

FLIGHT DISRUPTION PREDICTION FOR KLIA

LEONG YI HONG

SESSION 2022/2023

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
JULY 2023

FLIGHT DISRUPTION PREDICTION FOR KLIA

BY

LEONG YI HONG

SESSION 2022/2023

THIS PROJECT REPORT IS
PREPARED FOR

FACULTY OF COMPUTING AND INFORMATICS
MULTIMEDIA UNIVERSITY
IN PARTIAL FULFILLMENT
FOR

BACHELOR OF COMPUTER SCIENCE
B.CS (HONS) DATA SCIENCE

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY


JULY 2023

Copyright of this report belongs to Universiti Telekom Sdn. Bhd. as qualified by Regulation 7.2 (c) of the Multimedia University Intellectual Property and Commercialisation Policy. No part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Universiti Telekom Sdn. Bhd. Due acknowledgement shall always be made of the use of any material contained in, or derived from, this report.

© 2023 Universiti Telekom Sdn. Bhd. ALL RIGHTS RESERVED.

DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.



Name of candidate: LEONG YI HONG
Faculty of Computing & Informatics
Multimedia University
Date: 14/7/2023

ACKNOWLEDGEMENT

I'd like to express my heartfelt gratitude to Prof. Ts. Dr. Ting Choo Yee, my Final Year Project supervisor, who has provided invaluable guidance, support, and encouragement throughout this project. His invaluable insights, expertise, and patience were critical in assisting me in completing this project. I am also grateful to a Malaysia Airlines employee who provided me with the flight datasets that served as the basis for this study. I'm also grateful to my colleagues and peers, who provided insightful feedback and suggestions throughout the project. Their assistance and advice were crucial in shaping the project's final outcome. I would like to extend my sincere thanks to everyone who has been a part of this project, for their unwavering support and for helping me make this project a reality.

ABSTRACT

Flight disruption has been a major issue for both the airlines and passengers. This research includes a comprehensive analysis on flight disruption prediction at Kuala Lumpur International Airport (KLIA). The research makes use of an extensive database that includes flight schedules, weather data, and foot traffic data, and conducts analysis using data mining, machine learning, and data visualization approaches. In this study, Feature Selection approach is attempted to extract a set of optimal features before creating predictive models to improve the prediction of flight disruption. The models investigated in this study were Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbors and Logistics Regression, allowing airlines to proactively manage and minimize their impact. This research helps to the creation of effective methods and tools for airlines to optimize their operations, decrease costs, and improve customer happiness by putting these program-specific skills and knowledge to use. The findings showed that optimal variables were flight departure time, airline, departure times of the day, arrival time, and flight duration, and the Random Forest classifier outperformed the rest of the models with an accuracy of 93.1%.

Contents

1	Introduction	8
1.1	Problem Statement	11
1.2	Objectives	14
1.3	Project Scope	15
1.4	Chapter Outline	17
2	Literature Review	20
2.1	Factors Causing Flight Disruption	21
2.2	Machine Learning Techniques for Flight Disruption Prediction . .	30
3	Method	37
3.1	Dataset	38
3.1.1	Flight Dataset	38
3.1.2	Weather Dataset	40
3.1.3	Foot Traffic Dataset	43
3.1.4	Flight Model Dataset	47
3.2	Data Preprocessing	51
3.2.1	Flight Dataset	51
3.2.2	Weather Dataset	58
3.2.3	Merging	61
3.3	Exploratory Data Analysis	66
3.4	Modelling	69
3.5	Prediction Visualization	74
4	Analysis and Findings	78
4.1	The Results of Exploratory Data Analysis	78
4.1.1	Airline Performance	78
4.1.2	Weather Influence	84
4.1.3	Busyness	88
4.1.4	Flight Characteristics	93
4.1.5	Evaluation of Findings	98
4.2	Feature Selection	100
4.3	Modelling Results	102
4.3.1	Modelling Results with Normal Data	102
4.3.2	Modelling Results with SMOTE Data	103

4.3.3	Modelling Result Summary	104
4.4	Predictive Visualization	106
5	Conclusion	113
	References	118

List of Tables

2.1	Factors Causing Flight Disruption by Other Authors	22
2.2	Methods Used By Other Authors to Solve Flight Disruption	32
3.1	Flight Dataset Features	39
3.2	Weather Dataset Features	42
3.3	Foot Traffic Dataset Features	45
3.4	Flight Model Dataset Features	49
3.5	Flight Dataset Feature After Pre-Processing	57
3.6	Weather Dataset Feature After Pre-Processing	60
3.7	Final Dataset After Merging	64
4.1	Top 30 Features	100
4.2	Result using Top 10 Features and Normal Data	102
4.3	Result using Top 30 Features and Normal Data	103
4.4	Result using Top 10 Features and SMOTE data	103
4.5	Result using Top 30 Features and SMOTE data	104

List of Figures

3.1	Flow of Methodology	37
3.2	Weather Dataset Download	41
3.3	Foot Traffic Dataset Download	44
3.4	Foot Model Dataset Download	48
3.5	Flight Dataset Pre-Processing	52
3.6	Weather Dataset Pre-Processing	59
3.7	Merging Process	61
3.8	Exploratory Data Analysis	67
3.9	Modelling Process	70
3.10	Count Plot before performing SMOTE	71
3.11	Count Plot after performing SMOTE	71
3.12	Upload Data Streamlit	75
4.1	Airline Performance Dashboard	79
4.2	Top 10 Airlines with the Highest Count of Delays	79
4.3	Top 10 Airlines with the Highest Total Delayed Time	80
4.4	Count of Delay and Sum Of Delayed time by Airline	81
4.5	Proportion of Delay Flight	82
4.6	Delay information	83
4.7	Weather Influence Dashboard	84
4.8	Departure Humidity vs. Delay Count	85
4.9	Departure Cloud Cover vs. Delay Count	85
4.10	Departure Temperature vs. Delay Count	86
4.11	Departure Wind Gusts vs. Delay Count	87
4.12	Busyness Dashboard	88
4.13	Map and Number of Delays by Departure Airport	89
4.14	Number of Delays by Departure Busyness	90
4.15	Departure Intensity Status and Proportion of Delayed Flights	91
4.16	Departure Intensity Status and Proportion of Delayed Flights	92
4.17	Flight Dashboard	93
4.18	Count of Delayed Flights by Flight Types	94
4.19	Average Delayed Time by Flight Type	94
4.20	Number of Delays by Aircraft Model	95
4.21	Count of Delayed Flights by Duration of Flight	96
4.22	Number of Delays by Departure Time	97
4.23	Upload Data Streamlit	106

4.24	Prediction Process Streamlit	107
4.25	Selection of Best Model Streamlit	108
4.26	Power BI in Streamlit	108
4.27	Prediction Dashboard Page 1	109
4.28	Prediction Dashboard Page 2	110
4.29	Prediction Dashboard Page 3	111

1 Introduction

Flight disruption is a significant issue faced by airlines worldwide, resulting in substantial economic losses for airline companies and inconvenience for passengers (Gui et al., 2020). Flight disruptions, defined as flights departing or arriving later than scheduled, can disrupt travel plans and evoke negative emotions among passengers. Various factors contribute to passenger disturbance, including the duration of flight delays, departure time in the schedule, the current time of delay, boarding gate crowd density, airline service quality, and management ability (Gu & Yang, 2019).

Flight disruptions have a significant financial and reputational impact on airlines, including costs related to fuel usage, airport fees, passenger compensation, and overtime pay (Yazdi, Kamel, Chabok, & Kheirabadi, 2020). These disruptions can arise from factors such as air traffic, weather conditions, aircraft issues, airport infrastructure, and unforeseen events. Although some disruptions are unavoidable, their occurrence can be minimized through effective prediction and proactive management.

As global COVID-19 restrictions ease, air travel is expected to increase, resulting in a higher probability of flight disruptions. China, for example, was one

of the first aviation markets heavily affected by the pandemic but has gradually recovered as the situation improved within mainland China (Czerby, Fu, Zheng, & Tae H., 2021). With the relaxation of quarantine requirements and other regulations, air travel is predicted to surge, leading to a corresponding increase in flight disruptions. Therefore, having a flight disruption prediction tool becomes crucial for airlines to anticipate and mitigate these issues before they escalate. Such a prediction tool can assist in categorizing and predicting potential disruptions, aiding in their resolution and minimizing their impact on passengers.

This project aims to develop a Flight Disruption Prediction tool that can accurately forecast potential disruptions and assist airlines in making proactive decisions. Data visualization plays a crucial role in understanding and presenting the outcomes of the predictive models. Therefore, the project will employ Data Visualization Methods, utilizing tools such as Power BI, to visually represent the insights gained from the data analysis and prediction models.

The first phase of the project involves Exploratory Data Analysis (EDA) through data visualization techniques. By visualizing the flight dataset, weather dataset, and foot traffic dataset, the project aims to uncover patterns, trends, and correlations among the variables. Through interactive visualizations, such as line charts, scatter plots, heat maps, and geospatial maps, the project will provide a com-

prehensive understanding of the relationships between flight disruptions, weather conditions, passenger demand, and other relevant factors. This EDA phase will lay the foundation for identifying significant variables and formulating hypotheses for further analysis.

Once the EDA is complete, the project will proceed to construct predictive models using data mining techniques. These models will leverage machine learning algorithms, such as decision trees, random forests, support vector machines, naive bayes, logistics regression, and k nearest neighbours to predict flight disruptions based on the identified variables. The performance of these models will be evaluated using appropriate evaluation metrics, ensuring their accuracy and reliability. By combining the strengths of data mining techniques and data visualization methods, this project aims to provide a comprehensive and practical solution for flight disruption prediction. The utilization of Power BI as a visualization tool enhances the interpretability and usability of the results, allowing stakeholders to gain valuable insights and effectively manage flight disruptions in a dynamic and fast-paced aviation environment.

