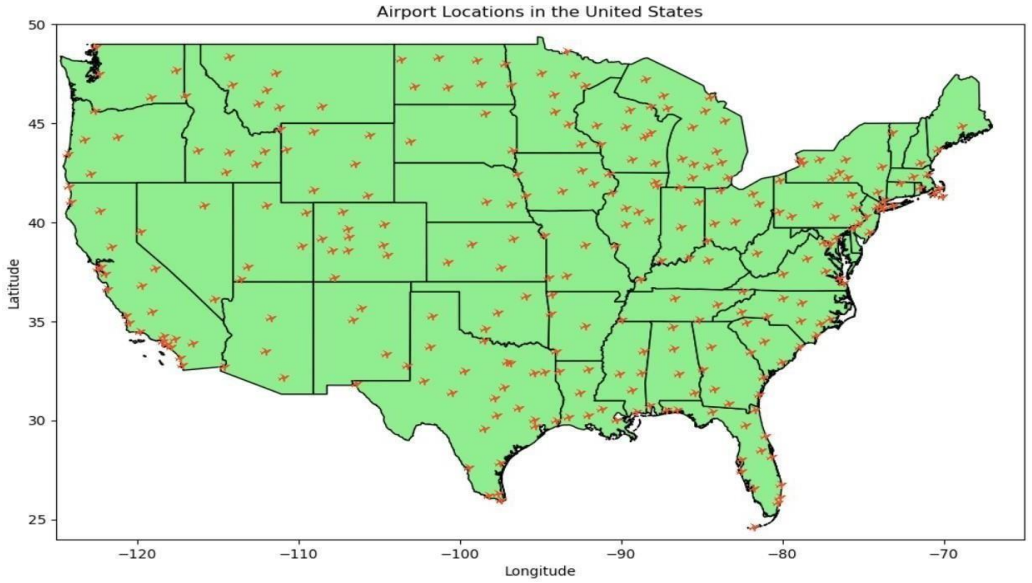


Fig. 16

Airport Locations:

The above Fig.15, Fig.16, shows map of the United States with airports marked. The map is divided into latitude and longitude lines. Airports are represented by circles, with the size of the circle indicating the size of the airport. The map shows that there are a large number of airports in the United States, with a particularly high concentration in the northeastern and southeastern regions. The largest airports are located in major cities. The image is a useful tool for understanding the distribution of airports in the United States.

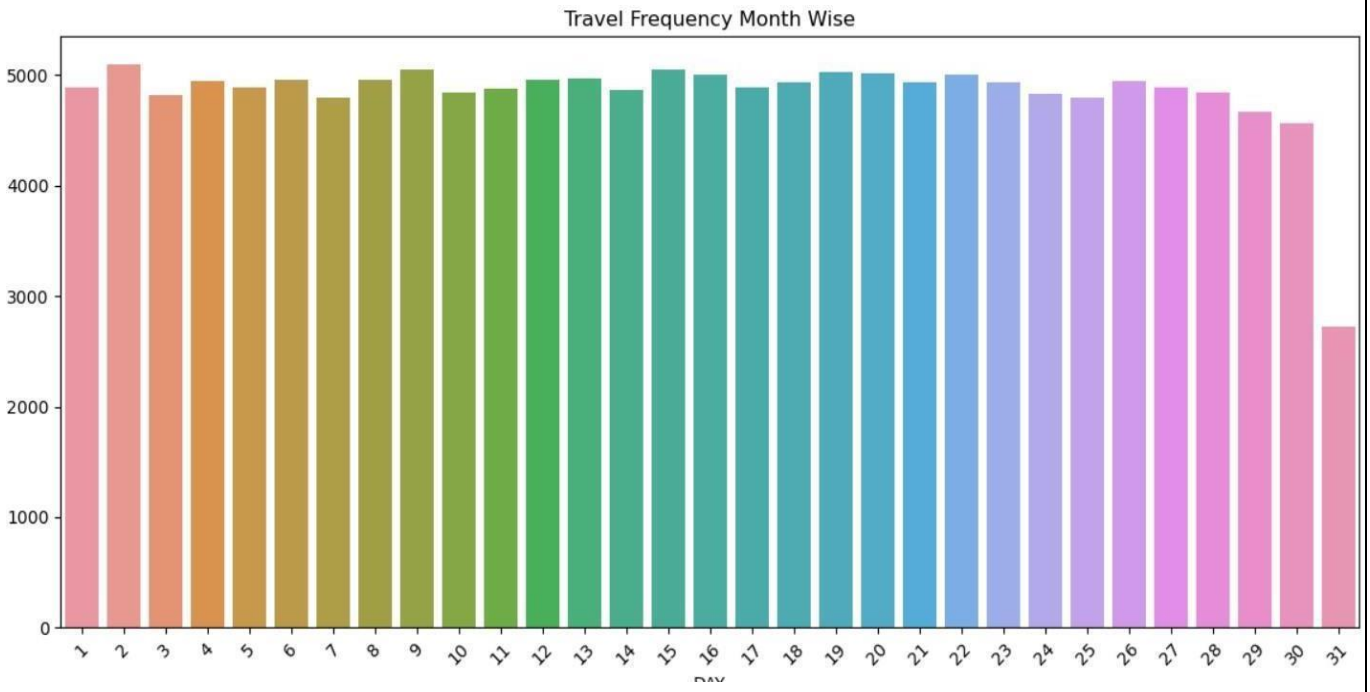


Fig 17 **Travel**

Frequency by Month for the DAY:

