

ANALYSIS OF FLIGHT DELAY DATA

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering

By

ABHISHEK RAJ (Reg.No - 40110019)

ABHINAV KUMAR (Reg.No - 40110011)



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Category-1 University by UGC

Accredited with Grade "A++" by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

APRIL - 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category - I University by UGC

Accredited "A++" by NAAC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Abhishek Raj** (Reg.No:40110019) and **Abhinav Kumar**(Reg.No:40110011) who carried out the project entitled "**Analysis of Flight Delay Data**" under my supervision from November 2023 to April 2024.

Internal Guide

Dr. A. PRAVIN, M.E., Ph.D.

HEAD OF THE DEPARTMENT
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Head of the Department
Jeppiaar Nagar, Ram Gandhi Salai,
Chennai - 600 119.

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on 15/04/24

15/4/24
Internal Examiner
15/4/24
External Examiner

DECLARATION

I, **Abhishek Raj (Reg.No-40110019)** hereby declare that the Project Report entitled "**ANALYSIS OF FLIGHT DELAY DATA**" done by me under the guidance of **Dr. A. PRAVIN ,M.E. ,Ph.D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 10/4/24

Abhishek Raj

PLACE: Chennai

SIGNATURE OF THE CANDIDATE



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Category - I University by UGC

Accredited "A++" by NAAC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Abhishek Raj** (Reg.No:40110019) and **Abhinav Kumar**(Reg.No:40110011) who carried out the project entitled "**Analysis of Flight Delay Data**" under my supervision from November 2023 to April 2024.

Internal Guide

Dr. A. PRAVIN, M.E., Ph.D.

HEAD OF THE DEPARTMENT
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Head of the Department
Department of Computer Science and Engineering,
Chennai - 600 119.

Dr. L. LAKSHMANAN, M.E., Ph.D.

Submitted for Viva voce Examination held on

15/04/24

15/4/24
Internal Examiner

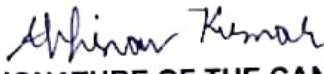
15/4/24
External Examiner

DECLARATION

I, **Abhinav Kumar (Reg.No-40110011)** hereby declare that the Project Report entitled "**ANALYSIS OF FLIGHT DELAY DATA**" done by me under the guidance of **Dr. A. PRAVIN ,M.E. ,Ph.D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering**.

DATE: 10/04/2024

PLACE: Chennai


SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **The Board of Management of Sathyabama Institute of Science and Technology** for their encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D., Dean**, School of Computing, **Dr. L. Lakshmanan M.E., Ph.D.**, Head of the Department of Computer Science and Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. A. Pravin, M.E, Ph.D.**, for his valuable guidance, suggestions and constant encouragement paved the way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

The aviation sector is significantly impacted by flight delays., leading to operational inefficiencies, customer dissatisfaction, and economic losses. This presents a comprehensive overview of flight delay data analysis, highlighting its importance in identifying patterns, predicting delays, and improving airline operations. The goal of the dashboards is to visualize which airlines have the most delays and potential causes. Accurately anticipating these flight delays helps airlines address probable causes of delays ahead of time, minimizing the detrimental effects and enabling customers to be well-prepared for the disruption to their travel plans. The aviation industry, marked by its dynamic and time-sensitive nature, constantly strives to enhance operational efficiency, improve passenger satisfaction, and address the persistent issue of flight delays. The prediction of flight delay can be considered as one of the most challenging problems to solve. Delay of an aircraft is not only a problem for an airline but also for the passengers. Flights can be delayed due to several reasons, the weather being the primary. Our focus is to predict the delay of the flights due to bad weather. Overall, this study contributes to the body of knowledge on flight delay management by offering a comprehensive analysis of delay data and proposing evidence-based strategies for minimizing disruptions in air travel. To elucidate the intricate relationships between delay factors, flight characteristics, and temporal variables, advanced statistical methods including regression analysis, time series modeling, and machine learning algorithms are employed. By dissecting the data at granular levels, the study elucidates the impact of factors such as flight route, time of day, aircraft type, and airport facilities on delay occurrences. Moreover, the research extends its analysis beyond immediate operational factors to consider broader systemic influences on flight delays. This includes investigating the role of air traffic management systems, airport capacity constraints, regulatory frameworks, and socio-economic factors in shaping delay patterns. Ultimately, this study contributes valuable insights to the field of aviation management by offering a nuanced understanding of flight delay dynamics and proposing evidence-based approaches.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	v
	LIST OF TABLES	viii
	LIST OF FIGURES	ix
	LIST OF ABBREVIATIONS	x
1	INTRODUCTION	1
	1.1 Introduction	1
2	LITERATURE SURVEY	4
	2.1 Literature Survey	4
	2.2 Inferences from Literature Survey	8
	2.3 Open problems in Existing System	9
3	REQUIREMENT ANALYSIS	11
	3.1 Feasibility Studies/Risk Analysis of the Project	13
	3.2 Software Requirements Specification Document	15
4	DESCRIPTION OF PROPOSED SYSTEM	16
	4.1 Datasets	17
	4.2 Selected Methodology/Process Model	19
	4.3 Architecture/Overall design of Proposed System	21
	4.4 Description of Software for Implementation and Testing Plan of the Proposed Model/System	22
	4.5 Project Management Plan	23
5	IMPLEMENTATION DETAILS	26
	5.1 Development and Deployment Setup	26
	5.2 Algorithms	27

5.2.1	Random Forest	27
5.2.2	Decision Tree	29
5.3	Testing	33
6	RESULTS AND DISCUSSIONS	34
6.1	Results	34
6.2	Discussions	34
6.3	Performance Analysis	36
7	CONCLUSION	38
7.1	Summary	38
7.2	Conclusion	38
7.3	Future Enhancements	39
	REFERENCES	41
	APPENDIX	43
A	SOURCE CODE	43
B	SCREENSHOTS	52
C	RESEARCH PAPER	54

LIST OF TABLES

TABLE NO.	TABLE NAME	PAGE NO.
5.1	Flight table	31
5.2	Airport table	32
5.3	Airline table	32

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
4.1	System Architecture for Analysis of flight delay data	21
5.1	Pie chart for Airline	33
5.2	Heatmap	33
5.3	Distribution of Air Time	34
5.4	Distribution of Elapsed Time	34
B.1	Taxi Out	54
B.2	Taxi Time	55

LIST OF ABBREVIATIONS

ABBREVIATION	EXPANSION
SVM	Support Vector Machine
KNN	K-Nearest Neighbors
LGBM	Light Gradient-Boosting Machine
NAS	National Airspace System
TP	True Positive
TN	True Negative
FN	False Negative
FP	False Positive

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

As more and more individuals opt to travel by air, an increasing number of flights are missing their planned departure schedules. This development worsens airport traffic and jeopardizes the financial stability of the airline industry. Travel delays are an indicator of aviation industry inefficiencies. The cost to airline companies and their passengers is high. As more and more individuals opt to travel by air, an increasing number of flights are missing their pla

Comparing how well machine learning classification methods perform in forecasting flight delays is the main aim of the study. John F. Kennedy International Airport, New York City served as the study's airport. Data pertaining to flights departing from JFK airport during a one-year period was being examined. The whole cost of air transportation delays travel, based on the Total Delay Impact Study's assessment.

One of the most important forms of contemporary adaptability is the air transportation system. It's critical to preserve perseverance and resilience in the face of growing passenger and aviation traffic congestion. The resources and land availability support airport infrastructure. The rules advancing technology and practice are designed to preserve things like capacity, safety, and efficiency.

As a consequence, the "National Airspace System" (NAS) prioritizes reducing the negative impact that improvisation has on the environment. Passengers may see their flight route, direction, altitude, and other relevant information while traveling thanks to contemporary technology. Nonetheless, air traffic controllers are always attempting to minimize Aeroplane arrival and departure delays.

Even though they worked in phases, the results are not what they wanted since there were hours-long delays that sometimes resulted in turmoil. A few significant

factors that contribute to delays include carrier, weather, maintenance, and security. Travel for business purposes and tourism are the two main drivers of the aviation industry. which, by 2030, is anticipated to quadruple. Air traffic is predicted to rise in the same multiple as a consequence of this increase. It is possible to build additional airports in order to reduce air traffic congestion. But the level of complexity keeps rising rapidly.

Therefore, modifying the current airports is the only method to reduce the delay. In light of the restricted quantity of land resources available, the latter option makes more sense. In essence, delay is the amount of time that an aircraft is delayed.

One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings is an upscale affair for the airlines, making issues in scheduling and operations for the end users therefore inflicting unhealthy name and client discontent.

As we all know that we have a tendency to not get the flight delay before departure as customers of the Airline Company neither the airline company's ground staff gets the airline delay prediction supported varied conditions. However, we all know that one in all the biggest reasons for delays in flights is the weather.

The models predicting flight delays have increased significantly in recent years, and significant progress has been made regarding feature selection and the use of machine learning algorithms. However, most of these studies have used small sample datasets along with small-scale models to predict delays. We believe that the potential for deploying machine learning models to predict arrival delays for multiple airports has not yet been fulfilled.

The machine learning models with a rich dataset of multiple origin and destination airports, including features related to air traffic data and weather. We also contribute by introducing mechanisms to tackle imbalanced data in the prediction workflow when doing this comparison.

In the realm of air travel, few things disrupt meticulously planned itineraries and cause passenger aggravation quite like flight delays. These delays not only cause personal inconvenience but also translate to financial losses for airlines and potentially missed business opportunities for travelers. To combat this persistent issue, a deeper understanding of the root causes behind flight delays is paramount.

This analysis embarks on a comprehensive investigation into a meticulously curated dataset of flight information. By meticulously dissecting this data, we aim to unveil the secrets behind these delays. We will embark on a multi-faceted exploration, examining not only the prevalence and severity of delays, but also the influence of various contributing factors. These factors encompass the airline itself, the origin and destination airports, the prevailing weather conditions on the day of travel, and even the specific day of the week. Through meticulous analysis, we hope to unearth any underlying patterns or trends that might be influencing these delays.

The ultimate goal of this exploration is to leverage the power of data analysis to achieve several key objectives. Ideally, this analysis will pave the way for the development of more accurate flight delay prediction models. Armed with such models, airlines and passengers alike can be better prepared for potential disruptions. Additionally, this analysis can serve as a springboard for formulating proactive strategies to minimize delays and improve on-time performance. Finally, by gleaning insights into the root causes of delays, airlines can enhance communication with passengers during disruptions, managing expectations and minimizing the overall frustration associated with delayed flights.

The modern world thrives on efficiency, and air travel is no exception. Yet, a persistent thorn in the side of this fast-paced industry is the ever-present threat of flight delays. These delays cause a domino effect of frustrations – missed connections, disrupted schedules, and a general sense of powerlessness for both passengers and airlines. The financial impact is significant as well, with airlines hemorrhaging money due to operational disruptions and passengers incurring additional expenses.

CHAPTER 2

LITERATURE SURVEY

2.1 LITERATURE SURVEY

1. Fujun Wang, Jun Bi, Dongfan Xie, Xiaomei Zhao, Quick taxi route assignment via real-time intersection state prediction with a spatial-temporal graph neural network, 2024

This article provides a perceptive examination of European and Western nations' aircraft delay statistics. Data scientists combine domain knowledge with scientific skills in data learning to investigate various data aspects related to aircrafts like sequential (month, season, time of day, day of week), scheduling (flight timetable, flight preparation), spatial (province airfield, urban), climate (upper limit, Visibility, Surface climate, Convective climate), operations (claim, capacity), features (Airport transportation, Airline Status, plane tenancy, plane model, price) .

2. T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices" ,2014

For the San Francisco–New York route, a desired model is created implementing a Linear Quantile Blended Regression technique, in which daily airfares are made available online. Two criteria are considered by the model: the total travel days and journey percentage which takes place both on weekdays and weekends.

3. Peng Hu, Jia Hou, Jianping Zhang, Qingyuan Yang, Lei Cao, Liwei Wang, Research on flight arrival delay prediction based on support vector machine, 2022

This research analyzes flight data to develop a model for forecasting aircraft arrival delays using flight plans, based on support vector machines. It builds and chooses feature vectors, such as “planned arrival time period, planned arrival in peak time period, planned flight time, planned departure time in day or night, the planned departure time in peak time period, and the passenger or cargo flight, by selecting one day's domestic flight plan and flight operation data. Next, it examines the information on the flight type, actual arrival time, and departure and arrival times

on the flight plan (passenger flight or cargo flight).” Furthermore, a prediction model is established.

4. K. Wang, J. Li and Y. Tian, "Airport Delay Prediction Method based on Improved Weather Impacted Traffic Index," 2019

It makes use of a machine learning method known as logistic regression to forecast Aeroplane delays. Using a variety of independent characteristics, this method trains a model to predict whether or not an aircraft would experience a delay. We used the “Microsoft Azure Learning Studio” platform to develop the algorithm. To ascertain the effect of weather conditions on delays of flights and improve the prediction for real-world situations, we additionally used a weather dataset and connected it with the airport information at the corresponding locations. Thirty percent of the dataset is used for testing after the model has been trained on seventy percent of it.

5. V. Natarajan, S. Meenakshisundaram, G. Balasubramanian and S. Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018

The eight attributes of the airline dataset were as follows: source, destination, number of stoppages, airline type, departure and arrival times, and other data. Regression machine learning models, like LGBM Regressor, Decision Tree Regressor, Random Forest Regression Tree, were used by the authors to conduct prediction. Random Forest Regression and Supervised Learning algorithm performs regression using ensemble learning techniques. The ensemble learning technique combines predictions from many machine learning algorithms to provide a forecast that is more accurate than one from a single model.

6. G. Ataman and S. Kahraman, "Flight Prediction in Using Linear Regression and Artificial Neural Network Hybrid Models," The Singapore Economic Review, 2021

Social media and deep learning data-driven airline ticket pricing prediction algorithms are shown here. The variables influencing ticket prices are discussed by the authors together with the state of airline ticket pricing at the moment. A Random Forest constructs several decision trees in the training phase, with each tree's prediction derived from the mean of the classes. The Random Forest

Regression is a powerful and reliable regression model. It often performs well on a variety of problems, especially those containing attributes with non-linear correlations.

7. J. C. Lok, Prediction Factors Influence Airline Fuel Price Changing Reasons, 2018.

Airlines need to be able to predict airfare trends down to the market segment level in order to modify their plans and allocation of resources for a certain route. It may not always be the case, but the scientific literature on business sector price prediction is predicated on the idea that the selected factors have a linear relationship. This is because the research uses biased conventional predictive approaches, such as linear regression.