In addition to the predictive modelling aspect, a Prediction Visualization module will be implemented using Streamlit and Power BI. This module will enable users to upload sample datasets to Streamlit, where the dataset will be pro-

cessed and prepared for prediction. The pre-trained prediction model, stored in a GitHub repository, will be retrieved and used to generate accurate predictions for the datasets. The results, including accuracy, precision, recall, and F1-score, will be calculated using the best-performing model based on the F1-score comparison. The final output will consist of the predicted datasets, which will be updated to Google Drive for further analysis.

The flight disruption prediction tool, combined with the Prediction Visualization module, will provide airlines with valuable insights into potential disruptions, allowing them to take proactive measures to minimize their impact. By leveraging data visualization techniques and machine learning algorithms, this project aims to enhance the efficiency and effectiveness of airline operations and improve the overall passenger experience in the face of increasing air travel demand and potential disruptions.

1.1 Problem Statement

The issue of flight disruptions poses a significant challenge for stakeholders in the aviation industry, including airport terminals, airline companies, and passengers. In recent years, the frequency of disruptions has increased, particularly following the impact of the pandemic, resulting in not only passenger inconvenience but also

substantial financial and reputational consequences for airlines. Consequently, reducing the occurrence and severity of disruptions has become a critical area of research and emergency management in civil aviation transportation (Dhanawade, Deo, Khanna, & Deolekar, 2019).

A key challenge in addressing disruptions lies in identifying the essential variables that contribute to them, encompassing weather conditions, passenger behaviour, technical errors, and other factors that ultimately lead to disruptions. Understanding these variables is paramount as it enables airlines to implement preventive measures aimed at minimizing disruptions. For instance, if an airline recognizes that a particular aircraft type is more susceptible to technical issues, it can take proactive steps to mitigate disruptions caused by that specific aircraft. Additionally, modelling airport delays based on weather and scheduled flight data can enhance system performance (K. Wang, Li, & Tian, 2019). Moreover, ensuring the cleanliness and accuracy of the data, along with conducting regular checks for errors or outliers, is imperative. Utilizing historical data also provides valuable insights for developing a robust predictive model.

Another challenge lies in constructing an effective predictive model for flight disruption prediction. Such a model needs to consider various factors, including past flight data, weather patterns, foot traffic information, and other relevant

data, to accurately predict disruptions. The model should be capable of analyzing the data and identifying patterns or trends indicative of a high likelihood of disruptions. Moreover, it should account for diverse variables that contribute to disruptions and provide an overall prediction of disruption likelihood. Regularly updating the model with fresh data and conducting routine testing are essential practices. Additionally, having a clear understanding of the model's assumptions, limitations, strengths, and weaknesses is crucial to identify potential inconsistencies and undertaking appropriate measures to address them.

To add on, the challenge lies in uncovering the hidden relationships between the multiple factors that contribute to flight disruptions. Flight disruptions, including delays, cancellations, and diversions, have a significant impact on either airlines or passengers. However, a good understanding of the connection and dependencies between the causes of these disruptions is still a complex task. By investigating variables such as weather conditions, foot traffic conditions, aircraft issues, airline issues, airport issues and unforeseen events, It aims to understand the underlying dynamic of flight disruptions. This will help airlines to implement countermeasures and help airlines or airports to effectively manage and minimise flight disruptions, resulting in smoother operations and better experiences for passengers.

In conclusion, flight disruptions represent a significant concern for all stakeholders in the aviation industry. Addressing this issue involves identifying critical factors, developing a predictive model, and optimizing its parameters. By understanding the causes and consequences of flight disruptions, airlines can enhance their ability to anticipate, prevent, and effectively manage disruptions. This research aims to contribute to the field of flight disruption prediction by developing a Flight Disruption Prediction tool specifically for Kuala Lumpur International Airport.

1.2 Objectives

The objectives are shown below:

- To identify the variables contributing to flight disruption in Kuala Lumpur International Airport.
- To uncover the hidden relationships between variables.
- To construct a predictive model to predict flight disruption in Kuala Lumpur International Airport.

1.3 Project Scope

The primary goal of this project is to develop accurate flight disruption prediction tools by analyzing a comprehensive dataset provided by an airline company. Flight disruptions, characterized by delays or cancellations, can have significant financial and operational implications for airlines. By leveraging advanced data analysis techniques and machine learning algorithms, this project aims to predict potential disruptions and assist airlines in making proactive decisions to minimize their impact.

In order to achieve this goal, multiple data sources will be utilized, including flight data, weather datasets, and foot traffic information. These diverse datasets will provide a holistic view of the factors that contribute to flight disruptions. The Python programming language will be used as the primary tool for data processing, statistical analysis, and predictive modelling, given its versatility and extensive libraries for data manipulation and machine learning.

The project will employ data mining techniques such as SMOTE (Synthetic Minority Over-sampling Technique) to address any imbalances in the dataset and ensure accurate predictions. Feature extraction will be performed to identify the most relevant variables that contribute to flight disruptions. Machine learning algorithms, including decision trees, random forests, support vector machines, naive

bayes, logistics regression and k nearest neighbours will be trained on the data to create predictive models.

Exploratory data analysis and statistical methods will be conducted to uncover hidden relationships and validate the performance of the predictive models. This phase will involve visualizing the datasets using tools such as Power BI, enabling interactive and intuitive data exploration. By visualizing the relationships between flight disruptions, weather conditions, passenger demand, and other relevant factors, valuable insights can be gained to inform decision-making processes.

The predictive models and visualizations developed in this project will provide airlines with actionable insights to improve operational efficiency, resource allocation, and overall passenger experience. By accurately forecasting potential disruptions, airlines can take proactive measures such as adjusting schedules, optimizing crew allocation, and communicating timely information to passengers. Ultimately, the aim is to enhance the overall performance and reliability of airline operations, leading to improved customer satisfaction and reduced financial losses due to flight disruptions.

1.4 Chapter Outline

The first chapter of this research report provides an introduction to the topic of flight disruption prediction at Kuala Lumpur International Airport. It begins by outlining the purpose and significance of the research, which is to study the causes and consequences of flight disruptions at Kuala Lumpur International Airport and develop a predictive model to forecast potential disruptions. This chapter sets the stage for the subsequent chapters by highlighting the relevance of the research to Kuala Lumpur International Airport and the broader aviation industry.

The second chapter of the report conducts a comprehensive literature review of relevant studies on flight disruption prediction. This chapter explores previous research on the causes and consequences of flight disruptions, as well as the methods and techniques employed for predicting disruptions. By reviewing existing literature, this chapter identifies gaps in current research and emphasizes the need for further investigation in this area.

The third chapter delves into the theoretical framework that underpins the development of a flight disruption prediction tool. It explores key concepts, theories, and models related to flight disruptions and predictive modelling. By examining the causes of flight disruptions and understanding the methodologies employed in predictive modelling, this chapter identifies the variables contributing to disrup-

tions at Kuala Lumpur International Airport and lays the groundwork for subsequent analysis.

The fourth chapter of the report outlines the research methodology employed in this study. This chapter explains the data sources and tools utilized for data collection and analysis, as well as the procedures and techniques employed to conduct the research. It also justifies the chosen methodology and how it effectively addresses the research questions.

The fifth chapter focuses on the implementation of the flight disruption prediction system at Kuala Lumpur International Airport. This chapter presents a detailed plan for putting the system into action during FYP2. It outlines the necessary steps, resources, and timeline for implementation, taking into consideration the practicality and feasibility of the proposed solution. Additionally, this chapter discusses potential challenges and risks associated with implementation and outlines strategies to address them effectively.

The sixth chapter evaluates the findings of the research and their implications for Kuala Lumpur International Airport and the aviation industry. It summarizes the main results and insights derived from the analysis of flight disruption data and the predictive model. This chapter also highlights the contributions of the research to the field of flight disruption prediction and suggests avenues for future research.

The final chapter serves as the conclusion of the research report. It provides a concise summary of the entire study, emphasizing the achievement of research objectives and the significance of the findings. This chapter also discusses the practical implications of the developed flight disruption prediction tool for Kuala Lumpur International Airport and the broader aviation industry. Additionally, it suggests potential areas for future research to further enhance the accuracy and applicability of flight disruption prediction.

2 Literature Review

The increasing demands of customers in the aviation industry have a profound influence on the number of flights and the likelihood of flight disruptions worldwide. Over the past century, airlines have transformed from simple postal carriers to complex and fascinating businesses (Almaameri & Mohammed, 2022). Consequently, any delays that occur during flight schedules can have significant disruptive effects on all flight operations.

The workers of the airport, the flight crew, the passengers of the aircraft, and the airports themselves will all be affected by the flight disruptions, the financial losses will affect all the involved parties. Apart from that, due to flight disruptions, there will be a lot of unnecessary usage of airport resources, and jet fuel combustion will also cause damage to the airports and the environment. A flight disruption application is needed to be implemented to avoid such things from happening. An accurate flight disruption prediction system will be able to optimize airport operations, increasing overall customer satisfaction and providing smoother processes to all travellers. According to Li and Jing (2021), by using a high-accuracy prediction model is a good way to mitigate delays, reduce economic costs for airlines and minimise passenger dissatisfaction.

2.1 Factors Causing Flight Disruption

Various factors will cause flight disruptions, some of the factors are the weather, aircraft, air traffic and aircraft issues (Gu & Yang, 2019). The unexpected occurrences of flight will introduce a vital challenge for stakeholders in the aviation industry. To add on, The duration of delay can vary, which will cause a ripple effect on the subsequent flights, potentially leading to more flights being missed and causing further delays.