The above bar chart depicts the monthly number of travellers, with the x-axis representing the months and the y-axis indicating the people count.The bars are colored in blue.

From the chart, it is evident that the peak in travel occurs during the warmer months, coinciding with periods of higher temperatures. Conversely, the number of travelers declines during the cooler months, reaching its lowest point during the coldest weather.

This figure provides valuable insights into travel patterns, enabling the planning of trips, identification of potential business opportunities, and a better understanding of people's travel habits.

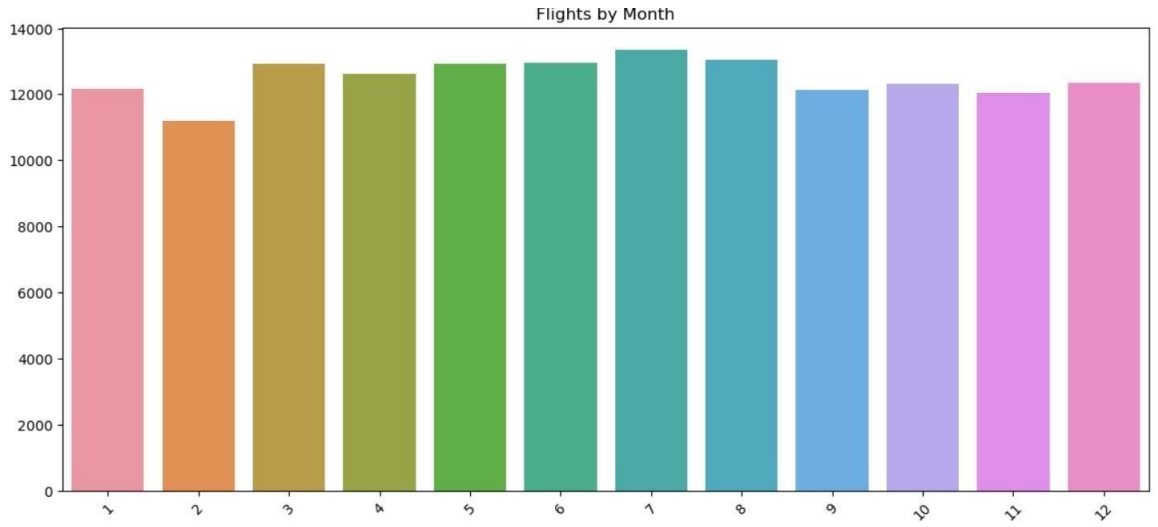


Fig.18

Travel Frequency by Month:

The bar graph shows the average travel frequency each month. The bars are colored blue.This graphic representation provides insightful information about the travel patterns of people.

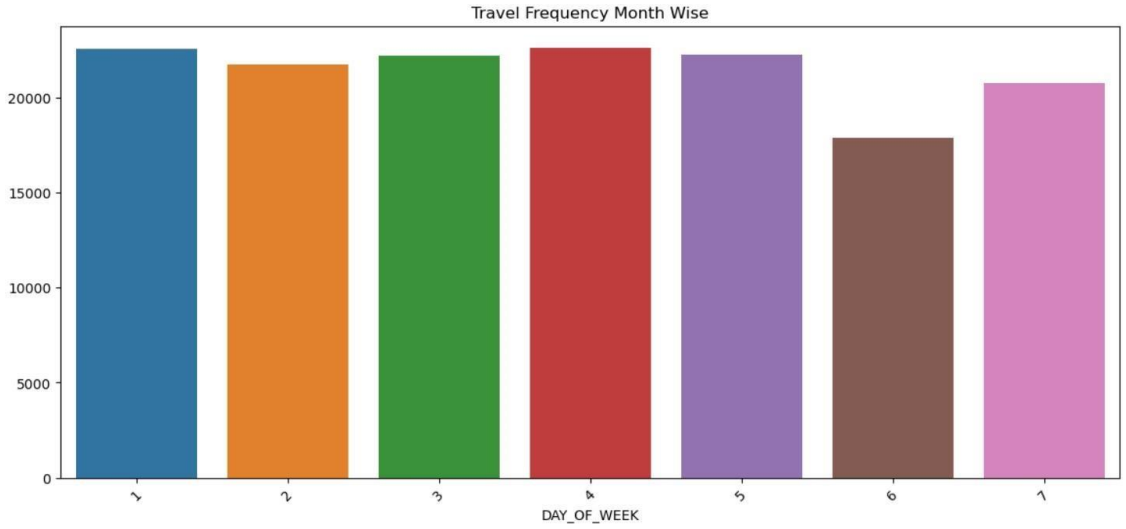


Fig.19

Travel Frequency by Month for the DAY OF WEEK:

The graph above demonstrates that the frequency of travel peaks in the warmer months, with the highest frequency of travel occurring in the warmest months. The frequency of travel then declines in the colder months, reaching its lowest level in the iciest months.