8. T. M. Vowles, "Airfare pricing determinants in hub-to-hub markets," Journal of Transport Geography, 2006.

It suggested a method to estimate "the average ticket price based on input and output pairs by integrating two databases, including macroeconomic data, and using machine learning techniques like support vector machines and XGBoost. With the latest R-squared performance measures, the framework attains a remarkable 0.869 accuracy" .

9. K. Rama-Murthy, "Modeling of United States airline fares—using the official airline guide (OAG) and airline origin and destination survey (DB1B)," 2006.

It demonstrates that customers may have an operator presented to them who will make the most of their buying time. Using the partial least squares method, a model is built. In her survey report Supriya Rajankar made use of a small dataset of flights between Bombay and Delhi to anticipate Aeroplane fares using the ML algorithms. The methods used include linear regression, Support Vector Machine (SVM) and K- Nearest-Neighbors (KNN).

10. Tianyi wang, Samira Pouyanfar, Haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach", 2022

The ticket pricing plan has developed into a sophisticated set of laws, like numerical simulations that decide price marketing methods, since the airline firm

was privatized. Studies have depicted that these principles are affected by different conditions, even if the majority of their details are still unknown.

11. William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the international conference on autonomous agents and multi-agent systems,2013.

In order to estimate the average ticket price based on source and destination pairings,they suggested a framework in this study that combines two databases with macroeconomic data and machine learning algos like XGBoost and support vector machines.

12. J. Stavins, "Price discrimination in the airline market: The effect of market concentration," Review of Economics and Statistics,2001.

This article offers a framework for examining various routes and/or carriers as well as broad trends in airline pricing behavior. By providing customers with the essential information they need to choose the best time to purchase a ticket, the aim is to balance the desire to save money with any time limitations the customer may have.

13. B. Mantin and B. Koo, "Dynamic price dispersion in airline markets," Transportation Research Part E: Logistics and Transportation Review,2009.

They conducted research to test the theory that as airline market competition rises,pricing discrimination would also rise. They estimated price discrimination using marginal implicit pricing of ticket limits, which carriers usually utilize to price discriminate, using a wide cross section of tickets supplied by many airlines on different routes.

14. P. Malighetti, S. Paleari, and R. Redondi, "Has Ryanair's pricing strategy changed over time? an empirical analysis of its 2006–2007 flights," Tourism Management, 2010

This paper examines the mechanisms behind the dynamic price dispersion, or daily airfare fluctuations across ticket histories. Empirical analyses reveal that factors pertaining to demand characteristics, such as income, population, and the proportion of business passengers, and competitive pressures resulting from the

presence of low-cost carriers, have a significant impact on dynamic price dispersion. However, competition intensity has no such effect.

15.Zhen Guo, Bin Yu, Mengyan Hao, Wensi Wang,” A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient”

Regarding the instance of Ryanair, the low-cost leader in Europe, this chapter makes reference to two primary fields of literature on aviation: the analysis of the low-cost business model and the research of dynamic pricing approaches. This chapter seeks to identify the competitive and contextual elements, as well as the key components of Ryanair’s business model, within this framework. This chapter seeks to identify the competitive and contextual elements, as well as the key components of Ryanair’s business model, within this framework.

2.2 INFERENCES FROM LITERATURE SURVEY

The literature survey indicates a rigorous and multifaceted approach to analyzing aircraft delay statistics, combining both qualitative and quantitative methods. The focus on a specific route suggests a tailored understanding of factors influencing delays and airfare dynamics in that context.

The development of a predictive model underscores the practical implications of the research, potentially benefiting travelers by providing real-time air predictions. The consideration of temporal factors and travel patterns in the model highlights the nuanced nature of pricing dynamics in the aviation industry.

The literature survey reflects a comprehensive exploration of predictive modeling techniques in the aviation domain, ranging from flight delay prediction to airline ticket pricing. There is a strong emphasis on leveraging machine learning algorithms, such as SVM, logistic regression, and ensemble methods, to analyze and predict various aspects of flight operations.

The incorporation of weather data indicates an acknowledgment of external factors that can impact flight delays, highlighting the importance of considering both

internal and external variables in predictive modeling. The use of practical platforms like Microsoft Azure Learning Studio suggests a focus on real-world implementation and scalability of the developed algorithms. Overall, the literature survey demonstrates a concerted effort to harness data-driven approaches for improving operational efficiency and decision-making in the aviation industry.

2.3 OPEN PROBLEMS IN EXISTING SYSTEM

Based on the literature survey , some open problems in the existing system include:

Limited Scope of Features: The features used for prediction in both the flight delay and ticket pricing models appear to be somewhat limited. There may be additional relevant factors not considered in the analysis, such as airport infrastructure, air traffic control operations, or geopolitical events, which could significantly impact flight delays and ticket prices.

Model Generalization: While various machine learning algorithms are explored for prediction tasks, there might be challenges in generalizing these models across different geographical regions or airlines. The models developed may be specific to certain routes, airlines, or time periods, limiting their applicability in diverse contexts.

Weather Data Integration: Although weather data is incorporated into the flight delay prediction model, there may be room for improvement in how this data is integrated and utilized. Weather conditions can be highly variable and complex, and more sophisticated methods may be needed to accurately capture their impact on flight operations.

Data Quality and Availability: The effectiveness of predictive models relies heavily on the quality and availability of data. Issues such as missing or erroneous data, inconsistent data formats, or limited historical data could undermine the accuracy and reliability of the models developed.

Real-time Prediction: While the models aim to predict flight delays and ticket prices, there may be challenges in providing real-time predictions that account for rapidly changing conditions. Integrating real-time data streams and developing algorithms capable of quickly adapting to new information could enhance the practical utility of these models.

Interpretability and Explainability: Some of the machine learning models used, such as Random Forest Regression, may lack interpretability, making it difficult to understand the underlying factors driving predictions. Improving the interpretability and explainability of models could increase trust and usability, particularly in decision-making contexts.

Ethical Considerations: There may be ethical implications associated with the use of predictive models in the aviation industry, particularly regarding issues such as fairness, bias, and privacy. Ensuring that models are developed and deployed in a responsible manner, with appropriate safeguards in place.

CHAPTER 3

REQUIREMENT ANALYSIS

Requirement analysis for the analysis of flight delay data involves identifying the necessary components, resources, and procedures to effectively conduct the study. Here's a breakdown of key requirements:

1. Data Sources:

Flight Data: Comprehensive records of flight schedules, departures, arrivals, and delays are essential. This includes information on flight numbers, departure and arrival times, origin and destination airports, aircraft type, and carrier details.

Weather Data: Historical weather data for each airport and flight route, including factors such as temperature, precipitation, wind speed, and visibility, is crucial for assessing weather-related delays.

Air Traffic Data: Data on air traffic density, congestion, and routing patterns provide insights into airspace utilization and potential causes of delays.

Airport Infrastructure Data: Information on airport facilities, runway configurations, capacity constraints, and operational procedures help understand the impact of airport infrastructure on delay occurrences.

Aircraft Maintenance Records: Details on aircraft maintenance schedules, repairs, and downtime contribute to analyzing delays stemming from mechanical issues.

2. Data Quality and Consistency:

Ensuring the accuracy, completeness, and consistency of the data sources is paramount. This involves data cleaning, validation, and reconciliation to address errors, missing values, and discrepancies. Standardization of data formats and terminology across different sources facilitates seamless integration and analysis.

3. Analytical Tools and Techniques:

Statistical Software: Utilizing statistical analysis software such as R, Python (with libraries like Pandas, NumPy, and SciPy), or specialized tools for data manipulation, visualization, and modeling. Machine Learning Algorithms: Employing machine learning techniques, including regression analysis, time series forecasting, classification, and clustering, to identify patterns and predict delay occurrences.

4. Regulatory Compliance and Ethical Considerations:

Adhering to data privacy regulations and ethical guidelines when accessing and handling sensitive information, such as passenger data or proprietary airline records. Ensuring data security measures are in place to safeguard against unauthorized access, breaches, or data misuse.

5. Documentation and Reporting:

Documenting data sources, methodologies, assumptions, and limitations to ensure transparency and reproducibility of the analysis. Generating comprehensive reports or presentations summarizing the findings, insights, and recommendations for stakeholders and decision-makers.

Identifying Mitigating Factors: Through comprehensive analysis, the project aims to identify potential strategies and interventions for mitigating flight delays. This could involve implementing new technologies, optimizing scheduling and routing procedures, improving communication and coordination among stakeholders, or investing in infrastructure upgrades. Overall, the aim of the analysis of the flight delay data project is to leverage data-driven approaches to enhance the reliability, efficiency, and overall performance of air travel systems. By gaining a deeper understanding of the factors contributing to delays and implementing targeted interventions, the aviation industry can work towards minimizing disruptions.

3.1 FEASIBILITY STUDIES/RISK ANALYSIS OF THE PROJECT

Feasibility studies and risk analysis are crucial components of any project, including one focused on flight delay prediction. Here's an outline of what such studies might entail. By conducting comprehensive feasibility studies and risk analysis, project stakeholders can better understand the opportunities and challenges associated with flight delay prediction and develop strategies to mitigate risks and maximize project success.

The keys involved in the feasibility analysis are:

- Technical feasibility
- Operational feasibility
- Financial feasibility
- Legal and Ethical Feasibility
- Model performance
- Risk Analysis of the project

1. Technical feasibility

Assess the availability and quality of data required for flight delay prediction, including historical flight data, weather data, airport congestion data, etc. Evaluate the feasibility of implementing machine learning algorithms and predictive models to analyze and predict flight delays. Determine the computational resources and infrastructure needed to process and analyze large volumes of data in real-time.

2. Operational feasibility

Analyze the operational workflow of integrating flight delay prediction into existing airline and airport systems. Identify potential challenges or barriers to adoption, such as resistance from stakeholders or regulatory constraints. Assess the feasibility of deploying predictive models in a real-world operational environment, considering factors like data availability, model maintenance, and scalability.

3. Financial feasibility

Estimate the costs associated with data acquisition, model development, infrastructure setup, and ongoing maintenance. Conduct a cost-benefit analysis to determine the potential return on investment (ROI) of implementing flight delay prediction systems, considering factors such as improved operational efficiency, reduced costs due to fewer delays, and enhanced customer satisfaction.

4. Legal and Ethical Feasibility

Identify legal and regulatory requirements related to data privacy, security, and compliance with aviation regulations. Assess potential ethical implications of using predictive models for flight delay prediction, such as fairness, transparency, and accountability. Ensure compliance with relevant laws and regulations governing the use of predictive analytics in the aviation industry.

5. Model performance

Predictive models may fail to accurately forecast flight delays, leading to suboptimal decision-making. Conduct thorough model validation and testing using historical data, implement ensemble methods, and continuously monitor and recalibrate models to improve performance.

6. Risk Analysis of the project

Inaccurate or incomplete data could lead to unreliable predictions. Implement data quality checks, data preprocessing techniques, and redundancy measures to address data quality issues. Disruption to airline operations or customer dissatisfaction due to inaccurate or unreliable predictions. Develop contingency plans and communication strategies to manage expectations and mitigate the impact of unforeseen events. Non-compliance with aviation regulations or data privacy laws could result in legal consequences or reputational damage. Stay abreast of regulatory requirements, collaborate with legal experts to ensure compliance, and implement robust data security and privacy measures.

Conduct fairness assessments, promote transparency in model development and decision-making processes, and prioritize ethical considerations throughout the project lifecycle. By conducting comprehensive feasibility studies and risk analysis,

project stakeholders can better understand the opportunities and challenges associated with flight delay prediction and develop strategies to mitigate risks and maximize project success.

Engage with stakeholders, including airlines, airports, and regulatory agencies, to ensure alignment with operational requirements and regulatory standards. Continuously monitor external factors and adjust analysis methodologies as needed to adapt to changing conditions.

3.2 SOFTWARE REQUIREMENTS SPECIFICATION DOCUMENT

Hardware specification :

- Microsoft Server enabled computer, preferably workstations
- Higher RAM of about 8GB or above
- Processor of frequency 1.5Hz or above

Software Specifications:

- Python 3.5 and higher
- Google Colab
- IDE are Jupyter Notebook, Pycharm Notebook.
- Libraries used are Pandas, Numpy, Sci-kit Learn, Matplotlib, Seaborn

CHAPTER 4

DESCRIPTION OF PROPOSED SYSTEM

The description of a proposed system for the analysis of flight delay data outlines the key components, functionalities, and objectives of the system. Here's a detailed description:

The proposed system is a comprehensive analytics platform designed to analyze flight delay data and provide actionable insights to stakeholders in the aviation industry. It integrates various data sources, analytical tools, and visualization techniques to facilitate the exploration, interpretation, and visualization of flight delay patterns and trends. The system retrieves flight delay data from multiple sources, including airline databases, airport records, weather services, air traffic control systems, and regulatory agencies. It employs data integration techniques to harmonize disparate datasets, ensuring consistency and accuracy in the analysis.

Prior to analysis, the system performs data preprocessing tasks, including data cleaning, normalization, and transformation, to address missing values, errors, and inconsistencies. It applies quality control checks to validate the integrity and reliability of the data, ensuring its suitability for analysis. The system employs a range of analytical techniques, including descriptive statistics, regression analysis, time series modeling, and machine learning algorithms, to uncover patterns and correlations within the flight delay data. It identifies significant factors contributing to flight delays, such as weather conditions, air traffic congestion, aircraft type, and operational inefficiencies, through exploratory data analysis and predictive modeling. The system generates interactive visualizations, including charts, graphs, maps, and dashboards, to present the analysis findings in a visually compelling and easy-to-understand manner.

It enables users to customize and interact with visualizations, explore data subsets, and drill down into specific delay factors or geographic regions of interest. The system generates comprehensive reports summarizing the analysis

results, insights, and recommendations for stakeholders. Leveraging historical flight delay data, the system develops predictive models to forecast future delay occurrences and trends. It assesses the accuracy and reliability of the predictive models through validation techniques and provides probabilistic forecasts of delay probabilities.

The system supports real-time monitoring of flight delays, continuously ingesting and analyzing incoming data streams to detect and respond to delay events as they occur. It generates alerts and notifications to stakeholders, enabling timely intervention and decision-making to mitigate the impact of delays. The system facilitates collaboration among stakeholders by providing shared access to analysis results, reports, and visualizations. It promotes knowledge sharing and best practices exchange through collaborative features such as discussion forums, document sharing, and workflow management.

Designed for scalability, the system can accommodate large volumes of data and scale up to support the analysis of diverse datasets and complex analytical tasks. It offers flexibility in terms of customization, allowing users to tailor analysis workflows, algorithms, and visualization templates to their specific needs and preferences. The system incorporates robust security measures to protect sensitive flight delay data and ensure compliance with data privacy regulations and industry standards. It implements user authentication, authorization, and encryption mechanisms to safeguard data integrity and confidentiality.