Moreover, the departure and arrival time is also an important factor to be considered. Flights that have been scheduled during busy hours are always affected by flight disruptions due to the increase in air traffic in the air and aircraft congestion. Therefore, there is a need to have a flight schedule which is carefully planned and can be managed effectively to reduce flight disruptions.

Furthermore, the crowd density will also increase the flight disruptions rate. The insufficient resources to accommodate passengers' volume will bring potential delays to the boarding process, security screening, and check-in procedures. Therefore, a smooth management of boarding procedures and usage of resources will be able to reduce the flight disruptions caused by crowd density.

In addition, the quality of airline service and management ability play a significant role. Airlines with robust management systems and efficient operations are

better equipped to handle disruptions and minimize their impact on passengers. Prompt communication, effective contingency plans, and alternative solutions for affected passengers are vital components of efficient management.

It is important to note that the propagation of delays can have significant consequences if not properly controlled. If delays are not effectively managed, they can spread and impact other flights, resulting in a domino effect of disruptions (H. Zhou, Li, Jiang, Cai, & Xue, 2022). Therefore, proactive measures and effective strategies are needed to contain and minimize the impact of disruptions. By understanding and addressing the factors that contribute to flight disruptions, stakeholders can work towards reducing their occurrence and mitigating their impact, leading to smoother operations and improved customer experiences in the aviation industry.

Table 2.1: Factors Causing Flight Disruption by Other Authors

Author	Weather	Airport	Airline	Network	Security	Aviation Control	Passenger	Time	Passenger Load	Aircraft
(Li & Jing, 2021)	X			X						
(H. Zhou et al., 2022)	X	X	X							
(Y. Wang & Li, 2020)	X		X		X	X	X			

(F. Zhou, Jiang, Lu, & Wang, 2022)				X						
(Yazdi et al., 2020)		X						X	X	
(Cai, Li, Fang, & Zhu, 2021)				X						
(Jiang, Liu, Liu, & Song, 2020)	X									
(Kalyani et al., 2020)	X									
(Khaksar & Sheikholeslami, 2017)	X									X
(Shu, 2021)	X		X		X	X				
(Yanying, Mo, & Haifeng, 2019)		X	X					X		
(Jiang, Miao, Zhang, & Le, 2020)	X	X						X		X
(J. Wang & Pan, 2022)								X		
(Hu, Zhang, & Li, 2021)								X		
(Liu, Sun, Liu, Yang, & Gui, 2020)	X	X		X				X		
(T. Wang, Lin, & Gao, 2021)	X									
(K. Wang et al., 2019)	X									
(Mang & Chen, 2020)	X									
(Gui et al., 2020)	X	X						X		
(Anees & Huang, 2021)	X		X		X	X				

(Almaameri & Mohammed, 2022)	X									
(H. Wang, 2022)	X		X					X		
(Wu, Cai, Yan, & Li, 2019)		X	X							X
(Yiu, Ng, Kwok, Tung Lee, & Mo, 2021)	X									

Weather conditions are among the primary factors contributing to flight disruptions due to their inherent variability. Various weather elements, such as strong winds, thunderstorms, fog, and heat waves, can significantly impact flight operations. In fact, the weather has been identified as a crucial parameter influencing flight delays, making it an essential aspect for predicting and managing flight disruptions (Kalyani et al., 2020).

Strong winds pose challenges for aircraft during takeoff and landing, as they can deflect planes off course and compromise pilot control. Particularly, crosswinds can cause planes to drift sideways during landing, posing risks, especially if runways are insufficient in length. A notable example is the transport disruption caused by Storm Eunice in February 2022, resulting in numerous flight cancellations and delays across the UK (Beth, 2022). Similarly, at Japan's Narita International Airport, multiple passenger planes had to abort landings due to strong

crosswinds, leading to subsequent delays and diversions (Keoni, 2023).

Thunderstorms present risks to aircraft due to potential damage to the electrical system and reduced visibility caused by heavy rain. Turbulence associated with thunderstorms can also make the flight uncomfortable and unsafe for crew members. As a precautionary measure, flight crews actively avoid areas affected by thunderstorms, which can inevitably lead to flight disruptions. Proper planning and coordination are crucial in mitigating the impact of lightning and high winds, ensuring the safety and productivity of aviation workers (Schuldt, Nicholson, Adams, & Delorit, 2021). Furthermore, research by Borsky and Unterberger (2019) reveals that flights experiencing weather shocks can face additional delays of up to 23 minutes, depending on the type and intensity of the weather event.

Fog presents visibility challenges, impeding pilots' ability to see the runway clearly and potentially causing flight disruptions. In cases of low visibility, pilots may need to rely on specialized instruments or divert to airports with better visibility conditions. However, the diversion of aircraft to unfamiliar airports can create a negative experience for passengers and crew members, resulting in flight disruptions. Conditions like fog, dust, smoke, and haze commonly encountered in low-visibility situations pose significant problems for aviation operations. Dense fog, in particular, has detrimental effects on flight operations, including delays,

cancellations, and diversions, leading to severe impacts on passengers and substantial economic losses for airlines (Kulkarni et al., 2019). Moreover, overly pessimistic visibility predictions can result in an unnecessary reduction in airport capacity (Dietz, Kneringer, Mayr, & Zeileis, 2018).

Apart from weather conditions, the intensity of passenger traffic at airports is a crucial factor to consider when analyzing flight disruptions. Several factors, such as security delays, baggage handling, passenger check-in, immigration and customs processes, airport infrastructure, and airport capacity, are closely linked to the volume of passengers at airports. Urban airport areas, characterized by complex flows of people, materials, and aircraft, present various risk factors that can contribute to flight delays during operations (Y. Wang & Li, 2020).

Foot traffic emerges as an intriguing factor to investigate flight disruptions. High passenger traffic density can lead to challenges in security checks, check-in procedures, immigration processes, and more, ultimately causing disruptions for airlines and airports. Security checks, although essential for ensuring passenger and personnel safety, can result in disruptions if a large number of individuals attempt to go through security simultaneously. Issues with the security screening process or excessive foot traffic within the airport can exacerbate these disruptions. The increased foot traffic can also inconvenience passengers, forcing them

to spend more time on security screening than anticipated. Implementing effective strategic management of airport demand and capacity becomes essential in reducing congestion and delays (Xu, Wang, Wang, & Delahaye, 2022).

Furthermore, delays in the boarding process may occur when lengthy lineups form at check-in desks. Insufficient check-in counters to handle the passenger volume or procedural issues can contribute to such delays. Prolonged waiting times at the check-in counter cause passenger inconvenience and increase the likelihood of flight disruptions. Delays at immigration and customs checkpoints can also result in missed connecting flights or late arrivals at final destinations. Insufficient immigration and customs officials or complications with passengers' documents can contribute to these delays. Security breaches may lead to terminal clearance and subsequent re-boarding delays (Anees & Huang, 2021).

Late passengers are particularly prone to causing flight disruptions, requiring airport personnel to allocate valuable time and resources to accommodate them. Therefore, investigating foot traffic data at airports is crucial for understanding and managing flight disruptions effectively.

Flight disruptions can be influenced by various factors related to airport operations. One significant factor is air traffic control (ATC) delays. The congestion or issues within the air traffic control will contribute to the flight delay and di-

versions of aircraft, affecting flight operations and causing flight disruptions. For example, runway maintenance or closure will lead to flight disruptions, and the unscheduled work that closes the airport runways will reduce the capacity of the airport, and will potentially delay the flight from taking off or landing contributing to flight disruptions. The usage of ground handling services is also very vital to reduce flight disruptions, with an effect and adequate ground handling services such as aircraft servicing and baggage handling will smoothen the departure and arrival process, and minimise the risks of delays.

One of the factors that cause flight disruptions is the airlines themselves. An important point to be aware of is aircraft maintenance. The sudden breakdown of an aircraft will lead to flight disruptions, and the maintenance or repair of the aircraft will cause some to occupy the schedule of the aircraft can will leads to flight cancellation and flight delays. Therefore, regular maintenance and inspection are essential to reduce the breakdown of an aircraft and reduce flight disruptions. Crew members are also a factor contributing to flight disruption, unexpected sick leaves, crew fatigue and labour disputes will result in flight disruption. Additionally, the scheduling and planning of flight schedules, turnaround time between flights, and aircraft utilization will contribute to disruptions in flight operations. Moreover, coordination issues or disruptions within airline alliances or codeshare arrangements

can impact the operations of partner airlines, potentially causing flight disruptions.

Passenger-related factors can also contribute to flight disruptions. Overbooking flights beyond available capacity or miscalculating passenger loads can result in conflicts and disruptions when passengers are denied boarding or rearranged onto alternative flights. Security concerns, such as unruly passengers or suspicious items, can lead to flight delays or diversions as necessary security measures are taken. Non-compliant or disruptive passenger behaviour, including failure to follow safety instructions or causing disturbances during the flight, can also result in delays or diversions for the safety and security of all passengers and crew.

In conclusion, passenger intensity at airports, foot traffic, airport operations, airline-related aspects, network dynamics, and passenger loads will impact flight operations and cause flight disruptions. With a good strategy to manage the airport demand, the flow of passengers is vital to reduce the congestion of passengers, reducing delays and reducing the impact of flight disruptions on airlines, airports and passengers. Stakeholders in the aviation industry can work towards reducing disruptions and improving overall operations by addressing these factors.

2.2 Machine Learning Techniques for Flight Disruption Prediction

Flight disruption is a concern for airlines, passengers and airport operators causing impacts to the economy and causing inconvenience. Therefore, the interest to develop applications that can predict and prevent flight disruption is growing. A lot of approaches have been proposed in the literature, covering statistical modelling, machine learning and optimization techniques. These methods aim to solve flight disruptions by using data from different sources such as flight and weather data.