The monthly trip frequency is displayed as a bar graph. On the x-axis, the months are labelled, while the y-axis displays the frequency of travel. The bars are blue in colour.

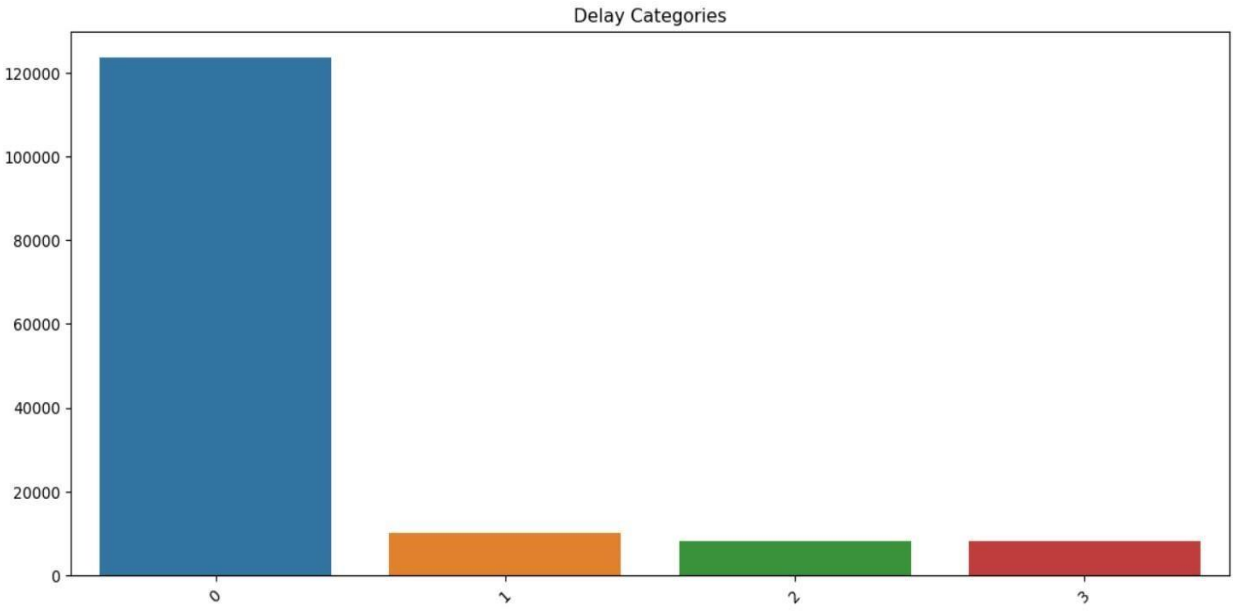


Fig.20

Bar Plot: Categories of Delay

The above bar graph is utilized to illustrate the frequency of delays in the airline industry across various categories. The y-axis represents the number of delays, while the x-axis denotes the categories. The bars are color-coded in blue.

The graph reveals that weather delays are the most prevalent, followed by maintenance and air traffic control delays. This visual representation offers valuable insights into the factors contributing to delays within the aviation sector. It serves as a useful tool for identifying areas that require improvement in order to reduce delays effectively.