4.1 Datasets

The Title of my dataset that I am using here Analysis of Flight Delay Data the Dataset I am using contains three datasets: Airlines - 15 rows, 2 columns Airports - 323 rows, 7 columns Flights - 28821 rows, 23 columns. The Airlines Columns are IATA Code and Airline the Airports Columns are "IATA Code, Airport, City, State, Country, Latitude, Longitude.

The Flights Columns are Month, Day of month, Day of week, OP Unique Carrier, Tail num, Destination, Departure Delay, CRS Elapsed time, Distance, CRS

Departure M, Departure time M, CRS arrival M, Temperature, Dew Point, Humidity, Wind, Wind Speed, Wind gust, Pressure, Condition, Sch Dep, Sch arr, TAXI Out.

Here's an outline of the experimental or materials and methods section for analyzing flight delay data:

Data Source: Specify the source of your flight delay data. This could be publicly available datasets (e.g., Bureau of Transportation Statistics), data obtained directly from airlines, or a combination of sources.

Data Description: Describe the key features (columns) present in your data. Examples include:

- Flight identifier
- Airline
- Origin and destination airports
- Scheduled departure and arrival times
- Actual departure and arrival times
- Delay reason (if available)
- Weather data (if available)

Data Preprocessing:

Data Cleaning: Identify and address missing values, outliers, or inconsistencies in the data. This might involve techniques like imputation, removal, or winsorization.

Exploratory Data Analysis (EDA):

Descriptive Statistics: Summarize key flight delay characteristics like average delay, delay frequency, and distribution using measures like mean, median, standard deviation, and histograms.

Visualization:

Create visualizations like boxplots, scatter plots, and heatmaps to explore relationships between delay and factors like airline, route, weather, and time of day.

Modeling and Analysis:

Choose a method depending on your goals, you might choose methods like
Regression analysis To understand how different factors influence delay duration.
Classification analysis To predict if a flight will be delayed or not (binary classification) or the category of delay duration (multi-class classification).

Survival analysis: To model the time it takes for a flight delay to occur (if interested in delay likelihood over time).

Model Training and Evaluation: Split your data into training and testing sets. Train your chosen model on the training data and evaluate its performance on the testing data using relevant metrics (e.g., accuracy, precision, recall, F1-score for classification).

Model Selection: If using a modeling approach, justify your choice of model based on the research question and data characteristics.

Software Tools: Specify the software tools you'll be using for data analysis and modeling (e.g., Python libraries like Pandas, Scikit-learn,).

4.2 SELECTED METHODOLOGY OR PROCESS MODEL

Analyzing a flight delay dataset involves several common steps. I'll provide a high-level overview of the process for both tools:

1. Data Preparation:

Import your flight delay dataset . Common file formats like CSV, Excel, or databases are usually supported. Ensure that your dataset is clean, with no missing values or outliers. You may need to do data cleaning operations within the tools.

2. Data Exploration:

Start by examining the dataset You can use the "Data" view to see your data tables, or in, the "Data Source" tab provides a similar view.

3. Data Modeling:

Define relationships between different tables in your dataset if applicable, e.g., connecting tables for flights, airports, airlines, and weather data. Create calculated columns or measures to derive useful insights from your data.

4. Data Visualization:

Both are renowned for their data visualization capabilities. You can create a wide range of charts, graphs, to represent your data effectively. Common visualizations might include bar charts showing delay times, scatter plots to explore relationships, and heatmaps for spotting patterns.

5. Time-Series Analysis:

Flight delay data often includes a time component. Analyze this over time using line charts, time series analysis, and trend analysis to identify patterns and seasonality.

6. Drill-Down and Filtering:

Provide interactive features like filtering and drill-down to allow users to explore the data in more detail. In both tools, you can set up slicers or filters to enable this.

7. Geospatial Analysis:

If the dataset includes location data (e.g., airport codes), we can create geospatial visualizations to analyze flight delays by location.

8. Aggregations and Summary Metrics:

Calculate summary statistics, such as average delay times, total delays by airline, or worst-performing routes.

9. Data Storytelling:

Create dashboards or presentations that tell a compelling story using the data.

Explain your findings, insights, and recommendations clearly.

10. Performance Optimization:

Optimize the performance of your reports and dashboards, especially if dealing with large datasets. Both Power BI and Tableau provide options for data optimization

11. Collaboration and Sharing:

Share your reports and dashboards with your team or stakeholders. You can publish them to the web or share them within the organization.

12. Continuous Monitoring:

Regularly update your data and reports to keep them relevant. Automated data refreshes can be set up in both Power BI and Tableau. The specific steps and features used will depend on the complexity of your dataset and the insights you want to extract. Both Power BI and Tableau offer extensive documentation.

4.3 ARCHITECTURE / OVERALL DESIGN OF PROPOSED SYSTEM

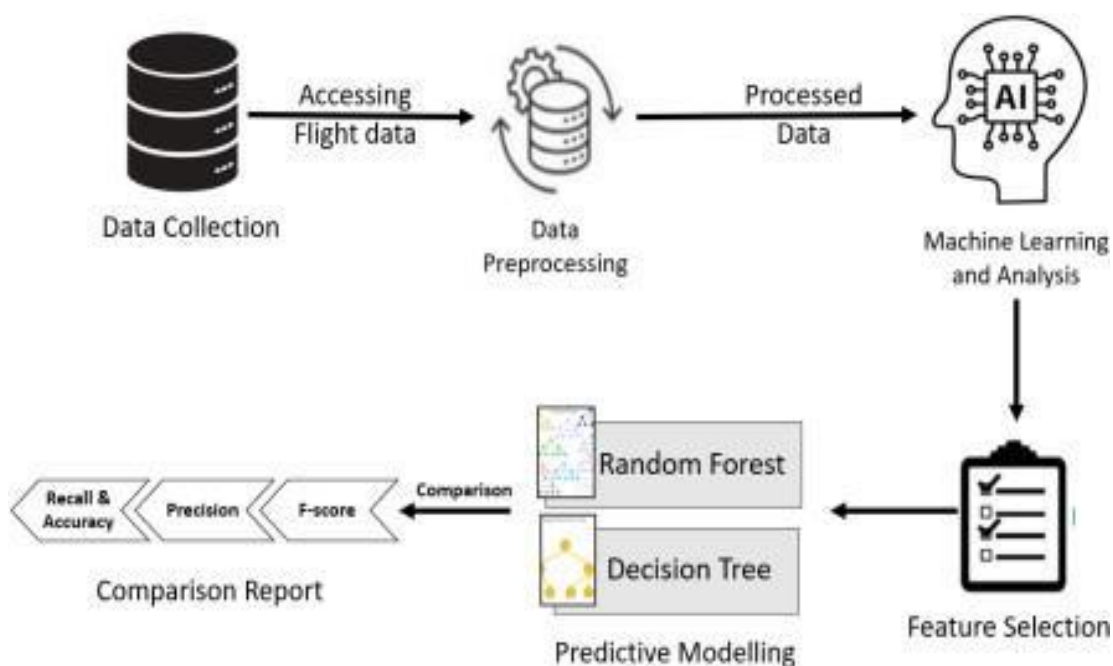


Fig 4.1: System Architecture for Analysis of Flight Delay Data

This architecture outlines the major components and their interactions in the project. It shows the flow of data from sources to the visualization tools while considering aspects like data security, performance, documentation, and change management.

4.4 DESCRIPTION OF SOFTWARE FOR IMPLEMENTATION AND TESTING PLAN OF THE PROPOSED MODEL/SYSTEM

Here's a description of the software for implementing and testing the proposed model/system for the analysis of flight delay data:

Software Description:

The software for implementing the flight delay data analysis system will be a comprehensive platform designed to collect, process, analyze, and visualize flight data to identify patterns and predict delays. The software will consist of several components.

1. Data Collection Module:

This module will gather data from various sources, including flight schedules, historical delay records, weather databases, airport information systems, and air traffic control data. Data collection processes will be automated and scheduled to ensure the timely acquisition of relevant data for analysis.

2. Data Preprocessing and Cleaning Module:

The preprocessing module will clean and preprocess raw data to address issues such as missing values, outliers, and inconsistencies. Techniques such as data imputation, outlier detection, and normalization will be applied to ensure data quality and consistency.

3. Analysis and Prediction Module:

This module will contain algorithms for analyzing flight delay data and predicting future delays. Machine learning techniques, such as regression, classification, and time series analysis, will be employed to identify patterns and trends in the data and make accurate predictions.

4. Visualization and Reporting Module:

The visualization module will generate interactive visualizations, charts, and graphs to present the analysis results in a user-friendly format. Users will be able to explore delayed trends, view predictions, and generate custom reports to support decision-making processes.

5. Integration and Deployment Module:

This module will facilitate the integration of the software with existing airline, airport, and aviation authority systems. Deployment processes will be automated to streamline the installation and setup of the software in different operational environments.

4.5 Project Management Plan

Developing a project management plan for the analysis of flight delay data involves outlining the project's objectives, defining the scope, allocating resources, establishing timelines, and implementing strategies to ensure successful execution. Here's a structured project management plan for this endeavor:

1) Project Initiation:

Clearly articulate the goals and objectives of the analysis, such as identifying causes of flight delays, predicting future delay occurrences, and proposing mitigation strategies. Identify key stakeholders, including airlines, airport authorities, regulatory bodies, and passengers, and establish communication channels to engage them throughout the project lifecycle. Develop a project charter outlining the project's scope, objectives, deliverables, timelines, and key stakeholders' roles and responsibilities.

2) Project Planning:

Define the scope of the analysis, including the types of flight delays to be investigated, the geographical coverage (e.g., regional, national, international), and the timeframe for data analysis. **Resource Allocation:** Identify and allocate necessary resources, including personnel, data sources, analytical tools, and computing infrastructure, to support the analysis. Develop a detailed WBS outlining the tasks,

subtasks, and dependencies involved in data collection, preprocessing, analysis, modeling, and reporting. **Timeline Development:** Create a project schedule with milestones, deliverables, and deadlines for each phase of the analysis, considering resource availability and project dependencies.

3) Risk Management:

Identify potential risks and uncertainties that could impact the project's success, such as data quality issues, resource constraints, technical challenges or regulatory changes. Assess the likelihood and potential impact of each identified risk and prioritize them based on severity and mitigation feasibility. Develop strategies and contingency plans to mitigate or manage identified risks, such as conducting data quality checks, establishing backup plans for critical resources, and alternative data.

4) Data Acquisition and Preparation:

Gather relevant flight delay data, weather data, air traffic data, airport infrastructure data, and any other required datasets from reliable sources, ensuring data completeness and accuracy. Cleanse, validate, and preprocess the collected data to address missing values, errors, inconsistencies, and outliers, ensuring data quality and consistency for analysis.

5) Data Analysis and Modeling:

Perform descriptive and inferential statistical analysis to explore the relationships between delay factors, flight characteristics, and temporal variables. **Predictive Modeling:** Develop predictive models using machine learning algorithms, such as regression analysis, time series forecasting, or classification, to predict future delay occurrences and identify significant predictors. **Interpretation and Validation:** Interpret the analysis results, validate the predictive models, and assess their accuracy and reliability using appropriate validation techniques.

6) Reporting and Communication:

Prepare comprehensive reports or presentations summarizing the analysis findings, insights, and recommendations in a clear and actionable manner. Engage key stakeholders through regular updates, progress meetings, and presentations to communicate project status, share insights, and solicit feedback. Facilitate

knowledge transfer by documenting the analysis methodologies, assumptions, and limitations, and providing training or workshops to stakeholders on interpreting the analysis results.

7) Project Closure:

Deliver final analysis reports, models, and documentation to stakeholders, ensuring smooth transition and handover of project deliverables. Conduct a post-project review to identify lessons learned, best practices, and areas for improvement, and document them for future reference. Complete all administrative tasks, such as archiving project files, closing out contracts, and updating project documentation, to formally close the project. By following this project management plan, the analysis of flight delay data can be effectively executed, leading to valuable insights and actionable recommendations for improving the efficiency and reliability of the air.

CHAPTER 5

IMPLEMENTATION DETAILS

The implementation details for analyzing flight delay data involve the technical aspects of building, deploying, and maintaining the analytical solution. Here's a structured overview of the implementation process:

Data Acquisition:

Establish connections to various data sources such as airline databases, airport records, weather services, air traffic control systems, and regulatory agencies. Develop scripts or workflows to automate the retrieval and ingestion of flight delay data into the analysis pipeline. Implement data quality checks to ensure the integrity, completeness, and consistency of the acquired data.

Data Preprocessing:

Cleanse and preprocess the raw flight delay data to address missing values, outliers, and inconsistencies. Perform data transformation, normalization, and feature engineering to prepare the data for analysis. Handle data integration tasks to merge and harmonize disparate datasets from multiple sources.

Exploratory Data Analysis (EDA):

Conduct exploratory analysis to gain insights into the distribution, patterns, and correlations within the flight delay data. Visualize key metrics and trends using charts, graphs, and interactive dashboards to facilitate data exploration and interpretation. Identify potential relationships between delay factors and explore temporal or spatial patterns in delay occurrences.

Statistical Analysis and Modeling:

Apply statistical techniques and machine learning algorithms to analyze flight delay data and identify significant delay factors. Develop predictive models to forecast future delay occurrences, using regression analysis, time series forecasting, or classification algorithms. Validate and optimize analytical models using

cross-validation, hyperparameter tuning, and performance evaluation metrics.

Visualization and Reporting:

Design and develop interactive visualizations and dashboards to present analysis results and insights in a user-friendly and accessible manner. Incorporate features for data exploration, filtering, and drill-down to enable users to interactively explore delay patterns and trends. Generate automated reports summarizing analysis findings, model performance, and actionable recommendations for stakeholders.

Deployment and Integration:

containerize analytical models, scripts, and dependencies using containerization technologies such as Docker for consistency and portability. Deploy containerized applications on cloud platforms (e.g., AWS, Azure, GCP) or on-premises servers using container orchestration tools like Kubernetes. Integrate the analytical solution with existing systems, applications, or workflows within the aviation industry to facilitate seamless data exchange and interoperability.

Monitoring and Maintenance:

Implement monitoring and logging mechanisms to track the performance, availability, and usage of the analytical solution. Set up alerts and notifications to detect anomalies, errors, or performance degradation in real-time and take proactive measures to address them. Establish regular maintenance procedures to update models, retrain algorithms, and optimize performance based on evolving data and requirements. **Documentation and Knowledge Sharing** Document the implementation details, including data sources, methodologies, algorithms, and system architecture, to facilitate knowledge transfer and reproducibility. Provide user documentation, tutorials, and training materials to onboard stakeholders and users to the analytical solution effectively.