Nowadays, researchers have made impressive progress in developing a model that is capable of predicting flight disruptions. Machine-Learning techniques used to analyze and interpret large volumes of data are employed in these models. By using information from various sources, such as flight historical data, weather data, and other relevant information, the model can predict potential flight disruption with improved accuracy. For example, the patterns and relationships between weather, flight routes and delays can be identified by using machine learning algorithms and making sure that countermeasure has been taken to prevent flight disruptions.

Additionally, the key factors contributing to flight disruptions have been explored by the researchers and identified by the researchers. This involves analyz-

ing different aspects such as weather, airport infrastructure, foot traffic, scheduling of crew and air traffic control. With a good understanding of the causes of flight disruptions, mitigation strategies to prevent flight disruptions can be developed. For example, optimization techniques can be used to optimize the crew schedules, minimise the delays and improve the efficiency.

All in all, It draws attention to the value of statistical modelling, machine learning, and optimisation methods for predicting and limiting disruptions. It also highlights the significance of integrating data from diverse sources and identifying major contributing elements to improve disruption management tactics. The aviation industry may improve operational reliability, reduce delays, and improve the overall passenger experience by expanding these approaches.

Table 2.2: Methods Used By Other Authors to Solve Flight Disruption

Author	Visualization	K Means	Bayesian Modelling	XGBoost	Logistic Regression	Linear Regression	Gradient Boosting Classifier	Multiplayer Perceptron	Decision Tree	Support Vector Machine	Random Forest	Arima Model	Neural Network	Causal Analysis	Levenburg-Marquat Algorithm	Social Network Analysis
(Li & Jing, 2021)											X					
(H. Zhou et al., 2022)													X			
(Y. Wang & Li, 2020)																X
(F. Zhou et al., 2022)														X		
(Yazdi et al., 2020)															X	
(Cai et al., 2021)													X			
(Jiang, Liu, et al., 2020)								X	X	X	X					
(Chakrabarty, 2019)							X									
(Kalyani et al., 2020)				X		X										
(Khaksar & Sheikholeslami, 2017)		X	X						X		X					
(Shu, 2021)			X		X	X										
(Yanying et al., 2019)			X		X				X	X						
(Jiang, Miao, et al., 2020)													X			
(Tao, Man, & Yanling, 2021)									X							
(J. Wang & Pan, 2022)												X				
(Hu et al., 2021)											X					
(Liu et al., 2020)									X							
(T. Wang et al., 2021)				X		X										
(K. Wang et al., 2019)											X					
(Mang & Chen, 2020)				X							X					

(Balamurugan, Maria, Baranidaran, MaryGladence, & Revathy, 2022)			X		X					X		X					
(Gui et al., 2020)														X			
(Anees & Huang, 2021)												X					
(Almaameri & Mohammed, 2022)										X	X			X			
(H. Wang, 2022)	X																
(Ballakur & Arya, 2020)														X			
(Wu et al., 2019)												X					
(Meel, Singhal, Tanwar, & Saini, 2020)			X		X		X		X		X						
(Yiu et al., 2021)		X	X						X		X		X				
(Huo, Keung, Lee, Ng, & Li, 2020)		X	X		X				X		X						
(Hopane & Gatsheni, 2019)		X						X	X	X							
(Pamplona, Weigang, de Barros, Shiguemori, & Alves, 2018)														X			

The table above discussed the various techniques used by other authors to solve the flight disruption problem at Kuala Lumpur International Airport. One of the most widely used methods by authors is the development of prediction models. These models will be able to analyze various datasets and identify potential disruptions, allowing for a timely intervention proactively. By understanding the factors, the root causes of flight disruptions will be able to identify and appropriate solutions will be able to apply. These models provide valuable insights and forecasts and provide a good decision-making choice in managing flight disruption.

tions. Airport authorities can effectively allocate resources, minimize the effect of flight disruptions, and improve operational efficiency and reliability.

Numerous studies have been conducted by researchers to predict flight disruptions using various modelling approaches. Jiang, Liu, et al. (2020) developed Support Vector Machine, Decision Tree, Random Forest, and Multilayer Perceptron models for flight delay prediction using Aviation Big Data. Kalyani et al. (2020) applied XGBoost and linear regression algorithms to develop a predictive model for flight delays. Khaksar and Sheikholeslami (2017) utilized Bayesian modelling, decision tree, cluster classification, random forest, and hybrid methods to estimate delay occurrences and magnitudes in a network. Shu (2021) proposed several machine-learning models based on flight information from 2018, while Yanying et al. (2019) predicted flight cancellations using logistic regression, support vector machine, naive Bayes, and decision tree algorithms.

In similar fashion, Balamurugan et al. (2022) employed machine learning algorithms such as logistic regression, decision tree regression, Bayesian Ridge, random forest regression, and gradient boosting regression to anticipate flight delays, while Meel et al. (2020) used machine learning models including logistic regression, decision tree regression, Bayesian Ridge, random forest regression, and gradient boosting regression to predict flight arrival delays. Yiu et al. (2021) adopted

various machine learning approaches, such as decision trees, random forests, k-nearest neighbour, naive Bayes, and artificial neural networks, to predict flight delays. Huo et al. (2020) compared and analyzed the prediction results of several machine-learning approaches using real data from the Hong Kong International Airport. Lastly, Almaameri and Mohammed (2022) utilized machine learning algorithms, including decision trees, support vector machine, k-means clustering, and multi-layered perceptron, to construct flight departure delay prediction models.

On top of that, another method used by authors is Exploratory Data Analysis (EDA), which involves visualizing the data and identifying crucial variables and characteristics that may potentially contribute to flight disruptions. The EDA will be able to identify patterns, anomalies and trends within the data and the strategies can be taken to reduce the risks of flight disruption and enhance the airport operations. For example, EDA can reveal unusual data elements, such as a flight at midnight usually tends to have more disruptions. These approaches help to improve airport operations by optimising resource allocation, making informed decisions, and mitigating disruption risks. Airports may efficiently resolve interruptions, improve passenger experience, and assure smoother and more reliable operations by adopting these strategies.

The findings from these studies help develop predictive models for flight interruptions, providing essential information towards enhancing operational efficiency and passenger experience. The wide range of machine learning algorithms and modelling techniques used in these studies demonstrates the researchers' efforts to investigate numerous approaches to accurately predict flight delays and cancellations. By leveraging the power of data analysis and machine learning, the understanding and management of flight disruptions can be enhanced, leading to more efficient and reliable air travel operations.

3 Method

This chapter presents the method employed to address the problem of flight disruptions at Kuala Lumpur International Airport using a data mining approach. Furthermore, it provides an in-depth analysis of the approach and the implementation plan to be utilised into generating the findings that verify the concept and describes how the working model or simulations function.

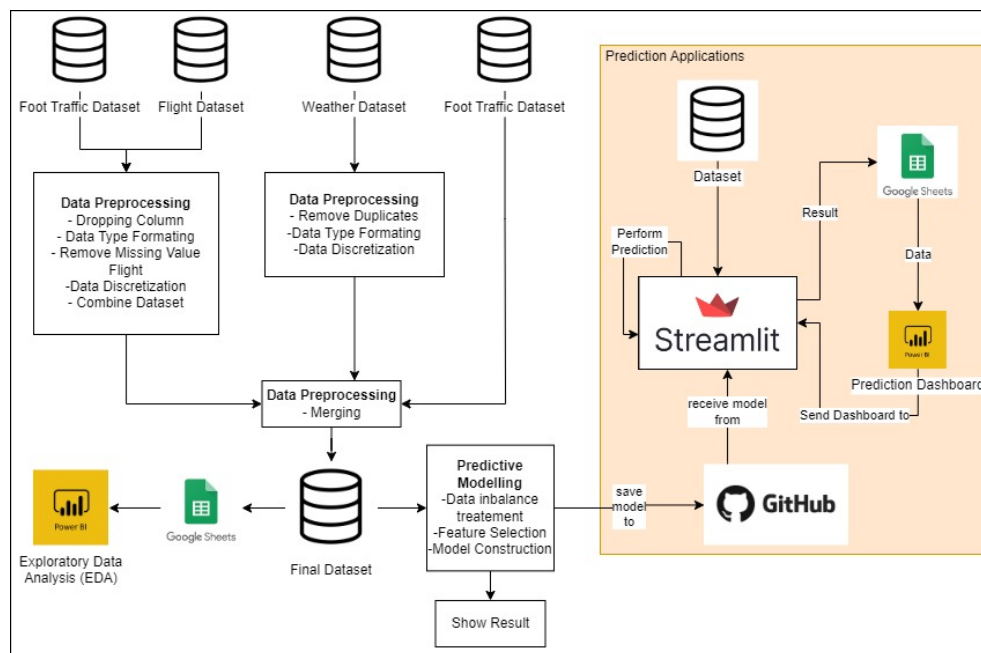


Figure 3.1: Flow of Methodology

3.1 Dataset

This research utilises three main datasets: the flight dataset, the weather dataset, and the foot traffic dataset, which provide crucial information for the analysis of flight disruptions at Kuala Lumpur International Airport.

3.1.1 Flight Dataset

The flight dataset is one of the main datasets that will be used throughout the research, the flight dataset contains comprehensive information about all the flights operating at Kuala Lumpur International Airport, and its nearby airports.

Each row in the dataset represents a unique flight, providing comprehensive insights into various aspects of flight operations. The dataset includes essential variables such as *airline_iata*, *airline_icao*, *flight_iata*, *flight_icao*, *flight_number*, departure airport (*Dep_iata*), departure time (*Dep_time*), arrival airport (*Arr_iata*), arrival time (*Arr_time*), code-sharing airline (*Cs_airline_iata*), code-sharing flight number (*Cs_flight_number*), code-sharing flight iata (*Cs_flight_iata*), flight status (*Status*), flight duration (*Duration*), and flight delay information (*Delayed*). The table below shows the example of data along with their features.