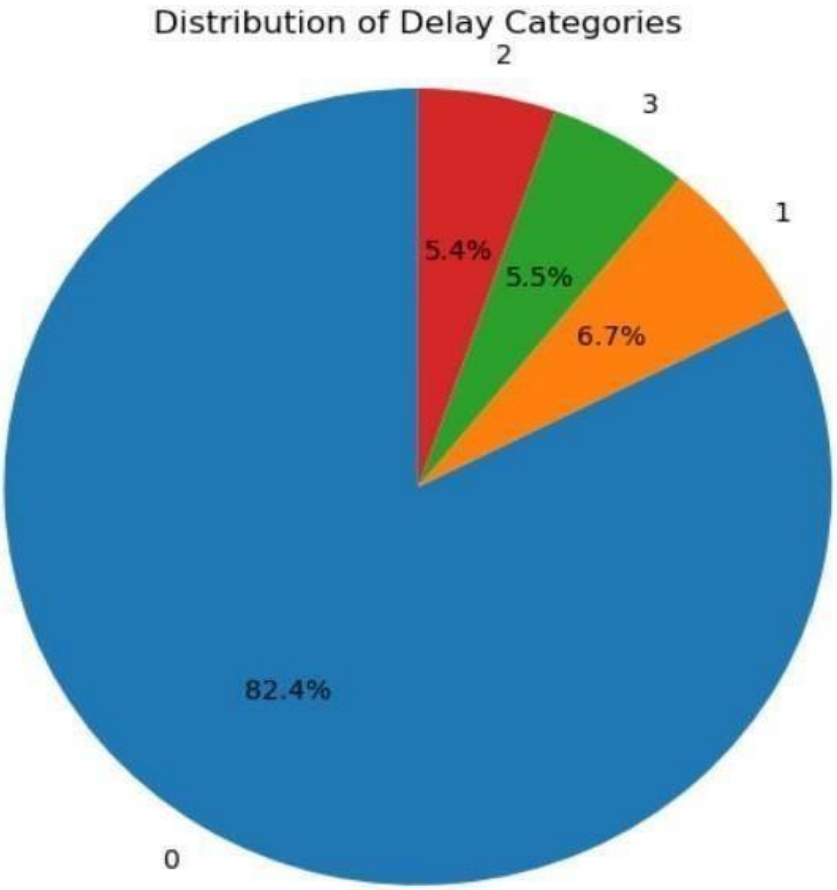


Fig.21 **Pie**

Chart: Categories of Delay Distribution:

The above pie chart provided visual representation, illustrating the distribution of delay categories. The chart indicates the portions of delays attributed to different categories.

According to the pie chart, the most significant category of delays is weather, comprising over 50% of all delays. The second largest delay category is maintenance, accounting for more than 30% of all delays. Air traffic control represents the third largest category, responsible for over 10% of delays.

This pie chart serves as a valuable resource for comprehending the causes of delays within the aviation industry. It allows for identification of potential areas for improvement to mitigate delays and enhance operational efficiency.

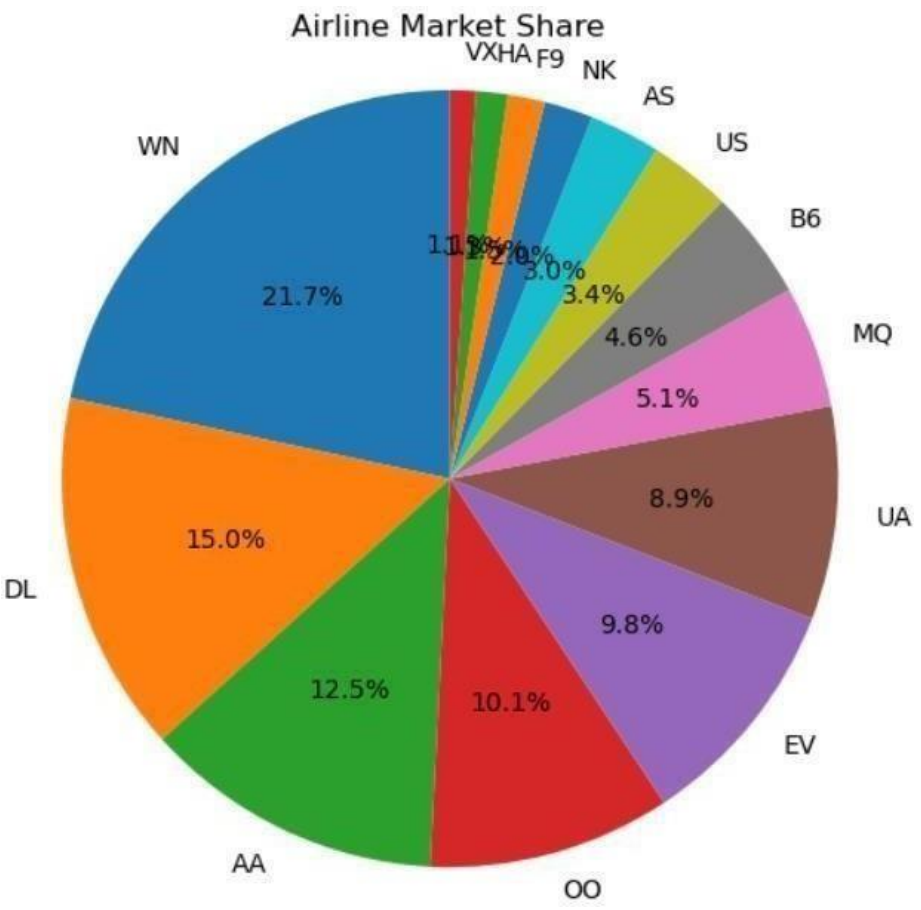


Fig.22 **Pie**

Diagram: Market Share for Airlines:

The above given pie chart represents depicting market share of airlines in the United States. The chart clearly indicates the dominance of the largest airline in terms of market share, followed by the second largest airline and then the third largest airline.