Foster a culture of knowledge sharing and collaboration by organizing workshops, seminars, or forums to disseminate insights and best practices derived from the analysis of flight delay data. By following these implementation details, organizations can build a robust analytical solution for analyzing flight delay data, enabling informed decision-making management in the aviation industry.

5.1 DEVELOPMENT AND DEPLOYMENT OF SETUP

The development and deployment setup for analyzing flight delay data requires a robust infrastructure and a well-defined workflow to ensure efficient development, testing, and deployment of analytical solutions. Here's a proposed setup:

1) Hardware Infrastructure:

High-performance workstations or laptops equipped with sufficient processing power, memory, and storage for data analysis tasks. Virtual or physical servers hosted on cloud platforms or on-premises infrastructure to deploy analytical models and applications.

2) Data Analysis and Modeling Tools:

Utilize software packages such as Python (with libraries like Pandas, NumPy, SciPy), R, or MATLAB for data analysis, statistical modeling, and machine learning. Integrated Development Environments (IDEs): Use IDEs like Jupyter Notebook, PyCharm, RStudio, or Visual Studio Code for coding, experimentation, and collaboration. Database Management Systems (DBMS): Employ relational databases (e.g., PostgreSQL, MySQL) or NoSQL databases (e.g., MongoDB, Cassandra) to store and manage flight delay data. Visualization Tools: Leverage visualization libraries and tools such as Matplotlib, Seaborn, Plotly, or Tableau for creating interactive visualizations and dashboards. Version Control: Use version control systems like Git and platforms such as GitHub or GitLab for collaborative development, code sharing, and version tracking.

3) Development Workflow:

Data Acquisition and Preprocessing: Retrieve flight delay data from various sources and perform preprocessing tasks such as cleaning, transformation, and feature engineering. Conduct exploratory analysis to understand data distributions, patterns, and correlations, and identify relevant features and variables. Develop predictive models, regression analyses, time series forecasting techniques, or machine learning algorithms to analyze flight delay data and identify significant factors. Assess the performance and accuracy of analytical models using cross-validation, holdout validation, or other validation techniques, and iterate on model refinement as necessary.

4) Monitoring and Maintenance:

Implement monitoring and logging mechanisms to track the performance, availability, and health of deployed analytical solutions. Set up alerts and notifications to detect and respond to anomalies, errors, or performance degradation in real-time. Establish regular maintenance schedules to update dependencies, retrain models, and optimize performance based on evolving data and requirements. By establishing a robust development and deployment setup, organizations can streamline the process of analyzing flight delay data, develop effective analytical solutions, and deploy them efficiently to support decision-making and operational improvements in the aviation industry.

5.2 ALGORITHM

5.2.1 Random Forest

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It belongs to the ensemble learning methods, which means it combines the predictions from multiple individual models to produce a more accurate and robust final prediction. Random Forest builds a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It's an ensemble learning method that combines multiple decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. It's known for its high accuracy, robustness to overfitting, and ability to handle large datasets with high dimensionality.

Random Forest offers several advantages:

It is less prone to overfitting compared to individual decision trees, thanks to the random sampling and feature selection. It can handle large datasets with high dimensionality. It provides estimates of feature importance, which can be useful for feature selection.

Precision = $TP / (TP + FP)$ Recall = $TP / (TP + FN)$ TP = True Positive

FN = False Negative FP = False Positive

True Positive (TP) is a term used in binary classification to describe the cases where the model correctly predicts the positive class (the condition being looked for is present), and it is actually present in the data. In other words, a true positive occurs when the model correctly identifies a positive instance from the dataset as positive.

True Negative (TN) is another term used in binary classification to describe the cases where the model correctly predicts the negative class (the condition being looked for is absent), and it is actually absent in the data. In other words, a true negative occurs when the model correctly identifies a negative instance from the dataset as negative.

False Positive (FP) is a term used in binary classification to describe the cases where the model incorrectly predicts the positive class (the condition being looked for is present), but it is actually absent in the data. In other words, a false positive occurs when the model incorrectly identifies a negative instance from the dataset as positive.

False Negative (FN) is a term used in binary classification to describe the cases where the model incorrectly predicts the negative class (the condition being looked for is absent), but it is actually present in the data. In other words, a false negative occurs when the model incorrectly identifies a positive instance from the dataset as negative.

5.2.2 Decision Tree

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It's a tree-like structure where each internal node represents a "decision" based on a feature, each branch represents the outcome of the decision, and each leaf node represents the prediction (or outcome) of the model. A decision tree is a flowchart-like tree structure where an internal node represents a feature (or attribute), the branch represents a decision rule, and each leaf node represents the outcome (or class label). It's a popular algorithm for classification and regression tasks in machine learning. Decision trees are easy to interpret and understand, making them valuable for both analysis and decision-making. They recursively split the dataset into subsets based on the most significant attribute until a stopping criterion is met, such as reaching a maximum depth or no further improvement in purity measures like Gini impurity or entropy.

Decision Trees have several advantages:

- They are easy to interpret and visualize, making them useful for understanding the decision-making process of the model.
- They can handle both numerical and categorical data.
- They implicitly perform feature selection by selecting the most important features at the top of the tree.

Table 5.1: Flight Table

	AIRLINE	Org_Airport_Name	Origin_city	Dest_Airport_Name	Destination_city	ORIGIN_AIRPORT	DESTINATION_CITY
0	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle	ANC	SEA
1	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle	ANC	SEA
2	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle	ANC	SEA
3	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle	ANC	SEA
4	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle	ANC	SEA

The Flight Table describes the airline, name of the origin airport,origin city,name of the destination airport and the destination city.

Table 5.2: Airport Table

	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 Table describes the International Air Transport Association [IATA CODE] ,Airports, city,State and Country which we have taken from the dataset.

Table 5.3: Airline Table

	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

The Airline Table describes the International Air Transport Association [IATA CODE] and Airline which we have taken from the dataset.

As shown in fig 5.1 , the diagram visualizes the pairwise correlations between numerical variables in the 'Flights' Data Frame. The heatmap provides a quick and visual way to identify relationships between different features, with the annotated values indicating the strength and direction of these correlations.

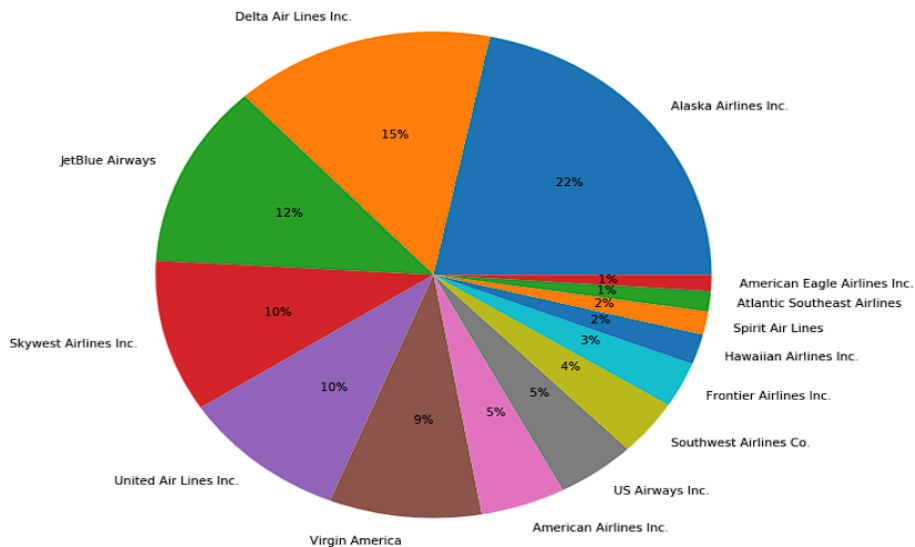


Fig 5.1 Pie chart of Airline

As shown in fig 5.2, we are creating a heatmap using the seaborn library to visualize the pairwise correlations between numerical variables in the 'Flights' DataFrame. The heatmap provides a quick and visual way to identify relationships between different features, with the annotated values indicating the strength and direction of these correlations

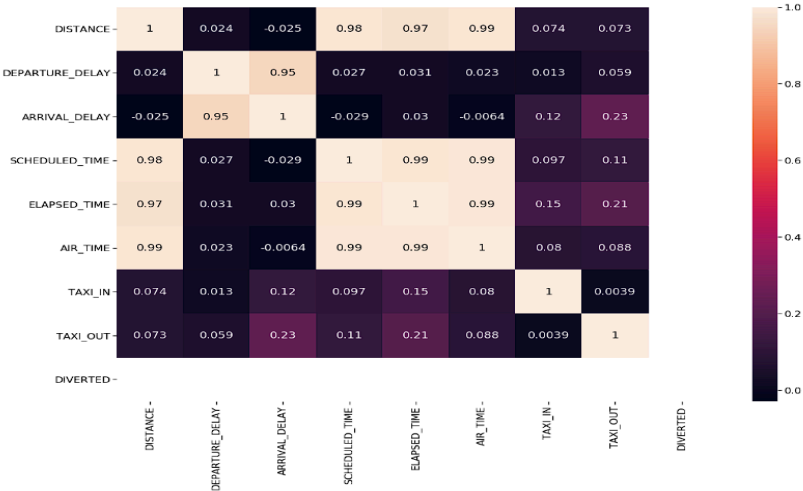


Fig 5.2 Heatmap

As shown in fig 5.3, the distribution plots to visualize the distribution of the 'AIR_TIME' variable in the 'Flights1' DataFrame. The plot helps you understand how the air time values are spread across different ranges and provides insights into the central tendency and variability of the 'AIR_TIME' variable in the dataset.

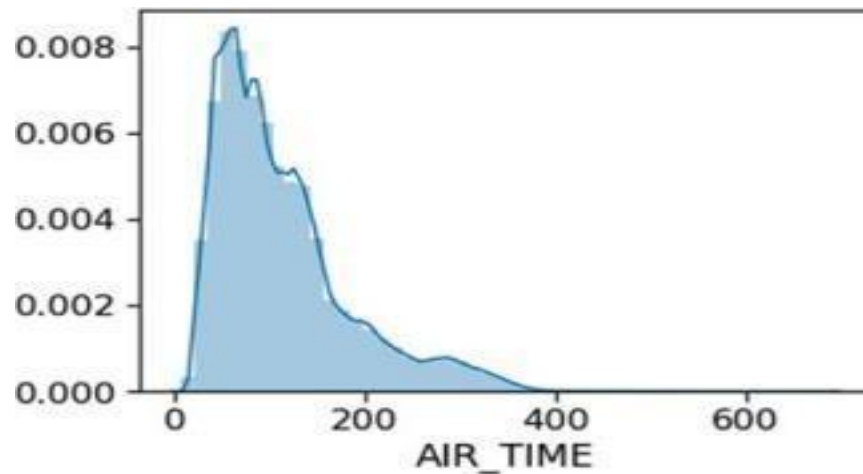


Fig 5.3 Distribution of Air time

As shown in fig 5.4, to visualize the distribution of the 'ELAPSED_TIME' variable in the 'Flights1' DataFrame. The plot provides insights into how the elapsed time values are distributed across different ranges, helping to understand the central tendency and variability of the 'ELAPSED_TIME' variable in the dataset.

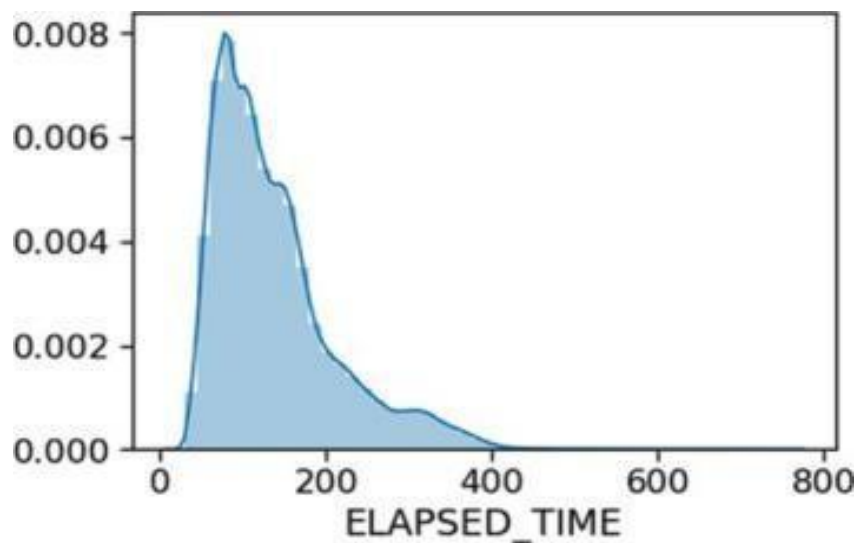


Fig 5.4 Distribution of elapsed time

5.3 TESTING

In the context of flight delay data analysis, testing refers to the process of evaluating the effectiveness of our findings and any models we might develop. Here are some key areas where testing would be crucial:

1)Hypothesis Testing: If we have specific hypotheses about the factors influencing delays (e.g., "delays are more frequent on certain airlines"), we can use statistical tests to assess their validity. This helps identify which factors have a statistically significant impact.

2)Model Performance: If we build a model to predict flight delays, testing is essential to gauge its accuracy. We can split the data into training and testing sets. The model is trained on the training data and then evaluated on the unseen testing data. Metrics like mean squared error or R-squared can assess how well the model predicts delays on new data.

3)Sensitivity Analysis: Testing can involve analyzing how sensitive our findings or model predictions are to changes in the data. For example, we might test how the model performs under different weather conditions or for specific origin/destination pairs.

4)Overfitting Prevention: During model development, testing helps prevent overfitting. This occurs when a model performs well on the training data but fails to generalize to unseen data. Techniques like cross-validation involve splitting the data into multiple folds for training and testing, providing a more robust evaluation of the model's generalizability.

By employing these testing approaches, we can ensure the robustness and validity of our analysis of flight delay data. This ultimately leads to more reliable insights and effective action plans for mitigating delays and improving air travel efficiency.

CHAPTER 6

RESULTS AND DISCUSSIONS

6.1 RESULTS

We can see that maximum arrival Delays are dependent on the Departure Delays of the Airport. The first part dealt with cleaning and exploration of the data set to get more insights and the second part dealt with the sitting of the model to predict the delays. For exploratory Analysis I used visualization tools like tableau and seaborn as well as matplotlib. The Second part dealt with the model fitting and predicting the Arrival delays of the airlines.