Table 3.1: Flight Dataset Features

<i>Airline_iata</i>	Indicates the IATA code of the airline operating the flight. (MH)
<i>Airline_icao</i>	Indicates the ICAO code of the airline operating the flight. (MAS)
<i>Flight_iata</i>	Indicates the IATA code of the flight. (MH194)
<i>Flight_icao</i>	Indicates the ICAO code of the flight. (MAS194)
<i>Flight_number</i>	Represents the flight number. (194)
<i>Dep_iata</i>	Indicates the IATA code of the departure airport. (KUL)
<i>Dep_time</i>	Represents the local departure time. (6/1/2023 19:55)
<i>Dep_time_utc</i>	Represents the UTC departure time. (6/1/2023 11:55)
<i>Arr_iata</i>	Indicates the IATA code of the arrival airport. (BOM)
<i>Arr_time</i>	Represents the local arrival time. (6/1/2023 22:35)
<i>Arr_time_utc</i>	Represents the UTC arrival time. (6/1/2023 17:05)
<i>Cs_airline_iata</i>	Indicates the IATA code of the codeshare airline. (NULL)
<i>Cs_flight_number</i>	Represents the codeshare flight number. (NULL)
<i>Cs_flight_iata</i>	Indicates the IATA code of the codeshare flight. (NULL)
<i>Status</i>	Represents the status of the flight. (Scheduled)
<i>Duration</i>	Represents the duration of the flight in minutes. (310)
<i>Delayed</i>	Indicates whether the flight was delayed or not. (NaN)

The Flight Dataset, provided by an employee working in the airline industry, forms a crucial component of this research. It contains 60,550 rows and encompasses 17 columns, representing a vast amount of detailed information about flights operating at Kuala Lumpur International Airport. These variables provide a comprehensive analysis of flight operations at Kuala Lumpur International Airport and they serve as an important indicator to identify the factors that are contributing to the flight disruptions.

On top of that, the historical data inside the flight dataset will be able to help with the development of the flight prediction model. As these models use the patterns and trends in the historical data to predict future disruptions, allowing the stakeholders to effectively manage the potential disruptions.

Overall, the flight dataset plays a critical role in understanding the dynamics of flight operations at Kuala Lumpur International Airport and provides the necessary information for identifying factors contributing to flight disruptions. It serves as the basis for developing predictive models that can aid in proactive decision-making and minimize the impact of disruptions on passengers and airport operations.

3.1.2 Weather Dataset

Apart from that, the weather dataset also provides valuable insights into understanding the impact of weather on flight disruption at Kuala Lumpur International Airport. The key variables included in this dataset include temperature, humidity, precipitation, cloud cover, patterns and specific weather condition. These variables will be useful in helping to understand how different weather factors affect flight disruption at Kuala Lumpur International Airport.

Algorithm 1 : Weather Dataset Download
<ol style="list-style-type: none"> 1. Read the Flight Data file. 2. Preprocess the Flight Data. 3. Group the Flight Data by Departure IATA Codes and Arrival IATA Codes. 4. Define the function 'getData()' to retrieve weather information. 5. Retrieve airport information (latitude and longitude) for each IATA code. 6. Loop through the Departure and Arrival IATA Codes. 7. Construct the request URL for weather data based on airport location and date range. 8. Send a request to the Weather Data API. 9. Retrieve the weather data. 10. Create a DataFrame to store the weather data. 11. Concatenate the weather DataFrames for Departure and Arrival. 12. Download the data 15. End the algorithm.

Figure 3.2: Weather Dataset Download

To obtain this dataset, several steps were undertaken. Firstly, the departure and arrival dates in the Flight Dataset were converted from string format to the standard date format to facilitate further processing. Subsequently, the flight data were grouped together based on the maximum and minimum dates to determine the latest and earliest dates available for each airport. This information was crucial for retrieving the corresponding weather datasets using the OpenMeteo API.

To crawl the weather data, an algorithm was created to iterate through each airport and retrieve the weather information for the specified dates. This algorithm utilized the latitude, longitude, latest date, and earliest date previously obtained.

By passing this information to the API, weather data in JSON format was retrieved for each airport.

The complete weather information for all 185 airports is subsequently appended into a single data frame. The final Weather Dataset comprises 14 features and contains a total of 237,408 rows, representing hourly weather observations for the selected airports. These observations enable us to analyze various key weather variables associated with flight operations. Some of the variables included in the dataset are:

Table 3.2: Weather Dataset Features

<i>Airport</i>	Indicates the airport location such as BKK (Bangkok).
<i>Time</i>	Specifies the date and time of the weather observation (e.g. 10-11-2022 8:20).
<i>Temperature</i>	reflects the temperature in degrees Celsius at the given hour.
<i>Humidity</i>	Indicates the relative humidity percentage at the given hour.
<i>SeaLevelPressure</i>	Represents the atmospheric pressure at sea level in hPa.
<i>Precipitation</i>	Specifies the amount of precipitation in millimetres.
<i>Rain</i>	Indicates whether rain was observed (0 for no rain 1 for rain).
<i>Snowfall</i>	Specifies the amount of snowfall in millimetres.
<i>WeatherCode</i>	Represents a numerical code indicating the weather conditions.
<i>CloudCover</i>	Reflects the percentage of cloud cover at a given hour.
<i>Windspeed</i>	Indicates the wind speed in kilometres per hour.
<i>WindDirection</i>	Represents the direction of the wind in degrees.
<i>WindGusts</i>	Specifies the speed of wind gusts in kilometres per hour.

The weather dataset, obtained from reliable sources via the Open Meteo API,

is made up of hourly weather observations for a variety of airports and provides a comprehensive view of weather conditions relevant to Kuala Lumpur International Airport. The importance of different weather will be able to discover after analyzing the dataset. This knowledge empowers decision-makers to develop effective strategies for managing weather-related disruptions, optimize flight schedules, and enhance overall operational efficiency.

Overall, a clear understanding of the relationship between weather conditions and flight disruptions is important to minimize flight disruptions and ensure a smooth and reliable travel experience for passengers. Airport operators, airlines, and other relevant entities can decide based on data, take appropriate approaches, and employ suitable approaches to reduce the impact of weather-related issues on flight operations at Kuala Lumpur International Airport by leveraging the insights provided by the weather dataset.

3.1.3 Foot Traffic Dataset

On top of that, the foot traffic dataset is an important dataset to be investigated as well as it provides valuable insights into the relationship between passenger demand and the flight disruptions in Kuala Lumpur International Airport. The foot traffic dataset was collected through the BestTimes API. The dataset includes

seven key variables or features, including the airport's name, longitude and latitude coordinates, the day of the week, the hour of observation, the intensity status, and the level of busyness. These variables enable a thorough examination of passenger traffic patterns and their impact on flight operations.

Algorithm 2 : Foot Traffic Dataset Download
<ol style="list-style-type: none">1. Start the program.2. Prompt the user to enter the Venue Name.3. Prompt the user to enter the Venue Address.4. Set the parameters for the API request, including the API key, venue name, and venue address.5. Send a POST request to the BestTime API to retrieve the forecast.6. Convert the response to JSON format.7. Extract the hour-by-hour analysis and day raw data from the JSON response.8. Create a dictionary to store the forecast data for each hour.9. Iterate through the days of the week.10. Iterate through the hours of the day.11. Populate the dictionary with venue name, day, hour, intensity level, intensity status, and busyness.12. Create a DataFrame from the dictionary.13. Save the main DataFrame as a CSV file using the venue name as the filename.14. End the program.

Figure 3.3: Foot Traffic Dataset Download

To acquire the Foot Traffic Dataset, a data collection process is performed. Firstly, an algorithm is created to download the foot traffic data manually. Using the list of airports generated during the weather data compilation process, each airport is searched individually on the BestTime.app website to obtain the corresponding venue name and venue address.

Once the venue name and venue address for an airport is obtained, an API is created to make requests and download the foot traffic dataset. The API utilizes the venue name and venue address as parameters to retrieve the JSON files containing the foot traffic data for that specific airport. The downloaded data is then appended and converted into a single data frame, which is subsequently saved as a file.

This process is repeated for all the available airports, resulting in individual foot traffic datasets for each airport. To consolidate the data into a comprehensive Foot Traffic Dataset, all the individual datasets are concatenated together. The final Foot Traffic Dataset comprises 19,656 rows of data, capturing foot traffic observations for different airports, days of the week, and hours.

The Foot Traffic Dataset includes six distinct features: 'Airport', 'Day', 'Hour', 'Intensity Status', 'Intensity Level', and 'Busyness'. These features provide insights into the foot traffic patterns and busyness levels at airports, enabling a better understanding of passenger demand and its potential impact on flight operations at Kuala Lumpur International Airport.

Table 3.3: Foot Traffic Dataset Features

Airport	Indicates the name or code of the airport (e.g., ADD)
Day	Specifies the day of the week for the foot traffic observation. (e.g., Monday)
Hour	Represents the hour of the day for the foot traffic observation. (ranging from 0 to 23)

Intensity Level	Reflects the intensity level of foot traffic, indicating the expected busyness at a given hour. (e.g., 0)
Intensity Status	Represents a qualitative description of the foot traffic intensity, (e.g., Average)
Busyness	Indicates the level of busyness (e.g., 45)

A lot of insights can be gained into how passenger demand influences the reliability of flights at Kuala Lumpur International Airport by monitoring the datasets. By identifying patterns and trends in passenger foot traffic, it is possible to learn essential information about the busiest and least busy times at the airport. A lot of effective decisions related to resource allocation, staffing, and scheduling to better accommodate passenger demand and optimize flight operations can be made by airport operators, airlines, and other stakeholders after investigating the dataset.