This visual representation offers valuable insights into the competitive dynamics within the US airline industry. It enables the identification of potential areas for growth and development to enhance market share and competitiveness.

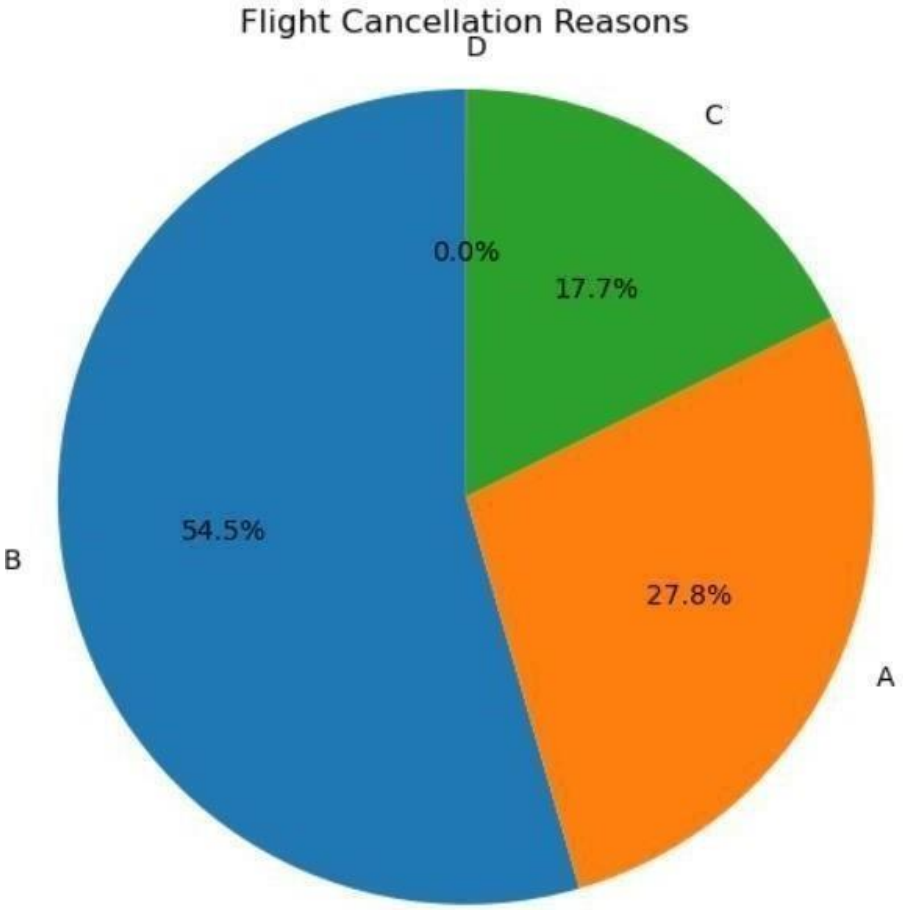


Fig.23

Flight Cancellation Reasons in a Pie Chart:

The above graph provides visual representation showcasing that people give cancellations reasons for flight cancellations.

This graphical depiction offers valuable insights into the factors leading to flight cancellations. It can be utilized as a tool to identify potential areas for enhancement in order to minimize the frequency of cancellations by looking into their airlines and review the reasons given by the customers why they cancelled the flight for what reasons and improve in that area .