We can see that departure delay is the main problem which is creating Delay in the aviation industry. Departure Delays can be caused due Security Delay, Airline System Delays, Airlines Delay etc. The Delays affect the revenue of the company to a great extent so the delays has to be reduced as much as possible so as to increase the profitability in the Airline Industry. Customer Satisfaction will also be greatly enhanced if the delays can be brought down as low as possible.

6.2 DISCUSSION

The discussions for the analysis of flight delay data delve into the interpretation, implications, and broader context of the findings. Here's a breakdown of potential discussion points:

1) Interpretation of Findings:

Discuss the significance of the identified delay factors and their impact on flight operations. Highlight any patterns, trends, or correlations observed in the data analysis. Interpret the results of predictive modeling efforts, including the accuracy of forecasts and the factors driving delay predictions. Consider the implications of temporal and spatial variations in delay occurrences, such as peak periods, seasonal trends, and geographical clusters. Root causes and contributing factors explore the root causes behind different types of flight delays, including weather-related disruptions, air traffic congestion, operational inefficiencies, and infrastructure limitations. Analyze the relative contribution of various factors to

delay occurrences and assess their interconnectedness and dependencies.

2)Operational Implications:

Discuss how the analysis findings can inform operational decisions and strategies aimed at mitigating flight delays. This could include optimizing flight schedules, improving ground handling procedures, or implementing real-time monitoring systems. Consider the potential trade-offs between operational efficiency, safety, and customer satisfaction when implementing delay management strategies.

3)Impact on Stakeholders:

Assess the impact of flight delays on different stakeholders within the aviation ecosystem, including airlines, airports, passengers, and regulatory authorities. Discuss the economic implications of delays, including direct costs incurred by airlines and passengers, as well as broader economic effects on tourism, trade, and business productivity. Policy and Regulatory Considerations;

examine the role of regulatory frameworks, air traffic management systems, and government policies in mitigating flight delays. Propose recommendations for regulatory reforms or policy interventions to address systemic issues contributing to delays, such as airspace congestion or airport capacity constraints.

Technological Innovations and Solutions:

Explore emerging technologies and innovations that have the potential to reduce flight delays, such as advanced weather forecasting systems, predictive analytics, or air traffic automation. Discuss the feasibility and scalability of implementing technological solutions within the current aviation infrastructure. Collaboration and

4)Industry Initiatives:

Highlight the importance of collaboration between airlines, airports, air traffic control agencies, and other stakeholders in implementing effective delay management strategies. Discuss industry initiatives, partnerships, or best practices aimed at improving operational efficiency and minimizing delays.

5)Future Directions and Research Opportunities:

Identify areas for further research and exploration, such as the impact of climate change on weather-related delays, the integration of unmanned aerial vehicles (drones) into airspace management, or the optimization of airport capacity utilization. Consider the potential for cross-disciplinary collaboration with fields such as meteorology, transportation engineering, and data science to address complex delay challenges.

6.3 Performance Analysis:

Performance analysis for the analysis of flight delay data involves evaluating the effectiveness, accuracy, and efficiency of the analysis process and its outcomes. Here are key components of performance analysis for this endeavor:

1)Accuracy of Predictive Models:

Evaluate the accuracy of predictive models developed to forecast flight delays. Use metrics such as mean absolute error (MAE), mean squared error (MSE), or root mean squared error (RMSE) to assess the deviation between predicted and actual delay times. Validate the models using cross-validation techniques to ensure their robustness and generalizability across different datasets and time periods. Compare the performance of different modeling approaches (e.g., regression, time series forecasting, machine learning algorithms) to identify the most effective technique for predicting flight delays.

2)Identification of Significant Delay Factors:

Assess the ability of the analysis to identify and quantify significant factors contributing to flight delays, such as weather conditions, air traffic congestion, aircraft type, and airport infrastructure. Conduct sensitivity analysis to evaluate the impact of individual delay factors on overall delay occurrences and assess their relative importance.

3)Temporal Analysis:

Analyze the temporal patterns of flight delays over different time scales (e.g., hourly, daily, monthly, seasonal) to identify peak delay periods, trends, and fluctuations. Evaluate the accuracy of temporal forecasting models in predicting

future delay occurrences and trends.

4)Spatial Analysis:

Perform spatial analysis to examine the geographical distribution of flight delays across different airports, regions, and flight routes. Identify spatial hotspots or clusters of delay occurrences and assess their underlying causes (e.g., airport congestion, weather patterns).

5)Operational Efficiency Analysis:

Measure the impact of the analysis findings on enhancing operational efficiency and reducing delay occurrences within the aviation system. Quantify potential cost savings or revenue benefits resulting from improved scheduling, resource allocation, and delay management strategies informed by the analysis.

6)Real-time Monitoring and Feedback:

Develop mechanisms for real-time monitoring of flight delays and performance metrics derived from the analysis. Establish feedback loops with stakeholders to gather insights, validate analysis results, and refine predictive models and strategies based on real-world observations and feedback.

7)Comparative Analysis:

Compare the performance of the analysis against industry benchmarks or best practices in delay management and operational efficiency. Benchmark the accuracy and effectiveness of predictive models and delay mitigation strategies against similar studies or initiatives in the aviation domain.

8)User Satisfaction and Adoption:

Solicit feedback from stakeholders, including airlines, airport authorities, and passengers, to assess their satisfaction with the analysis outcomes and the usability of the derived insights and recommendations. Monitor the adoption of analysis-driven strategies and interventions within the aviation industry and evaluate their impact on mitigating flight delays and improving overall service quality

CHAPTER 7

CONCLUSION

7.1 SUMMARY

The project analysis of flight delay data involves utilizing various data analysis techniques and tools to understand, predict, and mitigate flight delays. It encompasses collecting and preprocessing large datasets containing information on flight schedules, historical performance, weather conditions, air traffic, and other relevant factors. Through advanced analytics and predictive modeling, the project aims to identify patterns, trends, and factors contributing to flight delays, enabling stakeholders such as airlines, airports, and regulatory bodies to make informed decisions and implement proactive measures to minimize disruptions to air travel. Key components of the analysis may include developing predictive models using machine learning algorithms, visualizing data to gain insights, integrating real-time data streams for timely decision-making, and leveraging technology such as AI, blockchain, and optimization algorithms to enhance the accuracy and effectiveness of delay prediction and management strategies. Ultimately, the goal of the project analysis is to improve the efficiency, reliability, and overall passenger experience within the aviation industry while addressing the complex challenges associated with flight delays.

7.2 CONCLUSION

In conclusion, the project analysis of flight delay data presents a multifaceted approach to addressing one of the most significant challenges in the aviation industry. By leveraging advanced data analytics, machine learning, and emerging technologies, stakeholders can gain valuable insights into the causes and patterns of flight delays, empowering them to implement proactive measures for mitigation and optimization. Through the integration of historical data, real-time information streams, and predictive modeling techniques, the analysis facilitates informed decision-making at various levels of the aviation ecosystem. Airlines can improve operational efficiency by optimizing flight schedules, maintenance planning, and resource allocation.

7.3 FUTURE ENHANCEMENTS

Predicting future enhancements in the analysis of flight delay data involves considering current trends in technology, data science, and aviation industry needs. Here are some potential future enhancements:

1) Blockchain for Data Integrity: Implementing blockchain technology to ensure the integrity and security of flight delay data, allowing for transparent and tamper-proof recording of delays and related information.

2) Application of AI in Maintenance Planning : Utilizing AI algorithms to predict potential maintenance issues that could lead to flight delays. By analyzing historical maintenance data and combining it with real-time sensor data from aircraft, airlines can proactively schedule maintenance to minimize disruptions

3) Real-time Data Integration: Integration of real-time data streams from various sources such as weather satellites, air traffic control systems, and aircraft sensors to provide more accurate and timely predictions of flight delays.

4) Integration of Passenger Data: Incorporating passenger data such as booking patterns, preferences, and historical flight experiences into delay prediction models to provide personalized recommendations and assistance in case of delays.

5) Environmental Impact Analysis: Incorporating environmental factors such as carbon emissions and fuel consumption into delay analysis to understand the environmental impact of flight delays and optimize operations for sustainability. Develop mechanisms for real-time monitoring of flight delays and performance metrics derived from the analysis. Establish feedback loops with stakeholders to gather insights, validate analysis results, and refine predictive models and strategies based on real-world observations and feedback.

6)Comparative Analysis:

Compare the performance of the analysis against industry benchmarks or best practices in delay management and operational efficiency. Benchmark the accuracy and effectiveness of predictive models and delay mitigation strategies against similar studies or initiatives in the aviation domain.

7)User Satisfaction and Adoption:

Solicit feedback from stakeholders, including airlines, airport authorities, and passengers, to assess their satisfaction with the analysis outcomes and the usability of the derived insights and recommendations. Monitor the adoption of analysis-driven strategies and interventions within the aviation industry and evaluate their impact on mitigating flight delays and improving overall service quality

REFERENCES

- [1] Fujun Wang ,Jun Bi ,Dongfan Xie ,Xiaomei Zhao,Quick taxi route assignment via real-time intersection state prediction with a spatial-temporal graph neural network, 2024
- [2] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices",2014
- [3] Peng Hu, Jia Hou, Jianping Zhang, Qingyuan Yang, Lei Cao, Liwei Wang,Research on flight arrival delay prediction based on support vector machine, 2022
- [4] Rahul Nigam, K. Govinda A. Kakaraparthi and V. Karthick, "A secure and cost-effective platform for employee management system using lightweight standalone framework over Diffie Hellman's key exchange algorithm," ECS Trans., vol. 107, no. 1, pp. 13663–13674, Apr.2022.
- [5] Wikipedia "Mean Squared Error", Available: https://en.wikipedia.org/wiki/Mean_squared_error
- [6] G. Ataman and S. Kahraman, "Flight Prediction in Using Linear Regression and Artificial Neural Network Hybrid Models," The Singapore Economic Review. pp. 1–19, 2021. doi: 10.1142/s0217590821500521.
- [7] J. C. Lok, Prediction Factors Influence Airline Fuel Price Changing Reasons. 2018.
- [8] T. M. Vowles, "Airfare pricing determinants in hub-to-hub markets," Journal of Transport Geography, vol. 14, no. 1, pp. 15–22, 2006.
- [9] K. Rama-Murthy, "Modeling of United States airline fares—using the official airline guide (OAG) and airline origin and destination survey (DB1B)," Ph.D. dissertation, Virginia Tech,2006.
- [10] Tianyi wang, Samira Pouyanfar, Haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"
- [11] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in proceedings of the 2013 international conference on autonomous agents and multi-agent systems.
- [12] J. Santos Dominguez-Menchero, Javier Rivera and Emilio TorresManzanera "Optimal purchase timing in the airline market".

- [13] J. Stavins, "Price discrimination in the airline market: The effect of market concentration," *Review of Economics and Statistics*, vol. 83, no. 1, pp. 200–202, 2001.
- [14] B. Martin and B. Koo, "Dynamic price dispersion in airline markets," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no.6, pp. 1020–1029, 2009.
- [15] P. Malighetti, S. Palesi, and R. Redondi, "Has Ryanair's pricing strategy changed over time? an empirical analysis of its 2006–2007 flights," *Tourism Management*, vol. 31, no. 1, pp. 36–44, 2010
- [16] Zhen Guo, Bin Yu, Mengyan Hao, Wensi Wang," A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient"
- [17] M. Bardach, E. Gringinger, M. Schrefl and C. G. Schuetz, "Predicting Flight Delay Risk Using a Random Forest Classifier Based on Air Traffic Scenarios and Environmental Conditions," 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020, pp. 1-8,
- [18] V. Natarajan, S. Meenakshisundaram, G. Balasubramanian and S. Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1081-1086,
- [19] K. Wang, J. Li and Y. Tian, "Airport Delay Prediction Method based on Improved Weather Impacted Traffic Index," 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Kunming, China, 2019, pp. 73-78
- [20] Z. Guo et al., "A Big Service with Network Represent Learning for Quantified Flight Delay Prediction," 2020 IEEE International Conference on Web Services (ICWS), Beijing, China, 2020, pp. 289-293, doi: 10.1109/ICWS49710.2020.00044.
- [21] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016, pp. 1-6.

APPENDIX

A. SOURCE CODE

```
import datetime, warnings, scipy
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
data = pd.read_csv("flights(1).csv")
airports = pd.read_csv('airports (1).csv')
airlines = pd.read_csv('airlines (1).csv')
data.info()
data.shape
data.describe()
data.head()
data.tail()
airport.isnull().sum()
airport = airport.dropna(subset = ['LATITUDE','LONGITUDE'])
airport.isnull().sum()
airport.head(10)
airlines
Data_NULL = data.isnull().sum()*100/data.shape[0]
Data_NULL
data1=data.dropna(subset=['TAIL_NUMBER','DEPARTURE_TIME','DEPARTURE_
DELAY','TAXI_OUT','WHEELS_OFF','SCHEDULED_TIME','ELAPSED_TIME','AIR
_TIME','WHEELS_ON','TAXI_IN','ARRIVAL_TIME','ARRIVAL_DELAY'])
data1.shape
data1.isnull().sum()
data11=data1.dropna(subset=['AIR_SYSTEM_DELAY','SECURITY_DELAY',
'AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY'])
```

```

data11 = data11.drop(['YEAR','MONTH','DAY', 'DAY_OF_WEEK', 'TAIL_NUMBER',
'SCHEDULED_DEPARTURE', 'DEPARTURE_TIME', 'SCHEDULED_TIME',
'SCHEDULED_ARRIVAL', 'ARRIVAL_TIME', 'DIVERTED', 'CANCELLED',
'CANCELLATION_REASON', 'FLIGHT_NUMBER', 'WHEELS_OFF',
'WHEELS_ON', 'AIR_TIME'],axis = 1)
data11.info()
Flight_Delays = data11
data1.drop(['CANCELLATION_REASON','AIR_SYSTEM_DELAY','SECURITY_DE
LAY','AIRLINE_DELAY', 'LATE_AIRCRAFT_DELAY','WEATHER_DELAY'],axis =1)
data2.isnull().sum()
data2.shape
data2.info()
data2.DEPARTURE_TIME.isnull().sum()
data2.DEPARTURE_TIME.dtype
data2.DEPARTURE_TIME
def Format_Hourmin(hours):
if hours == 2400:
hours = 0
else:
hours = "{0:04d}".format(int(hours))
Hourmin = datetime.time(int(hours[0:2]), int(hours[2:4]))
return Hourmin
data2['Actual_Departure']=data1['DEPARTURE_TIME'].apply(Format_Hourmin)
data2['Actual_Departure']
data2.columns
data2['Date'] = pd.to_datetime(data2[['YEAR','MONTH','DAY']]) data2.Date
data2['Day'] = data2['Date'].dt.dayofweek
data2['Actual_Departure']=data1['DEPARTURE_TIME'].apply(Format_Hourmin)
data2['Scheduled_Arrival']=data1['SCHEDULED_ARRIVAL'].apply(Format_Hourmin)
data2['Scheduled_Departure']=data1['SCHEDULED_DEPARTURE'].apply(Format_Ho
urmin)