Furthermore, assessing the relative importance of passenger demand compared to other factors such as weather conditions or aircraft performance is crucial for understanding the complex dynamics that contribute to flight disruptions. A more comprehensive understanding of the interrelations between these variables can be obtained by studying the foot traffic dataset alongside the flight and weather datasets. This comprehensive analysis aids stakeholders in identifying potential areas for improvement, implementing targeted strategies to manage peak passenger demand, and ultimately improving the overall reliability and efficiency of flight

operations at Kuala Lumpur International Airport.

The Foot Traffic Dataset serves as a valuable resource for understanding passenger demand and its impact on flight operations at Kuala Lumpur International Airport. This dataset is obtained through BestTime.app API, which provides foot traffic data and forecasts the busyness of public businesses at different hours of the week. In this case, the API is utilized to collect foot traffic data specifically for airports.

3.1.4 Flight Model Dataset

Furthermore, the Flight Model dataset will be incorporated into the existing datasets in order to provide additional information and improve the analysis of flight disruptions at Kuala Lumpur International Airport. The key variables inside this dataset include aircraft model, turboprop information, and widebody aircraft information.

The foot traffic dataset was collected through the FlightStats API. The Flight Model Dataset is derived from the Flight Dataset by utilizing the flight number and airline IATA code as key identifiers. This dataset provides additional information about the flight models associated with each flight such as plane model, aircraft's turboprop information and aircraft's wide body information. It is important to note that not all flight models may be included in the dataset due to API constraints and

limitations with historical data availability.

Algorithm 3 : Flight Model Dataset Download
<ol style="list-style-type: none">1. Read the flight data from an Excel file into a DataFrame.2. Drop duplicate rows based on the combination of airline IATA code and flight number.3. Remove rows with missing values in the flight IATA column.4. Select only the columns 'airline_iata', 'flight_number', and 'flight_iata' from the DataFrame and reset the index.5. Initialize empty lists to store the aircraft model, turboProp status, and widebody status.6. Iterate over each row in the flight DataFrame.7. Retrieve the airline IATA code and flight number from the current row.8. Make an API request to retrieve the flight status using the airline IATA code, flight number, year, month, and day.9. Parse the JSON response.10. The resulting flight DataFrame contains the flight IATA, aircraft model, turboProp status, and widebody status.

Figure 3.4: Foot Model Dataset Download

To acquire the Flight Model Dataset, several steps are performed. Firstly, the Flight Dataset is read, and duplicate entries of airline IATA code and flight number are dropped to avoid repetition. Rows with missing flight names or number values are also dropped. Instead of using the flight time for data retrieval, a single day is chosen as the base to assume that the flight models for the same flight number remain consistent.

An algorithm is then created to loop through all the flights in the dataset. For each flight, the airline IATA code, flight number, and the base date are passed into the API to retrieve additional flight model information. The API returns data including the Model Name, Turbo Prop status, and Wide Body status. If the flight

model information is not available, a 'None' value is appended to the dataset for the corresponding flight.

To ensure a comprehensive dataset, the algorithm also loops through seven days to cover a wider range of flights. This accounts for cases where flights may only operate once or twice a week. By collecting data for multiple days, the dataset can capture a more complete representation of flight models.

The compiled data forms the Flight Model Dataset, which contains 1,503 unique flights, with each flight entry consisting of four features: *flight_iata*, *Aircraft_Model*, *Aircraft_isTurboProp*, and *Aircraft_isWidebody*. This dataset serves as a valuable resource for understanding the specific flight models associated with each flight, providing insights into the characteristics and attributes of the aircraft used in the operations.

Table 3.4: Flight Model Dataset Features

Flight_IATA	Represents the flight number (MH194)
Aircraft_Model	Indicates the specific model of the aircraft used for the flight, providing details about the type of aircraft. (A330-300)
Aircraft_isTurboProp	Specifies whether the aircraft for the flight is a Turbo Prop or not. (FALSE)
Aircraft_isWideBody	Indicates whether the aircraft for the flight is a Wide-body aircraft or not. (TRUE)

The usage of this dataset enables the exploration of relationships between dif-

ferent flight models and their impact on flight disruption. The potential factors leading to flight disruption can be identified by examining the performance of characteristics of different flight models such as technical issues, maintenance requirements, or operational considerations to certain aircraft types.

Finally, the integration of the Flight Model dataset, along with the current flight, weather, and foot traffic information, will contribute to a comprehensive analysis of flight disruptions at Kuala Lumpur International Airport. This research aims to develop predictive models, uncover underlying patterns, and provide actionable insights to improve operational strategies and the overall travel experience for passengers by leveraging the combined power of these datasets.

3.2 Data Preprocessing

The data preprocessing stage is critical in preparing the flight dataset for in-depth analysis and modelling. It entails a number of critical steps to ensure data quality, consistency, and relevance to the research objective.

3.2.1 Flight Dataset

In the data preprocessing phase for the flight dataset, several important steps are performed to ensure the data is properly prepared for analysis and modelling. Each step serves a specific purpose and contributes to the overall quality and suitability of the dataset.

The first step involves dropping certain columns from the flight dataset, such as *airline_iata*, *flight_icao*, *flight_number*, *cs_airline_iata*, *cs_flight_number*, *cs_flight_iata*, and *status*. These columns are excluded as they do not provide relevant information for the analysis and modelling of flight disruptions at Kuala Lumpur International Airport. By removing these columns, we focus only on the essential features that contribute to predicting flight disruptions accurately.

Next, any rows with missing departure time or arrival times are dropped from the dataset. This ensures that we have complete and reliable data for analysis,

Algorithm 4 : Flight Dataset Pre-Processing
<ol style="list-style-type: none"> 1. Read the flight data from an Excel file into a DataFrame. 2. Select the relevant columns from the DataFrame: 'flight_iata', 'airline_icao', 'dep_iata', 'dep_time', 'dep_time_utc', 'arr_iata', 'arr_time', 'arr_time_utc', 'duration', and 'delayed'. 3. Drop rows with missing values in the 'dep_time_utc' and 'arr_time_utc' columns. 4. Define two helper functions: 'roundHour' to round the time to the nearest hour and 'getHour' to extract the hour from a time. 5. Define additional helper functions: 'getDay' to get the day of the week and 'getWeekend' to determine if a day is a weekend day. 6. Extract the date and time components from the 'dep_time_utc' and 'dep_time_local' columns. 7. Round the time columns using the 'roundHour' function. 8. Extract the hour and day of the week from the rounded time and date columns. 9. Determine the part of the day (morning, afternoon, evening, night) based on the hour. 10. Determine if the day is a weekend day. 11. Define a function 'delay' to convert the 'delayed' column to a boolean column indicating if the flight was delayed. 12. Fill missing values in the 'delayed' column with 0 (no delay). 13. Apply the 'delay' function to convert the 'delayed' column to the 'delayStatus' column. 14. Load airport data using the 'airportsdata' module and store airport information for departure and arrival airports in separate lists. 15. Add columns for departure airport name, city, latitude, and longitude, as well as arrival airport name, city, latitude, and longitude. 16. Use the FlightRadar24 API to retrieve airline information based on the ICAO code and update the 'airline_icao' column to 'Airline' using a mapping. 17. Load the flight model data from a CSV file. 18. Merge the flight data with the flight model data based on the 'flight_iata' column. 19. Use the 'geopy' library to retrieve the country information for departure and arrival cities. 20. Create a mapping from city to country. 21. Map the city-to-country mapping to the departure and arrival cities in the flight data to obtain the departure and arrival countries. 22. Determine the flight type (domestic or international) based on whether the departure and arrival countries are the same.

Figure 3.5: Flight Dataset Pre-Processing

as flights with missing time information would not be useful for predicting disruptions accurately. By eliminating these incomplete records, we maintain data integrity and consistency throughout the dataset.

The format of the departure time columns (*dep_time_utc* and *dep_time_local*) is then transformed from string to DateTime format. This conversion allows us to perform accurate time-based calculations and operations on the data. By working with standardized datetime values, we can analyze flight patterns, identify temporal trends, and make precise predictions based on time-related factors.

Furthermore, the *dep_date_utc* and *dep_date_local* columns are extracted from the *dep_time_utc* and *dep_time_local* columns, respectively. These additional columns store the date information separately, allowing us to analyze flight disruptions based on specific dates and track temporal patterns that may influence disruptions.

To facilitate the analysis and grouping of flights, the *dep_time_utc* and *dep_time_local* columns are formatted to represent time only, excluding the date component. This simplifies the data by focusing on the temporal aspect of flights, enabling easier comparison and analysis of flight times across different days.

To capture the impact of rounded departure times, a new column called *dep_TimeRound_utc* and *dep_TimeRound_local* is created. This column stores the rounded hour data for each flight, where the precise departure time is rounded to the nearest hour. This rounding enables aggregation and grouping of flights based on common time intervals, providing insights into the overall distribution of flight disruptions within specific hour ranges. These columns are created in order to facilitate the merging with other datasets later on.

Additionally, new columns are introduced to extract hour and day information from the departure time. The *dep_Hour_utc* and *dep_Hour_local* columns store the hour information, representing the hour component of the departure time. For

example, a departure time of 8:00:00 would be represented as 8. The *dep_day_utc* and *dep_day_local* columns store the day information, specifying the day of the week for each flight (e.g., Monday, Tuesday, Wednesday, etc.). These columns enable an analysis of flight disruptions based on specific hours of the day and days of the week, revealing potential patterns and correlations.

To further categorize the time of departure, the *dep_partofhour_local* column is created. This column classifies the departure time into distinct parts of the day, such as Morning, Afternoon, Evening, and Night. This categorization allows us to examine whether flight disruptions vary based on different parts of the day, providing insights into potential time-related factors influencing disruptions.