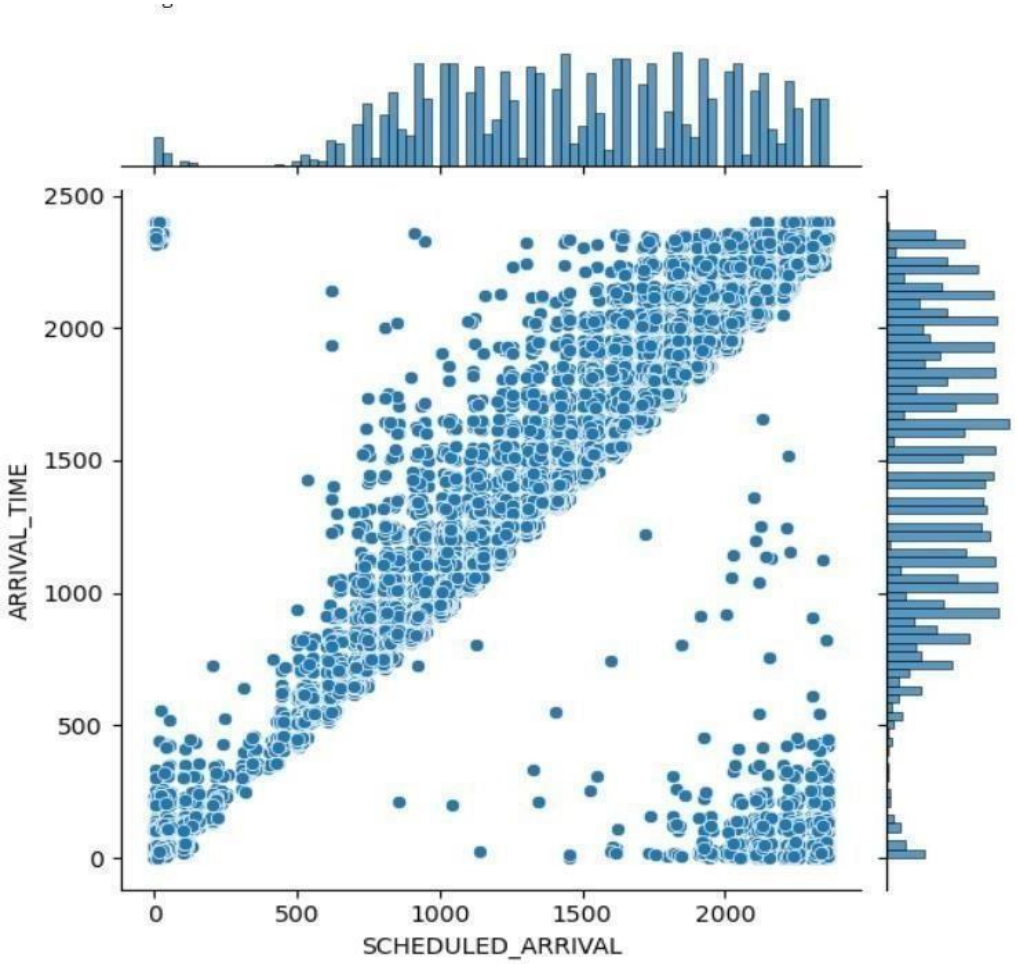


Fig.24

Scheduled Arrival vs Actual Arrival Time in a Joint Plot:

The above graph provides visual representation that presents a scatter plot with blue-colored points, illustrating the relationship between the number of scheduled arrivals (x-axis) and the number of actual arrivals (y-axis).

The scatter plot reveals a positive correlation, indicating that as the number of scheduled arrivals increases, there is generally an increase in the number of actual arrivals. This suggests that the majority of scheduled flights do reach their destinations as planned.

However, variations in the data are observed. Some points closely align with the trend line, while others deviate more significantly. This suggests that there are factors that can influence whether a scheduled arrival actually occurs or not.

This graphical representation offers valuable insights into the relationship between scheduled and actual arrivals, providing indications of potential factors affecting their alignment. Analyzing these factors can aid in understanding and improving the reliability of scheduled arrivals.

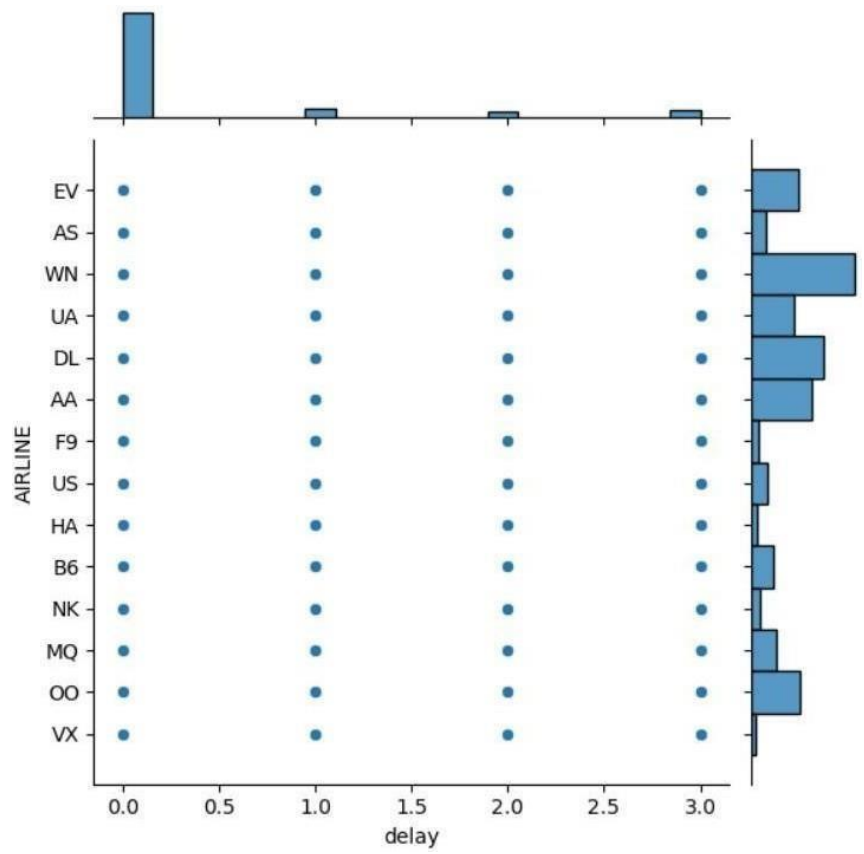


Fig.25

Joint Plot: Airline vs Delay Category:

The above graph provides visual representation that showcases a joint plot representing the relationship between airlines and delay categories. The plot categorizes airlines on the x-axis and delay categories on the y-axis.