```

```

data2['Actual_Arrival']=data1['ARRIVAL_TIME'].apply(Format_Hourmin)
data2 = data2.merge(airlines, left_on='AIRLINE', right_on='IATA_CODE', how='inner')
data2 = data2.drop(['AIRLINE_x','IATA_CODE'], axis=1)
data2
data2.rename(columns={'IATA_CODE_x':'Org_Airport_Code','AIRPORT_x':'Org_Airport_Name','CITY_x':'Origin_city','IATA_CODE_y':'Dest_Airport_Code','AIRPORT_y':'Dest_Airport_Name','CITY_y':'Destination_city'})
data2
# we are taking the required data into Account for visualization and the Analysis
ReqdData=pd.DataFrame(data2[['AIRLINE','Org_Airport_Name','Origin_city','Dest_Airport_Name','Destination_city','ORIGIN_AIRPORT','DESTINATION_AIRPORT','DISTANCE','Actual_Departure','Date','Day','Scheduled_Departure','DEPARTURE_DELAY','Actual_Arrival','Scheduled_Arrival','ARRIVAL_DELAY','SCHEDULED_TIME','ELAPSED_TIME','AIR_TIME','TAXI_IN','TAXI_OUT','DIVERTED',]]) data2.DEPARTURE_TIME.dtype
ReqdData=ReqdData.dropna(subset=['Actual_Departure','Actual_Arrival'])
ReqdData.info()
Flights = ReqdData
Flights plt.figure(figsize=(10, 10))
axis=sns.countplot(x=Flights['Origin_city'],data=Flights,order=Flights['Origin_city'].value_counts().iloc[:20].index,axis.set_xticklabels(axis.get_xticklabels(),rotation=90, ha="right") plt.tight_layout()
plt.show()
axis = plt.subplots(figsize=(10,14))
Name = Flights["AIRLINE"].unique()
size = Flights["AIRLINE"].value_counts()
plt.pie(size,labels=Name,autopct='%5.0f%%')
plt.show()
%%html
<divclass='tableauPlaceholder'id='viz1569936847474'style='position:relative'><noscript><ahref='#'><imgalt="src='https://public.tableau.com/static/7/images/9B/9 BD44NRJ9/1_rss.png' style='border: none'

```

```

/></a></noscript><objectclass='tableauViz'style='display:none;'><paramname='host_url'value='https%3A%2F%2Fpublic.tableau.com%2F'><paramname='embed_code_version' value='3' /> <param name='path' value='shared&#47;9BD44NRJ9'><paramname='toolbar'value='yes'>
<paramname='static_image'value='https:&#47;&#47;public.tableau.com&#47;Istatic&#47;images&#47;9B&#47;9BD44NRJ9&#47;1.png'><paramname='animate_transition'value='yes'><paramname='display_static_image' value='yes' /><param name='display_spinner' value='yes'><param name='display_overlay' value='yes' /><param name='display_count' value='yes' /><param name='filter' value='publish=yes' /></object></div>
<script type='text/javascript'>
var divElement = document.getElementById('viz1569936847474'); var vizElement =divElement.getElementsByTagName('object')[0];vizElement.style.width='100%';vizElement.style.height=(divElement.offsetWidth*0.75)+'px'; var scriptElement = document.createElement('script');
scriptElement.src='https://public.tableau.com/javascripts/api/viz_v1.js';
vizElement.parentNode.insertBefore(scriptElement, vizElement); </script>
axis = plt.subplots(figsize=(10,14))
sns.despine(bottom=True, left=True)
# Observations with Scatter Plot
sns.stripplot(x="ARRIVAL_DELAY", y="AIRLINE", data = Flights, dodge=True, jitter=True )
plt.show()
# Plot to show the Taxi In and Taxi Out Time
axis = plt.subplots(figsize=(10,14))
sns.set_color_codes("pastel")
sns.set_context("notebook", font_scale=1.5)
axis = sns.barplot(x="TAXI_OUT", y="AIRLINE", data=Flights, color="g") axis = sns.barplot(x="TAXI_IN",y="AIRLINE",data=Flights,color="r")axis.set(xlabel="TAXI_TIME (TAXI_OUT: green, TAXI_IN: blue)") axis = plt.subplots(figsize=(20,14))
sns.heatmap(Flights.corr(),annot = True)
plt.show()

```

%%html

```
<div class='tableauPlaceholder' id='viz1569938215526' style='position:
relative'><noscript><a href='#'><img alt='src='https://public.tableau.com/
static/images/PH/PHNKQPSZG/1_rss.png' style='border:
none' /></a></noscript><object class='tableauViz' style='display:none;'><param
name='host_url' value='https%3A%2F%2Fpublic.tableau.com%2F' /> <param
name='embed_code_version' value='3' /> <param name='path'
value='shared/PHNKQPSZG'
/> <param name='toolbar' value='yes' /><param name='static_image'
value='https://public.tableau.com/static/images/PH/
axis=sns.countplot(x=Flights['Origin_city'],data=Flights,order=Flights['Origin_city'].v
alue_counts().iloc[:20].index) a PHNKQPSZG/1.png'
/> <param name='animate_transition' value='yes' /><param
name='display_static_image' value='yes' /><param name='display_spinner'
value='yes'
/><param name='display_overlay' value='yes'
/><param name='display_count' value='yes' /></object></div> <script
type='text/javascript'>
var divElement = document.getElementById('viz1569938215526'); var vizElement
=divElement.getElementsByTagName('object')[0];vizElement.style.width='100%';vi
zElement.style.height=(divElement.offsetWidth*0.75)+'px';varscriptElement=docum
ent.createElement('script');scriptElement.src='https://public.tableau.com/javascripts
/api/viz_v1.js';vizElement.parentNode.insertBefore(scriptElement, vizElement);
</script>
Flights.head()
# Dropping unnecessary columns before prediction
Flights1=Flights.drop(['Org_Airport_Name','Origin_city','Dest_Airport_Name','Desti
nation_city'],axis = 1)
Flights1.columns
# To check the Distribution of Air Time
sns.distplot(Flights1['AIR_TIME'])
plt.show()
```

```

# To check the Distribution of Elapsed Time
sns.distplot(Flights1['ELAPSED_TIME'])
plt.show()

# To check the Distribution of Taxi IN
sns.distplot(Flights1['TAXI_IN'])
plt.show()

# To check the Distribution of Taxi out
sns.distplot(Flights1['TAXI_OUT'])
plt.show()

# importing Various regression algorithms
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.linear_model import Lasso, LinearRegression, Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import
RandomForestRegressor, AdaBoostRegressor, BaggingRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
Las = Lasso()
LinR = LinearRegression()
Rid = Ridge()
Rfc = RandomForestRegressor(random_state=2)
Dtc = DecisionTreeRegressor(random_state = 2)
Boost_Lin= AdaBoostRegressor(base_estimator=LinR, random_state=2)
Boost_las=AdaBoostRegressor(base_estimator=Las, random_state=2)
Boost_rid=AdaBoostRegressor(base_estimator=Rid, random_state=2)
Bg_Lin=BaggingRegressor(base_estimator=LinR, random_state=2)
Bg_las=BaggingRegressor(base_estimator=Las, random_state=2)
Bg_rid=BaggingRegressor(base_estimator=Rid, random_state=2)
le=LabelEncoder()

# Label encoding features to change categorical variables into numerical one

```

```

Flights1['AIRLINE']= le.fit_transform(Flights1['AIRLINE'])
Flights1['ORIGIN_AIRPORT']=le.fit_transform(Flights1['ORIGIN_AIRPORT'])
Flights1['DESTINATION_AIRPORT'] =
le.fit_transform(Flights1['DESTINATION_AIRPORT'])
Flights1['Day'] = le.fit_transform(Flights1['Day'])
Flights1=Flights1.drop(['Scheduled_Departure','Scheduled_Arrival','Actual_Arrival','
Date','Ac
axis.set_xticklabels(axis.get_xticklabels(), rotation=90, ha="right") plt.tight_layout()
plt.show()
tual_Departure'], axis = 1)
Flights1.info()
X = Flights1.drop('ARRIVAL_DELAY',axis = 1)
X.shape
y = Flights1['ARRIVAL_DELAY']
y.head()
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state = 2)
sc1=StandardScaler()
X_train_sc=sc1.fit_transform(X_train)
X_test_sc=sc1.transform(X_test)
for model, name in
zip([Las,LinR,Rid,Dtc,Rfc,Boost_Lin,Boost_las,Boost_rid,Bg_Lin,Bg_las,Bg_rid],
['Lasso','Linear Regression','Ridge','Random forest Regressor','Decision Tree
Regressor','Boosted Linear', 'Boosted Lasso','Boosted Ridge','Bagged Linear',
'Bagged Lasso', 'Bagged Ridge']):
model1 = model.fit(X_train_sc,y_train)
Y_predict=model1.predict(X_test_sc)
print(name)
print('Mean Absolute Error:', mean_absolute_error(y_test, Y_predict)) print('Mean
Squared Error:', mean_squared_error(y_test, Y_predict)) print('Root Mean
Squared Error:', np.sqrt(mean_squared_error(y_test, Y_predict)))
print('R2 : ',r2_score(y_test, Y_predict))
print()