Another important aspect of flight disruption analysis is considering whether flights occur on weekends or weekdays. To capture this information, the *dep_Weekend_local* column is introduced. By extracting the day information, we can identify whether the departure falls on a weekend or a weekday. This distinction provides insights into the potential influence of weekends on flight disruptions, as operational factors may differ between these periods.

In addition to the preprocessing steps performed on the departure time information, similar steps are also applied to the arrival time data in the flight dataset. These steps ensure consistency and enable comprehensive analysis of both depar-

ture and arrival aspects of flights.

Moving on, the *delayed* column, which represents the duration of flight delays, is filled with zeros for flights that are not delayed. This ensures consistency and facilitates analysis by providing a standardized value for flights without delays.

To further simplify the analysis of delays, a new column called *delayStatus* is created. This column discretizes the *delayed* column into a binary representation, where a value of True indicates a delayed flight, while False indicates a flight without delay. This discretization allows for easier classification and analysis of delayed flights, simplifying subsequent modelling and prediction tasks.

Additional information related to airports, such as airport name, airport city, airport latitude, and airport longitude, is appended to the flight dataset. This information enriches the dataset by providing context and spatial attributes for each flight's departure and arrival airports. The inclusion of this information allows for geospatial analysis and examination of potential airport-specific factors influencing flight disruptions.

Moreover, the *airline_icao* column is transformed by appending the full airline name to it. For example, abbreviations like MAS are changed to Malaysia Airlines. This renaming and transformation result in a more informative and descriptive column named *Airline*, which facilitates a better understanding and analysis of

flight disruptions by airlines.

To enhance the dataset further, the flight model information created in previous steps is appended to each flight using their respective *flight_iata* values. A left merge is performed between the flight model datasets and the flight dataset, ensuring that the flight model information is aligned with the corresponding flights. This integration enables analysis of flight disruptions in conjunction with the specific models of aircraft involved.

To classify flights as domestic or international, the city information derived earlier is utilized. By extracting the country information from the city column, we can determine the departure and arrival countries for each flight. The geopy library is leveraged to extract the country information accurately. With the arrival and departure countries determined, an algorithm is developed to check if the two countries are the same, allowing us to discretize the flights into international or domestic categories. This information is stored in a new column named *Flight_Type*, providing valuable insights into the distribution and characteristics of domestic and international flights in relation to disruptions.

In summary, the data pre-processing steps for the flight dataset involve carefully handling and transforming the data to ensure its quality, relevance, and suitability for subsequent analysis and modelling tasks. These steps include column

dropping, handling missing values, datetime formatting, feature extraction, rounding, categorization, discretization, appending additional information, and merging datasets. Each step contributes to a comprehensive and well-prepared dataset that enables accurate analysis and prediction of flight disruptions at Kuala Lumpur International Airport.

Below is the Flight dataset after performing data pre-processing:

Table 3.5: Flight Dataset Feature After Pre-Processing

<i>flight_iata</i>	The flight's unique identifier (IATA code)
<i>Airline</i>	The full name of the airline operating the flight
<i>dep_iata</i>	The IATA code of the departure city/airport
<i>dep_City</i>	The name of the departure city
<i>dep_Airport</i>	The name of the departure airport
<i>dep_Lat</i>	The latitude coordinate of the departure airport
<i>dep_Lon</i>	The longitude coordinate of the departure airport
<i>dep_Date_utc</i>	The UTC date of the departure
<i>dep_Time_utc</i>	The UTC time of the departure
<i>dep_Day_utc</i>	The day of the week of the departure (in UTC)
<i>dep_Hour_utc</i>	The hour of the departure (in UTC)
<i>dep_Day_local</i>	The day of the week of the departure (in local time)
<i>dep_Hour_local</i>	The hour of the departure (in local time)
<i>dep_Weekend_local</i>	Indicates if the departure is on a weekend or not
<i>dep_PartofHour_local</i>	The part of the day of the departure (Morning, Afternoon, Evening, Night)
<i>arr_iata</i>	The IATA code of the arrival city/airport
<i>arr_City</i>	The name of the arrival city
<i>arr_Airport</i>	The name of the arrival airport
<i>arr_Lat</i>	The latitude coordinate of the arrival airport
<i>arr_Lon</i>	The longitude coordinate of the arrival airport
<i>arr_Date_utc</i>	The UTC date of the arrival
<i>arr_Time_utc</i>	The UTC time of the arrival

<i>arr_Day_utc</i>	The day of the week of the arrival (in UTC)
<i>arr_Hour_utc</i>	The hour of the arrival (in UTC)
<i>arr_Day_local</i>	The day of the week of the arrival (in local time)
<i>arr_Hour_local</i>	The hour of the arrival (in local time)
<i>arr_Weekend_local</i>	Indicates if the arrival is on a weekend or not
<i>arr_PartofHour_local</i>	The part of the day of the arrival (Morning, Afternoon, Evening, Night)
<i>duration</i>	The duration of the flight in minutes
<i>delayed</i>	The delay duration in minutes
<i>delayStatus</i>	Indicates if the flight is delayed or not
<i>Aircraft_Model</i>	The model of the aircraft used for the flight
<i>Aircraft_isTurboProp</i>	Indicates if the aircraft is a turbo propeller aircraft
<i>Aircraft_isWidebody</i>	Indicates if the aircraft is a widebody aircraft
<i>Flight_Type</i>	Indicates whether the flight is domestic or international

3.2.2 Weather Dataset

A few simple pre-processing steps were taken to ensure that the weather dataset was in a format suitable for further analysis. Since the crawled dataset was relatively clean, the preprocessing processes focused on organizing and formatting the data for consistency.

The first step involved removing any duplicate entries from the dataset to ensure data integrity and avoid redundancy. This step is crucial as duplicate entries can lead to biased analysis and inaccurate results.

Next, the time column in the dataset was formatted into a DateTime format. By converting the time values into a standardized format, it becomes easier to manip-

Algorithm 5 : Weather Dataset Pre-Processing
<ol style="list-style-type: none"> 1. Read the Weather data from an Excel file into a DataFrame. 2. Drop duplicate rows from the weather table. 3. Convert the 'Time' column to datetime format using the format '%Y-%m-%dT%H:%M'. 4. Extract the date component from the 'Time' column and store it in a new 'Date' column in the format '%d-%m-%Y'. 5. Extract the hour component from the 'Time' column and overwrite the 'Time' column with the hour values. 6. Rename the 'Time' column to 'Hour'.

Figure 3.6: Weather Dataset Pre-Processing

ulate and analyze the temporal aspects of the data. This step ensures consistency and compatibility with other time-related variables in the analysis.

To enhance the data's usability and facilitate analysis based on date information, a new column called *Date* was created. This column extracted the date information from the time column, allowing for specific date-based analysis and insights. One of the reasons for including the date information is to facilitate the merging of the Weather dataset with the Flight dataset later on. The date information serves as a common key for merging the two datasets, enabling a comprehensive analysis that combines weather conditions with flight data.

Furthermore, to focus the analysis on the hourly variations of the weather, the time column was further processed to only store the hour information. By extracting and isolating the hour component of the time, a new column called *Hour* was created. This step simplifies the analysis by enabling a more granular examination of the weather patterns on an hourly basis. Similar to the date information, includ-

ing the hour information facilitates the merging of the Weather dataset with the Flight dataset. The hour information serves as another common key for aligning the weather conditions with specific flight times.

These preprocessing steps were performed to ensure the Weather dataset is organized, consistent, and ready for analysis. By removing duplicates, formatting the time column, and creating additional columns for date and hour information, the dataset becomes more manageable and suitable for exploring the temporal aspects of weather patterns. These steps not only contribute to the overall quality and integrity of the data but also enable a seamless integration with the Flight dataset for a comprehensive analysis of the impact of weather on flight disruptions.

Below is the Weather dataset after performing data pre-processing:

Table 3.6: Weather Dataset Feature After Pre-Processing

Airport	The airport code indicates the location of the weather measurement.
Date	The date on which the weather data was recorded.
Hour	The specific hour of the day when the weather data was recorded.
Temperature	The temperature is measured in degrees Celsius.
Humidity	The relative humidity is expressed as a percentage.
SeaLevelPressure	The atmospheric pressure at sea level is measured in millibars.
Precipitation	The amount of precipitation measured in millimetres.
Rain	The indicator of rain occurrence (0 indicates no rain, 1 indicates rain).
Snowfall	The amount of snowfall measured in millimetres.

WeatherCode	The numerical code representing the weather conditions.
Cloudcover	The extent of cloud coverage is measured as a percentage.
WindSpeed	The wind speed is measured in kilometres per hour.
WindDirection	The direction from which the wind is blowing, measured in degrees.
WindGusts	The maximum gust speed of the wind is measured in kilometres per hour.

3.2.3 Merging

The merging process involves combining the Flight Dataset with the Weather Dataset and the Foot Traffic Dataset to create a comprehensive dataset that incorporates relevant information about flights, weather conditions, and foot traffic at airports. The merging is performed in a series of steps to ensure the appropriate matching of data based on specific criteria.

Algorithm 6 : Merge Pre-Processing
<ol style="list-style-type: none"> 1. Read the Flight dataset from the Excel file "FYP2_FlightDataset_Final.xlsx". 2. Read the Weather dataset from the Excel file "FYP2_WeatherDataset_Final.xlsx". 3. Read the FootTraffic dataset from the CSV file "FYP2_FootTrafficDataset_Final.csv". 4. Create two separate copies of the Weather dataset for arrival and departure weather data: 'ArrivalWeather' and 'DepartureWeather'. The column names in each copy are prefixed with 'Arrival_' and 'Departure_', respectively. 5. Create two separate copies of the FootTraffic dataset for arrival and departure foot traffic data: 'ArrivalFootTraffic' and 'DepartureFootTraffic'. The column names in each copy are prefixed with 'Arrival_' and 'Departure_', respectively. 6. Merge the Flight dataset with the DepartureWeather dataset based on matching columns: 'dep_iata', 'dep_Date_utc', and 'dep_Hour_utc'. 7. Merge the resulting DataFrame with the ArrivalWeather dataset based on matching columns: 'arr_iata', 'arr_Date_utc', and 'arr_Hour_utc'. 8. Merge the resulting DataFrame with the DepartureFootTraffic dataset based on matching columns: 'dep_iata', 'dep_Date_utc', and 'dep_Hour_local'. 9. Merge the resulting DataFrame with the ArrivalFootTraffic dataset based on matching columns: 'arr_iata', 'arr_Date_utc', and 'arr_Hour_local'. 10. Drop unnecessary columns from the merged DataFrame, including the original airport codes and redundant columns from the merge operations.