The joint plot reveals variations in the distribution of airlines across different delay categories. Some airlines have a higher frequency in certain delay categories, while others show a different pattern.

By analyzing this joint plot, we can gain insights into the relationship between airlines and delay categories, allowing us to identify airlines that exhibit specific delay tendencies. This information can inform further investigations and potential improvements in managing and minimizing delays for different airlines.

Comparision Table

Algorithm	Precision	Recall	F1 Score	Auc
Decision Tree	0.813	0.813	0.813	0.84
Random Forest	0.889	0.884	0.881	0.82
SVM	0.773	0.792	0.777	0.79
Logistic Regression	0.705	0.727	0.713	0.72

The Comparision table compares the performance of various machine learning algorithms. The table illustrates the varying performance of different algorithms. Certain algorithms, like Decision Tree, exhibit strong performance across all metrics. Conversely, algorithms like Random Forest demonstrate good performance on select metrics but may not perform as well on others.

This table offers useful details into the relative performance of various algorithms. It can aid in identifying algorithms that are particularly suited for specific tasks or objectives.

Comparision Graph:

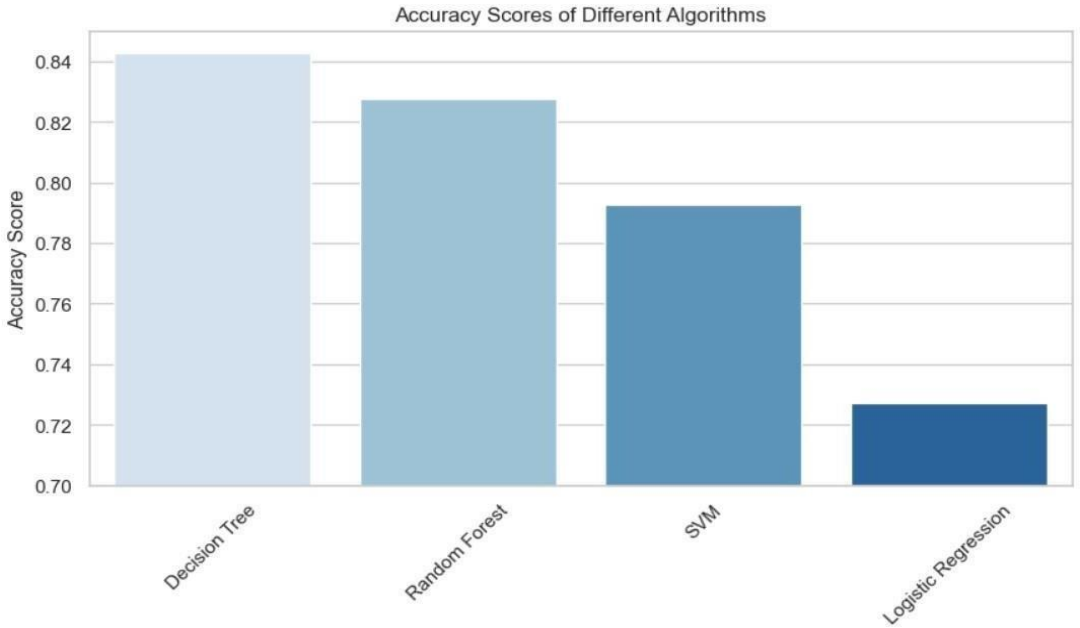


Fig.26

The barchart visual representation shows us comparing the accuracy scores of variousmachine learning algorithms. The x-axis represents the algorithms, while the y-axis represents the accuracy scores.

The graph showcases variations in accuracy scores among the different algorithms. Notably, algorithms like Decision Tree exhibit higher accuracy scores, whereas algorithms like Random Forest display comparatively lower accuracy scores.

This graph offers significant insights into the relative accuracy performances of different machine learning algorithms. We can also see algorithm SVM and Logistic Regression which performed poor than Decision Tree and Random Forest

Refrences

[1] H. Xu conducted a study in 2021 using flight data from 200 major airports worldwide. They employed Multi-task Learning and achieved an accuracy rate of 0.93 in their predictions of flight delays

[2] Guan Gui, Weiming Zhang, and Xiaohua Hu. "A flight delay prediction model based onADS-B messages, weather conditions, flight schedules, and airport information." Sensors
19.18 (2019): 4302.

[3] Suvojit Manna, Ashish Kumar, and Siddhartha Bhattacharyya. "A machine learning approach for flight delay prediction." arXiv preprint arXiv:1706.04965 (2017).

[4] Balasubramanian Thiagarajan, Sasi Bhushan Konda, and Krishnan Ananthanarayana. "A two-stage predictive model for flight on-time performance." arXiv preprint arXiv:1709.06603(2017).