```

```

for model, name in
zip([Las,LinR,Rid,Dtc,Rfc,Boost_Lin,Boost_las,Boost_rid,Bg_Lin,Bg_las,Bg_rid],
['Lasso','Linear Regression','Ridge','Random forest Regressor','Decision Tree
Regressor','Boosted Linear', 'Boosted Lasso','Boosted Ridge','Bagged Linear',
'Bagged Lasso', 'Bagged Ridge']):
model1 = model.fit(X_train_sc,y_train)
Y_predict=model1.predict(X_test_sc)

print(name)
plt.scatter(y_test, Y_predict)
plt.title("Model Analysis")
plt.xlabel("Truth")
plt.ylabel("Prediction")
plt.show()

axis = plt.subplots(figsize=(10,14))
Name = Flight_Delays["AIRLINE"].unique()
size = Flight_Delays["AIRLINE"].value_counts()
plt.pie(size,labels=Name,autopct='%5.0f%%')
plt.show()

Flight_Delays
Flight_Delays['ORIGIN_AIRPORT']=Flight_Delays['ORIGIN_AIRPORT'].astype(str)
Flight_Delays['DESTINATION_AIRPORT']=Flight_Delays['DESTINATION_AIRPORT'].astype(str)
Flight_Delays['AIRLINE']=le.fit_transform(Flight_Delays['AIRLINE'])
Flight_Delays['ORIGIN_AIRPORT']=le.fit_transform(Flight_Delays['ORIGIN_AIRPORT'])
Flight_Delays['DESTINATION_AIRPORT']=le.fit_transform(Flight_Delays['DESTINATION_AIRPORT'])
X = Flight_Delays.drop('ARRIVAL_DELAY',axis = 1)
X.shape
y = Flight_Delays['ARRIVAL_DELAY']

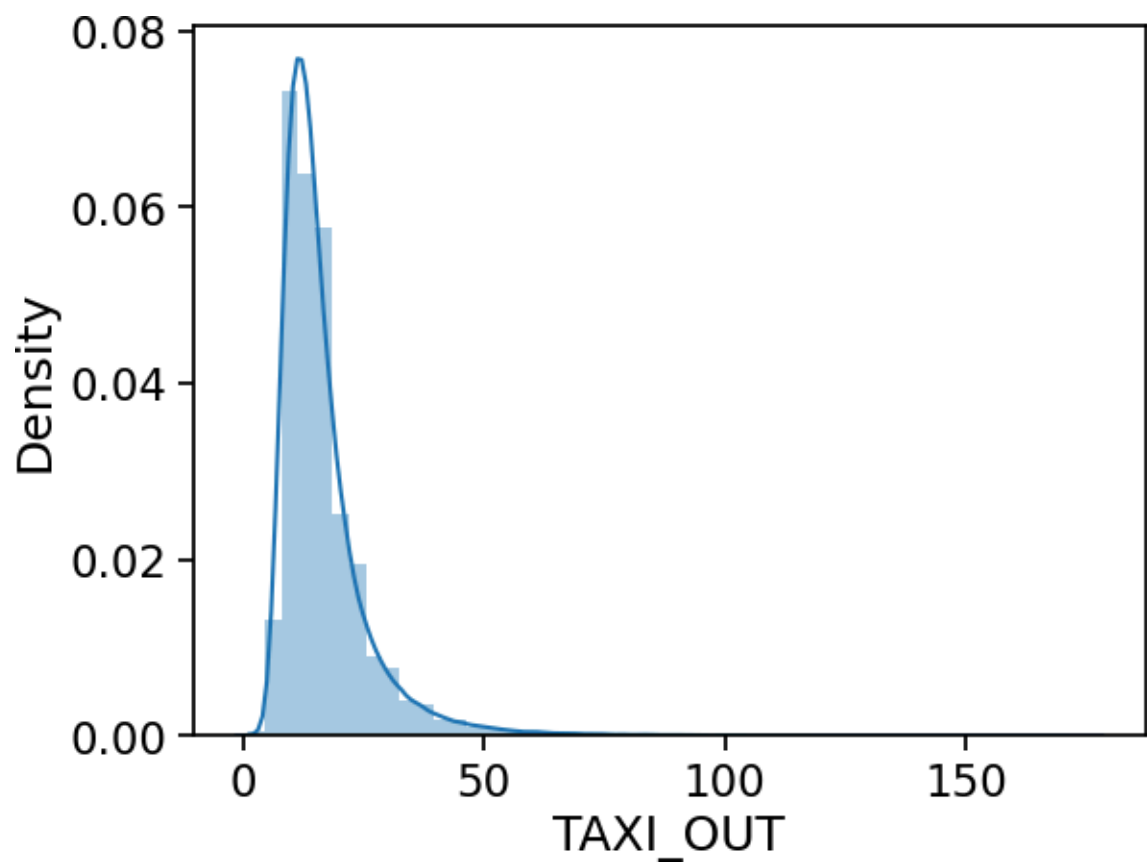
```

```

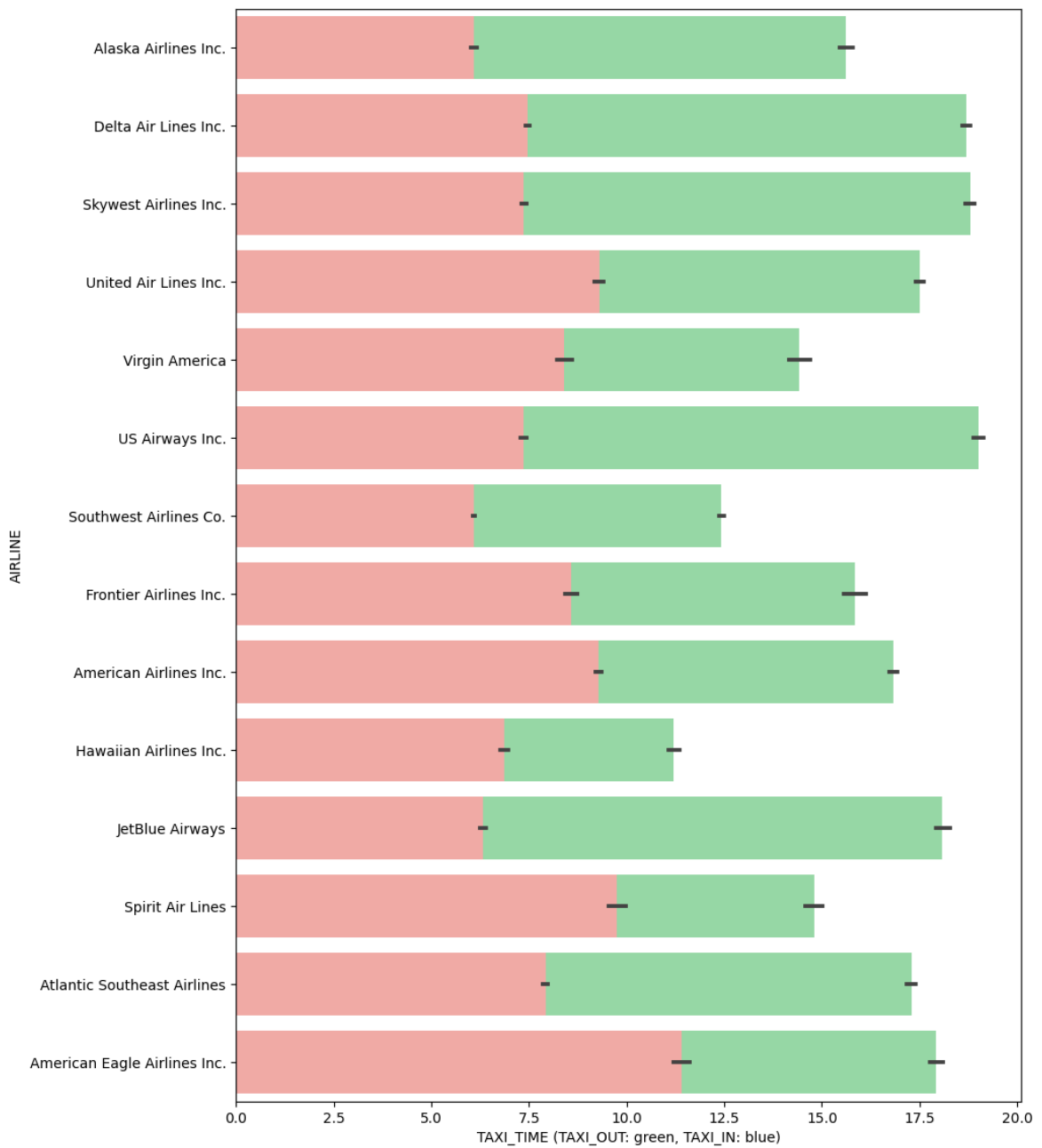
y.head()
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state = 2)
sc1=StandardScaler()
X_train_sc=sc1.fit_transform(X_train)
X_test_sc=sc1.transform(X_test)
for model, name in
zip([Las,LinR,Rid,Dtc,Rfc,Boost_Lin,Boost_las,Boost_rid,Bg_Lin,Bg_las,Bg_rid],
['Lasso','Linear Regression','Ridge','Random forest Regressor','Decision Tree
Regressor','Boosted Linear', 'Boosted Lasso','Boosted Ridge','Bagged Linear',
'Bagged Lasso', 'Bagged Ridge']):
model1 = model.fit(X_train_sc,y_train)
Y_predict=model1.predict(X_test_sc)
print(name)
print('Mean Absolute Error:', mean_absolute_error(y_test, Y_predict)) print('Mean
Squared Error:', mean_squared_error(y_test, Y_predict)) print('Root Mean
Squared Error:', np.sqrt(mean_squared_error(y_test, Y_predict)))
print('R2 : ',r2_score(y_test, Y_predict))
print()
for model, name in
zip([Las,LinR,Rid,Dtc,Rfc,Boost_Lin,Boost_las,Boost_rid,Bg_Lin,Bg_las,Bg_rid],
['Lasso','Linear Regression','Ridge','Random forest Regressor','Decision Tree
Regressor','Boosted Linear', 'Boosted Lasso', 'Boosted Ridge', 'Bagged Linear',
'Bagged Lasso', 'Bagged Ridge']):
model1 = model.fit(X_train_sc,y_train)
Y_predict=model1.predict(X_test_sc)
print(name)
plt.scatter(y_test, Y_predict)
plt.title("Model Analysis")
plt.xlabel("Truth")
plt.ylabel("Prediction")
plt.show()

```

B. SCREENSHOTS



B.1 Taxi Out



B.2 Taxi Time

C. RESEARCH PAPER

Analysis of Flight Delay Data

Abhinav Kumar

“Department of Computer Science and
Engineering Sathyabama Institute of
Science and Technology, Chennai, India”
abkr1712@gmail.com

Abhishek Raj

Department of Computer Science and
Engineering Sathyabama Institute of
Science and Technology, Chennai, India
abhishekkumarbxr101@gmail.com

Dr.A.Pravin

Department of Computer Science and
Engineering Sathyabama Institute of
Science and Technology, Chennai, India
pravin.cse@sathyabama.ac.in

Abstract—The aviation sector is significantly impacted by flight delays, leading to operational inefficiencies, customer dissatisfaction, and economic losses. This presents a comprehensive overview of flight delay data analysis, highlighting its importance in identifying patterns, predicting delays, and improving airline operations. The goal of the dashboards is to visualize which airlines have the most delays and potential causes. Accurately anticipating these flight delays helps airlines address probable causes of delays ahead of time, minimizing the detrimental effects and enabling customers to be well-prepared for the disruption to their travel plans. The aviation industry, marked by its dynamic and time-sensitive nature, constantly strives to enhance operational efficiency, improve passenger satisfaction, and address the persistent issue of flight delays.

Keywords—Machine Learning, Data Visualization, Random Forest Algorithm, Decision Tree

I INTRODUCTION

As more and more individuals opt to travel by air, an increasing number of flights are missing their planned departure schedules. This development worsens airport traffic and jeopardizes the financial stability of the airline industry. Travel delays are an indicator of aviation industry inefficiencies. The cost to airline companies and their passengers is high.

Comparing how well machine learning classification methods perform in forecasting flight delays is the main aim of the study. Data pertaining to flights departing from JFK airport during a one-year period was being examined. The whole cost of air transportation delays travel, based on the Total Delay Impact Study's assessment.

One of the most important forms of contemporary adaptability is the air transportation system. It's critical to preserve perseverance and resilience in the face of growing passenger and aviation traffic congestion. The resources and land availability support airport infrastructure. The rules advancing technology and practice are designed to preserve things like capacity, safety, and efficiency.

As a consequence, the “National Airspace System” (NAS) prioritizes reducing the negative impact that

improvisation has on the environment. Passengers may see their flight route, direction, altitude, and other relevant information while traveling thanks to contemporary technology. Nonetheless, air traffic controllers are always attempting to minimize Aeroplan arrival and departure delays.

Even though they worked in phases, the results are not what they wanted since there were hours-long delays that sometimes resulted in turmoil. A few significant factors that contribute to delays include carrier, weather, maintenance, and security. Travel for business purposes and tourism are the two main drivers of the aviation industry. which, by 2030, is anticipated to quadruple. Air traffic is predicted to rise in the same multiple as a consequence of this increase. It is possible to build additional airports in order to reduce air traffic congestion. But the level of complexity keeps rising rapidly. Therefore, modifying the current airports is the only method to reduce the delay. In light of the restricted quantity of land resources available, the latter option makes more sense. In essence, delay is the amount of time that an aircraft is delayed or canceled. It's probably commercial aviation.

II. LITERATURE REVIEW

The article provides a perceptive examination of European and Western nations' aircraft delay statistics. Data scientists combine domain knowledge with scientific skills in data learning to investigate various data aspects related to aircrafts like sequential (month, season, time of day, day of week), scheduling (flight timetable, flight preparation), spatial (province airfield, urban), climate (upper limit, Visibility, Surface climate, Convective climate), operations (claim, capacity), features (Airport transportation, Airline Status, plane tenancy, plane model, price) [1]. For the San Francisco–New York route, a new technique is created in which daily airfares are made available online. Two criteria are considered by the model: the total travel days and journey percentage which takes place both on weekdays and weekends [2].

The research analyzes flight data to predict, based on support vector machines. It builds and chooses feature vectors, such as planned arrival time period,

planned arrival in peak time period, planned flight time, planned departure time in day or night, the planned departure time in peak time period, and the passenger or cargo flight, by selecting one day's domestic flight plan and flight operation data. Next, it examines the information on the flight type, actual arrival time, and departure and arrival times on the flight plan (passenger flight or cargo flight). Furthermore, a prediction model is established. [3].

It makes use of a machine learning method known as logistic regression to forecast Aeroplane delays. Using a variety of independent characteristics, this method trains a model. We used the "Microsoft Azure Learning Studio" platform to develop the algorithm. To ascertain the effect of weather conditions on delays of flights and improve the prediction for real-world situations, we additionally used a weather dataset and connected it with the airport information at the corresponding locations. Thirty percent of the dataset is used for testing after the model has been trained on seventy percent of it. [4].

The eight attributes of the airline dataset were as follows: source, destination, number of stoppages, airline type, departure and arrival times, and other data. Regression machine learning models, like LGBM Regressor, Decision Tree Regressor, Random Forest Regression Tree, were used by the authors to conduct prediction. The ensemble learning technique combines predictions from many algorithms to provide a forecast.. [5].

Social media and deep learning data-driven airline ticket pricing prediction algorithms are shown here. The variables influencing ticket prices are discussed by the authors together with the state of airline ticket pricing at the moment. A Random Forest constructs several decision trees in the training phase, with each tree's prediction derived from the mean of the classes. It often does well on a variety of problems, especially those containing attributes with non-linear correlations. [6].

Airlines need to anticipate airfare trends. It may not always be the case, but the scientific literature on business sector price prediction is predicated on the idea that the selected factors have a linear relationship. This is because the research uses biased conventional predictive approaches, such as linear regression. [7].

It suggested a method to estimate the average ticket price based on input and output pairs by integrating two databases, including macroeconomic data, and using machine learning techniques like support vector machines and XGBoost. With the latest R-squared performance measures, the framework attains a remarkable 0.869 accuracy [8].

The given paper demonstrates that customers may have an operator presented to them who will make

the most of their buying time. Using the partial least squares method, a model built by Supriya Rajankar made use of a small dataset of flights between Bombay and Delhi to anticipate flight fares using the ML algorithms. The used methods include linear regression, and K-Nearest-Neighbors (KNN) [9].

The ticket pricing plan has developed into a sophisticated set of laws, like numerical simulations that decide price marketing methods, since the airline firm was privatized. Studies have depicted that these principles are affected by different conditions, even if the majority of their details are still unknown [10].

In order to calculate the mean ticket price based on source and destination pairings, Tianyi Wang, et al., suggested a framework in this study that combines two databases with macroeconomic data and algos like XGBoost and support vector machines [11].

Due to a lack of knowledge, purchasing plane tickets is a common process where people find it challenging to reduce costs. It is difficult to determine how changes in purchase time translate into variations in predicted cost, even with historical data accessible for scrutiny of a recent addition to several travel reservation services. They created an agent that can optimize purchase time on behalf of clients in order to solve this issue. [12].

The publication offers a framework for examining various routes and/or carriers as well as broad trends in airline pricing behavior. By providing clients who possess essential information they need to choose the best time to purchase a ticket, the aim is to balance the wish to reduce expenses with any time limitations the customer may have [13].

They conducted research to test the theory that as airline market competition rises, pricing discrimination would also rise. They estimated disparities in pricing using marginally implied pricing of ticket limits, which carriers usually utilize to price discriminate [14].

The journal examines the mechanisms behind the dynamic price dispersion, or daily airfare fluctuations across ticket histories. Empirical analyses reveal that factors pertaining to demand attributes, for example income, populace, and the proportion of travelers on business, and competitive pressures resulting from the presence of low-cost carriers, have a significant impact on dynamic price dispersion. However, competition intensity has no such effect [15].

Regarding the instance of Ryanair, the low-cost leader in Europe, this chapter makes reference to two primary fields of literature on aviation: the analysis of the low-cost business model and the research of dynamic pricing approaches. This chapter seeks to identify the competitive and contextual elements, as well as the key components of Ryanair's business

model, within this framework. This chapter seeks to identify the competitive and contextual elements, as well as the key components of Ryanair's business model, within this framework [16].

The document suggests a hybrid approach for predicting airline departure delays called "Random Forest Regression and Maximal Information Coefficient" (RFR-MIC). Regarding consistency of information, RFR-MIC are intrinsically combined. It also focuses on using flight data from many air routes to estimate aircraft departure delays. [17].

The flight delay cost can be done by proper flight planning. The Air Traffic Control(ATC) can open more runways to avoid flight delay. [18].

In recent studies it was found that 20% of domestic flights in the USA are delayed and reach their destination 15 minutes late. Because of the intricate structure of the air transport system, it becomes challenging to predict an accurate model. [19].

The author tries to lessen the influence of the climate on flights by improving the Impacted Traffic Index(WITI) model. Newer model is 92% higher than the previous WITI model.[20].

The given paper proposes a big service for predicting the delay during the departure of the flight. The accuracy was found to be 0.87 in predicting the delay of flight departure.[21].

III DESCRIPTION

The Title of my dataset that I am using here Analysis of Flight Delay Data the Dataset I am using contains three datasets Airlines - 15 rows, 2 columns Airports - 323 rows, 7 columns Flights - 28821 rows, 23 columns.

The Airlines Columns are IATA Code and Airline the Airports Columns are IATA Code, Airport, City, State, Country, Latitude, Longitude.