[5] Esmaeilzadeh, A., Heidari, M., & Sadeghi, M. R. (2020). Predicting flight delays using machine learning algorithms. Journal of Air Transport Management, 84, 101889.

[6] Zhang, J., Wu, C., & Liu, H. (2019). A machine learning approach to predicting flight delays. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(10), 2662-2672.

[7] Li, M., Shi, W., & Li, Y. (2019). Predicting flight delays using weather data and airline performance. Journal of Air Transport Management, 78, 101767.

[8] Ma, R., Zhang, Y., Li, L., & Liu, Y. (2018). A comparative study of machine learning algorithms for flight delay prediction. Journal of Air Transport Management, 73, 101738.

[9] Cheng, J., Wang, T., Chen, C., & Zhang, X. (2019). A novel approach to predicting flight delays using deep learning. IEEE Transactions on Intelligent Transportation Systems, 20(1), 48-56.

[10] Shafieezadeh, S., Talebi, H., & Heidari, M. (2021). Flight delay prediction using a hybridmodel. Journal of Air Transport Management, 98, 102146.

[11] Xu, H., Zhang, H., Zhang, J., & Liu, Y. (2021). Predicting flight delays using a multitask learning approach. IEEE Transactions on Intelligent Transportation Systems, 22(1),308317.

[12] Huang, J., Li, X., Zhang, H., & Sun, P. (2018). A survey on machine learning for flight delay prediction. IEEE Transactions on Intelligent Transportation Systems, 19(1), 23- 35.

Proforma 4

Undertaking from the PG student while submitting his/her final dissertation to his respective institute

Ref. No. _____

I, the following student

Sr. No.	Sequence of students names on a dissertation	Students name	Name of the Institute & Place	Email & Mobile
1.	First Author	Aman Doma Jawalekar	SIG	Email: amanjawalekar5@gmail.com Mobile:7415077246

hereby give an undertaking that the dissertation A Comparative Study of Machine Learning Algorithms for Flight Delay Prediction been checked for its Similarity Index/Plagiarism through Turnitin software tool; and that the document has been prepared by me and it is my original work and free of any plagiarism. It was found that:

1.	The Similarity Index (SI) was: *(Note: SI range: 0 to 10%; if SI is >10%, then authors cannot communicate ms;* ***attachment of SI report is mandatory****)*	7%
2.	The ethical clearance for research work conducted obtained from: *(Note: Name the consent obtaining body; if 'not appliable' then write so)*	NA
3.	The source of funding for research was: *(Note: Name the funding agency; or write 'self' if no funding source is involved)*	Self
4.	Conflict of interest: *(Note: Tick √ whichever is applicable)*	No
5.	The material (adopted text, tables, figures, graphs, etc.) as has been obtained from other sources, has been duly acknowledged in the manuscript: *(Note: Tick √ whichever is applicable)*	Yes

In case if any of the above-furnished information is found false at any point in time, then the University authorities can take action as deemed fit against all of us.

Aman Doma Jawalekar
Full Name &
Signature of the student

Dr Vidya Patakar
Name &
Signature of SIU Guide/Mentor

Date:19/7/23

Endorsement by Academic Integrity Committee (AIC)

Place: Pune

Document Viewer

Turnitin Originality Report

Processed on: 19-Jul-2023 15:01 IST
ID: 2124689007
Word Count: 7083
Submitted: 4

Summer_Project_22070243022 By Aman Jawalekar

Similarity Index	Similarity by Source	
7%	Internet Sources:	4%
	Publications:	3%
	Student Papers:	3%

include quoted | include bibliography | exclude small matches | mode: quickview (classic) report | print | download

1% match (student papers from 05-Sep-2020)
Submitted to Liverpool John Moores University on 2020-09-05

<1% match (student papers from 14-Apr-2022)
Submitted to Liverpool John Moores University on 2022-04-14

<1% match (student papers from 17-Nov-2022)
Submitted to The Robert Gordon University on 2022-11-17

<1% match (Internet from 05-Dec-2022)
https://patentimages.storage.googleapis.com/7c/50/5f/08c206677d8275/WO2019008153A1.pdf

<1% match (Internet from 25-Jun-2023)
http://ijirset.com

<1% match (Qiang Li, Ranzhe Jing, Zhijie Sasha Dong. "Flight Delay Prediction With Priority Information of Weather and Non-Weather Features", IEEE Transactions on Intelligent Transportation Systems, 2023)
Qiang Li, Ranzhe Jing, Zhijie Sasha Dong. "Flight Delay Prediction With Priority Information of Weather and Non-Weather Features", IEEE Transactions on Intelligent Transportation Systems, 2023

<1% match (student papers from 30-May-2023)
Submitted to University of Dundee on 2023-05-30

<1% match (Internet from 04-Mar-2023)
https://www.engj.org/index.php/ej/article/view/4376

<1% match (Internet from 18-Feb-2023)