The Flights Columns are Month, Day of month, Day of week, OP Unique Carrier, Tail num, Destination, Departure Delay, CRS Elapsed time, Distance, CRS Departure M, Departure time M, CRS arrival M, Temperature, Dew Point, Humidity, Wind, Wind Speed, Wind gust, Pressure, Condition, Sch Dep, Sch Arrival, TAXI Out.

IV. IMPLEMENTATION

This document serves as a technique for Analysis of flight Delay Data. Structural changes in the architecture of Analysis of flight Delay Data.

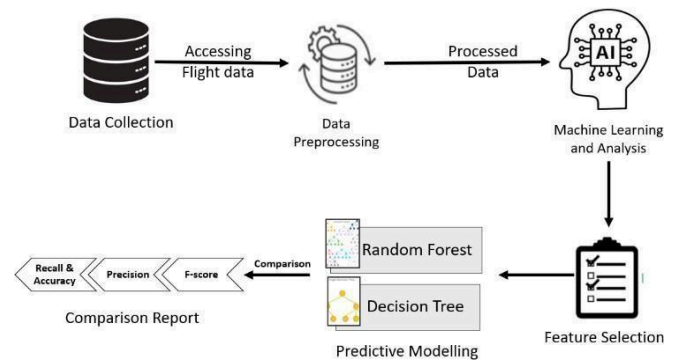


Fig 1. System Architecture

Figure 1 illustrates components of the system architecture, which are as follows:

1) **Data Collection:** For addressing research questions, test hypotheses, and assess results, data on variables of interest must be gathered and measured in a predetermined, methodical manner.

2) **Data Preprocessing:** It is the technique of transforming an unclean data collection from raw data. The dataset is preprocessed to check for missing values, inaccurate data, and other anomalies before it is put to the algorithm.

3) **Processed Data:** Information is the result of processing data. Data processing is the conversion of unprocessed data into usable data via a procedure. Data is processed to provide results that help solve an issue or make a situation better as it stands.

4) **Feature Selection:** It involves choosing a portion of pertinent features—variables and predictors.

5) **Random forest:** This popular ML approach, which blends the results of many decision trees to get result, is trademarked by Adele Cutler and Leo Breiman. Its popularity has been fueled by its adaptability, ease of use, and capacity to handle both classification and regression problems.

6) **F-measure or F-score** estimates the predictive performance.

7) **Precision:** The percentage of relevant occurrences among the recovered examples defines precision also known as the positive predictive value.

8) **Recall:** Also called the sensitivity is the fraction of retrieved relevant instances.

9) **Accuracy:** It is a measure of how often a ML model predicts a result correctly.

As part of data collection this data is collected based on airports, airlines and flights. The name of the selected airlines, airport and flights which I have taken is: i) Humidity ii) Temperature iii) Wind iv) pressure

The duration of flight delays increases when temperature rises within a given range, and the delay time deviates from the general data more readily.

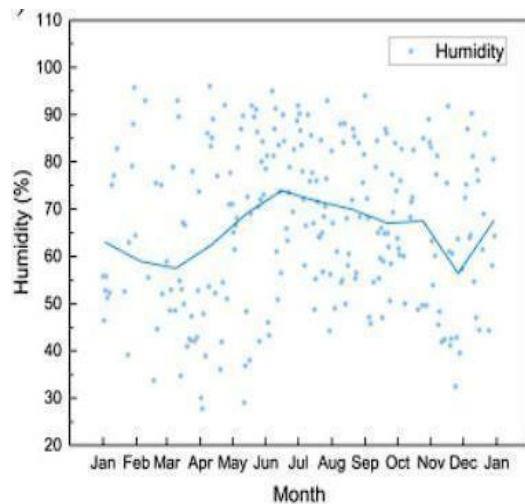


Fig 2. Humidity graph

Figure 2 illustrates how the scatter chart and broken line depict the changes in humidity over time. The data indicates a progressive increase from March to June and a gradual decrease from June to the following March. November's data is the lowest number for the whole year and deviates from the average.

The wind might reach 200 km/h or higher at flying altitude. It may seem that we are at hurricane levels. Nevertheless, the force of the wind on the aircraft is less than it would be at ground level because air density drops with height.

At cruising altitude, the cabin pressure is around 11-12 PSI (pounds per square inch), which is similar to the pressure we would find on a 6,000–8,000 feet mountain.

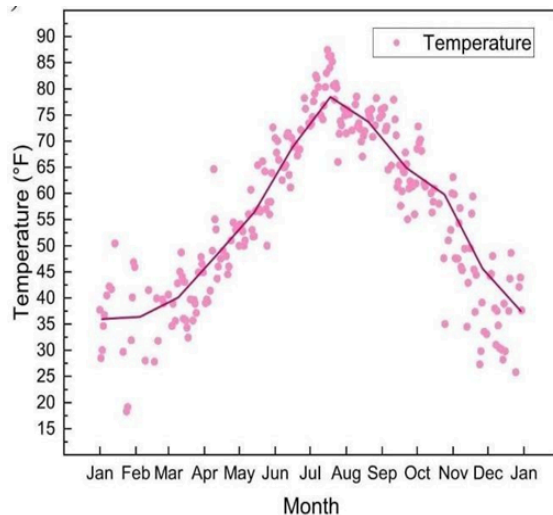


Fig 3. Temperature graph

The temperature peaks in July, climbs progressively from January to July, falls from July to December, and then falls considerably from November to December, as Fig. 3 illustrates.

	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

Table 1. Airport table

The Airport Table describes the International Air Transport Association [IATA CODE] ,Airports, City, State and Country. Which we have taken from Dataset.

Table 2. Airline table

	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

The Airline Table describes the International Air Transport Association [IATA CODE], Airline which we have taken from Dataset

Table 3. Flight table

	AIRLINE	Org_Airport_Name	Origin_city	Dest_Airport_Name	Destination_city
0	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle
1	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle
2	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle
3	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle
4	Alaska Airlines Inc.	Ted Stevens Anchorage International Airport	Anchorage	Seattle-Tacoma International Airport	Seattle

The Flight Table describes the airline, name of the

origin airport, and city, name of the airport destination and the city.

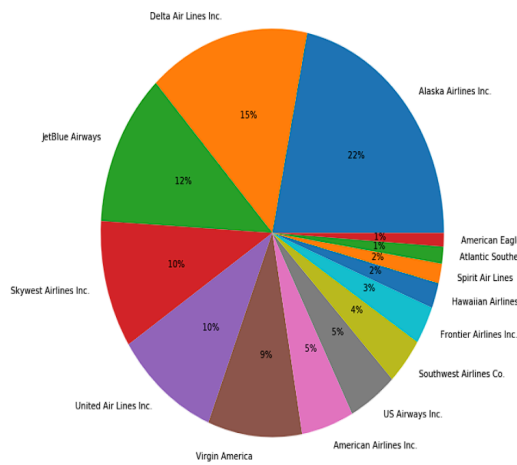


Fig 4. Pie chart of Airline

As shown in fig 4, the diagram visualizes the pairwise correlations between numerical variables in the 'Flights' Data Frame. The heatmap provides a quick and visual way to identify relationships between different features, with the annotated values indicating the strength and direction of these correlations

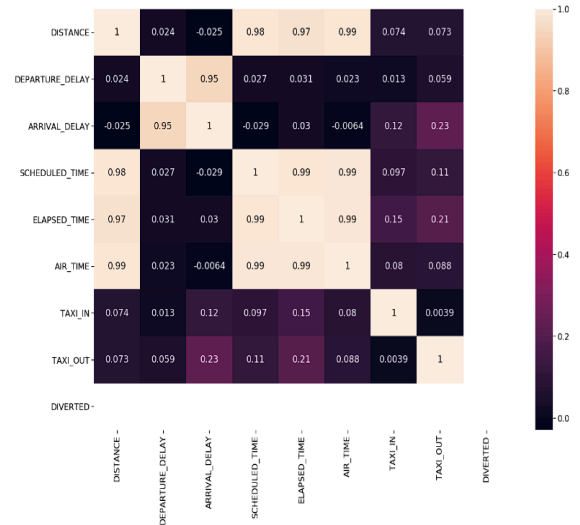


Fig 5. Heatmap

As shown in fig 5, we are creating a heatmap using the seaborn library to visualize the pairwise correlations between numerical variables in the 'Flights' Data Frame. The heatmap provides a quick and visual way to identify relationships between different features, with the annotated values indicating the strength and direction of these correlations

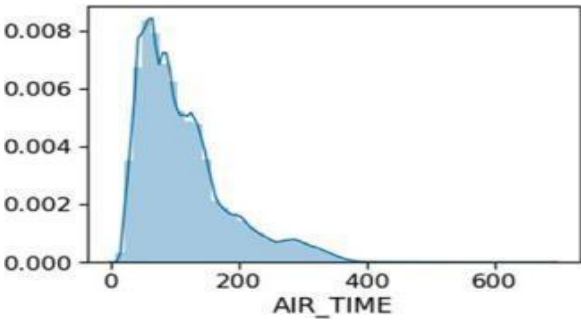


Fig 6. Distribution of air time

As shown in fig 6, the distribution plot to visualize the distribution of the 'AIR_TIME' variable in the 'Flights1' Data Frame. The plot helps you understand how the air time values are spread across different ranges and provides insights into the central tendency and variability of the 'AIR_TIME' variable in the dataset.

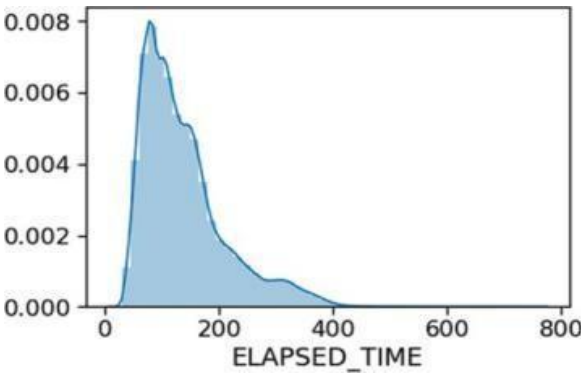


Fig 7. Distribution of elapsed time

V. COMPARATIVE ANALYSIS

Here the performances measures of existing works are given below:

TABLE 4. COMPARATIVE ANALYSIS OF EXISTING METHOD VS PROPOSED METHOD

Sl	Existing Method	Algorithm	Accuracy
1	M.Bardach (2020)	Random Forest	82.5%
2	S.Choi (2016)	KNN	61.69%
3	Proposed Work	Decision Tree	98%

Given that our suggested work has a 98% accuracy record, it should be able to assist us in making precise predictions for flight delay analysis.

Sampling approaches were used in S. Choi's work, which is why the kNN algorithm generated an accuracy of 61.69%. However, it was discovered that the accuracy was 82.42% if we used kNN without sampling approaches.

VI. CONCLUSION

Departure delays may arise from several factors such as airline delays, security delays, or airline system delays. Delays have a significant negative impact on the company's revenue, hence minimizing them is essential to raising airline industry profitability. Reducing the delays as much as possible can also significantly improve customer satisfaction.

REFERENCES

- [1] Fujun Wang Jun Bi Dongfan Xie Xiaomei Zhao
- [2] T. Janssen "A linear quantile mixed regression model for prediction of airline ticket prices"
- [3] Peng Hu, Jia Hou, Jianping Zhang, Qingyuan Yang, Lei Cao, Liwei Wang
- [4] Rahul Nigam, K. Govinda A. Kakaraparthi and V. Karthick, "A secure and cost-effective platform for employee management system using lightweight standalone framework over Diffie Hellman's key exchange algorithm," *ECS Trans.*, vol. 107, no. 1, pp. 13663–13674, Apr.2022.
- [5] Wikipedia "Mean Squared Error", Available: https://en.wikipedia.org/wiki/Mean_squared_error
- [6] G. Ataman and S. Kahraman, "Flight Prediction in Using Linear Regression and Artificial Neural Network Hybrid Models," *The Singapore Economic Review*. pp. 1–19, 2021. doi: 10.1142/s0217590821500521.
- [7] J. C. Lok, *Prediction Factors Influence Airline Fuel Price Changing Reasons*. 2018.
- [8] T. M. Vowles, "Airfare pricing determinants in hub-to-hub markets," *Journal of Transport Geography*, vol. 14, no. 1, pp. 15–22, 2006.
- [9] K. Rama-Murthy, "Modeling of United States airline fares—using the official airline guide (OAG) and airline origin and destination survey (DB1B)," Ph.D. dissertation, Virginia Tech, 2006.
- [10] Tianyi wang, Samira Pouyanfar, Haiman Tian and Yudong Tao "A Framework for airline price prediction: A machine learning approach"
- [11] William Groves and Maria Gini "An agent for optimizing airline ticket purchasing" in *proceedings of the 2013 international conference on autonomous agents and multi-agent systems*.
- [12] J. Santos Dominguez-Menchero, Javier Rivera and Emilio Torres Manzanera "Optimal purchase timing in the airline market".
- [13] J. Stavins, "Price discrimination in the airline market: The effect of market concentration," *Review of Economics and Statistics*, vol. 83, no. 1, pp. 200–202, 2001.
- [14] B. Mantin and B. Koo, "Dynamic price dispersion in airline markets," *Transportation Research Part E: Logistics and Transportation Review*, vol. 45, no.6, pp. 1020–1029, 2009.
- [15] P. Malighetti, S. Paleari, and R. Redondi, "Has Ryanair's pricing strategy changed over time? an empirical analysis of its 2006–2007 flights," *Tourism Management*, vol. 31, no. 1, pp. 36–44, 2010
- [16] Zhen Guo, Bin Yu, Mengyan Hao, Wensi Wang, "A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient"
- [17] M. Bardach, E. Gringinger, M. Schrefl and C. G. Schuetz, "Predicting Flight Delay Risk Using a Random Forest Classifier Based on Air Traffic Scenarios and Environmental Conditions," 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2020, pp. 1-8,
- [18] V. Natarajan, S. Meenakshisundaram, G. Balasubramanian and S. Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1081-1086, doi: 10.1109/CSCI46756.2018.00210.
- [19] K. Wang, J. Li and Y. Tian, "Airport Delay Prediction Method based on Improved Weather Impacted Traffic Index," 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Kunming, China, 2019, pp. 73-78
- [20] Z. Guo et al., "A Big Service with Network Represent Learning for Quantified Flight Delay Prediction," 2020 IEEE International Conference on Web Services (ICWS), Beijing, China, 2020, pp. 289-293, doi: 10.1109/ICWS49710.2020.00044.
- [21] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), Sacramento, CA, USA, 2016, pp. 1-6, doi: 10.1109/DASC.2016.7777956.