Figure 3.7: Merging Process

To begin with, the Weather Dataset is duplicated and split into two separate data frames: one for departure weather and one for arrival weather. This duplication allows for distinct weather information to be associated with both the departure and

arrival aspects of each flight. Additionally, the column names in the duplicated data frames are modified by adding a prefix of "Arrival" for the arrival weather columns and "Departure" for the departure weather columns.

Similarly, the Foot Traffic Dataset is duplicated and split into two data frames: one for departure foot traffic and one for arrival foot traffic. This division enables the foot traffic information to be linked separately to the departure and arrival components of each flight.

The first merging step involves performing a left join between the Flight Dataset and the Departure Weather data frame. This merging is based on matching the departure airport code (*dep_iata*), departure date (*dep_Date_utc*), and departure hour (*dep_Hour_utc*) from both datasets. The purpose of this merging is to incorporate the relevant weather information for the departure phase of each flight into the Flight Dataset.

The second merging step entails conducting a left join between the previously merged data and the Arrival Weather data frame. This merging is performed by matching the arrival airport code (*arr_iata*), arrival date (*arr_Date_utc*), and arrival hour (*arr_Hour_utc*). By doing so, the dataset now includes the corresponding weather information for the arrival phase of each flight.

Next, a left join is performed to merge the dataset with the Departure Foot Traffic data frame. This merging is based on matching the departure airport code

(*dep_iata*), departure day (*dep_Day_local*), and departure hour (*dep_Hour_local*).

The objective here is to incorporate the foot traffic information for the departure phase of each flight into the dataset.

In the fourth and final merging step, a left join is executed to merge the previously merged data with the Arrival Foot Traffic data frame. This merging is performed by matching the arrival airport code (*arr_iata*), arrival day (*arr_Day_local*), and arrival hour (*arr_Hour_local*). This step ensures that the dataset includes the corresponding foot traffic information for the arrival phase of each flight.

After the merging process, some columns that are repeated or no longer needed are removed from the dataset. These columns include identifiers such as *dep_iata* and *arr_iata*, as well as various airport, date, and hour columns that were duplicated during the merging process.

In the end, the merging process results in a merged dataset that combines the Flight Dataset, the Weather Dataset, and the Foot Traffic Dataset. The merged dataset contains a total of 55 columns and encompasses 60,549 data entries, each representing a unique flight and its associated information.

Overall, the merging process enhances the value and utility of the individual datasets, providing a rich source of information for studying and understanding the complex dynamics between flights, weather, and foot traffic in the aviation

industry.

Below is the table of the features along with its example data:

Table 3.7: Final Dataset After Merging

Feature Name	Example
<i>flight_iata</i>	MH194
<i>Airline</i>	Malaysia Airlines
<i>dep_City</i>	Kuala Lumpur
<i>dep_Airport</i>	Kuala Lumpur International Airport
<i>dep_Lat</i>	2.745579958
<i>dep_Lon</i>	101.7099991
<i>dep_Date_utc</i>	06-01-2023
<i>dep_Time_utc</i>	11:55:00
<i>dep_Day_utc</i>	Friday
<i>dep_Weekend_local</i>	FALSE
<i>dep_PartofHour_local</i>	Evening
<i>arr_City</i>	Mumbai
<i>arr_Airport</i>	Chhatrapati Shivaji International Airport
<i>arr_Lat</i>	19.08869934
<i>arr_Lon</i>	72.86789703
<i>arr_Date_utc</i>	06-01-2023
<i>arr_Time_utc</i>	17:05:00
<i>arr_Day_utc</i>	Friday
<i>arr_Weekend_local</i>	FALSE
<i>arr_PartofHour_local</i>	Night
<i>duration</i>	310
<i>delayed</i>	0
<i>delayStatus</i>	FALSE
<i>Aircraft_Model</i>	Airbus A330-300
<i>Aircraft_isTurboProp</i>	False
<i>Aircraft_isWidebody</i>	True
<i>Flight_Type</i>	International
<i>Departure_Temperature</i>	26.2
<i>Departure_Humidity</i>	86
<i>Departure_SeaLevelPressure</i>	1010.1

<i>Departure_Precipitation</i>	0
<i>Departure_Rain</i>	0
<i>Departure_Snowfall</i>	0
<i>Departure_WeatherCode</i>	2
<i>Departure_Cloudcover</i>	56
<i>Departure_WindSpeed</i>	6.7
<i>Departure_WindDirection</i>	324
<i>Departure_WindGusts</i>	14.4
<i>Arrival_Temperature</i>	25
<i>Arrival_Humidity</i>	72
<i>Arrival_SeaLevelPressure</i>	1017.1
<i>Arrival_Precipitation</i>	0
<i>Arrival_Rain</i>	0
<i>Arrival_Snowfall</i>	0
<i>Arrival_WeatherCode</i>	0
<i>Arrival_Cloudcover</i>	1
<i>Arrival_WindSpeed</i>	10.1
<i>Arrival_WindDirection</i>	358
<i>Arrival_WindGusts</i>	18.4
<i>Departure_Intensity Level</i>	0
<i>Departure_Intensity Status</i>	Average
<i>Departure_Busyness</i>	80
<i>Arrival_Intensity Level</i>	0
<i>Arrival_Intensity Status</i>	Average
<i>Arrival_Busyness</i>	80

By following these meticulous data preprocessing steps, the dataset is transformed into a comprehensive and well-structured dataset ready for further analysis and modelling. These steps ensure data quality, enhance the dataset with additional relevant information, and enable the exploration of various factors contributing to flight disruptions at Kuala Lumpur International Airport.

3.3 Exploratory Data Analysis

After merging the datasets, a crucial step in the research methodology is to perform Exploratory Data Analysis using Power BI for data visualization. Exploratory Data Analysis is critical for understanding the relationships and patterns in the merged dataset. Data visualisation techniques such as line charts, scatter plots, bar graphs, and heatmaps can provide valuable insights into variable distribution, trend identification, and exploration of potential correlations.

In the exploratory data analysis (EDA) phase, the merged dataset plays a crucial role in uncovering valuable insights and patterns. To facilitate the analysis and ensure data accessibility and update ability, the dataset is stored in Google Drive, a cloud-based storage platform. By utilizing Google Drive, the dataset can be easily shared, updated, and accessed by relevant parties.

One of the primary tools used for analyzing the dataset is a Power BI dashboard. The dashboard is created and linked to the dataset stored in Google Drive through the Google Cloud API. This integration allows for seamless data connectivity between the Power BI dashboard and the dataset, enabling real-time visualization and analysis of the data.

Storing the dataset in Google Sheets format offers several advantages. Firstly, it provides a convenient way to update the data. When new data becomes available,

there is no need to create an entirely new Power BI dashboard and establish a connection to another file. Instead, users can simply upload the updated dataset to the existing Google Drive location.

Algorithm 7 : Exploratory Data Analysis
<ol style="list-style-type: none"> 1. Set up the required credentials and authorization for accessing Google Sheets using the 'gspread' library. 2. Open a specific Google Sheets file named "FYP2_Visualization" using the 'client.open()' function. 3. Clear any existing data from the sheet using the 'sheet.clear()' method. 4. Read a new dataset from an Excel file using 'pd.read_excel()' and store it in a pandas DataFrame called 'newdata'. 5. Convert the data in the 'newdata' DataFrame to string format using the 'astype()' method. 6. Update the Google Sheets file with the modified DataFrame by calling 'sheet.update()' and passing the column names and values as a list. 7. Define the Azure AD and Power BI details, including the tenant ID, client ID, client secret, and authority URL. 8. Specify the group ID and report ID of the Power BI report you want to display. 9. Create an instance of the 'Report' class, providing the required credentials and report details. 10. Display the Power BI report using the 'display()' method of the 'Report' instance.

Figure 3.8: Exploratory Data Analysis

To automate the process of uploading the dataset to Google Sheets, a Python algorithm is developed. This algorithm interacts with the Google Sheets API and the Google Drive API, both of which are enabled in the Google Cloud Console. The necessary authentication credentials, obtained as a JSON file, are stored in the Google Drive to establish a secure connection between the Python script and the Google Drive.

Using the '**oauth2client.service_account**' library, the Python script can link to the Google Drive and upload the dataset. Before sending the data to Google Sheets, it is important to ensure that the data is properly formatted, typically in a string format, to maintain data integrity during the upload process.

After the data is updated in Google Sheets, the Power BI dashboard automat-

ically retrieves the updated data. However, to ensure the dashboard reflects the most recent information, an automatic data refresh mechanism should be set up. This can be achieved by configuring a daily or hourly refresh schedule, utilizing Power Automate for data refresh, or manually refreshing the data as needed.

By leveraging Power BI for data visualization and analysis, this research aims to deliver meaningful insights into the relationships between various factors contributing to flight disruptions at Kuala Lumpur International Airport. The interactive and visually appealing dashboards created in Power BI will allow users to explore the data, derive actionable insights, and make well-informed decisions based on the latest processed dataset obtained through Python data preprocessing.

3.4 Modelling

This section outlines the approach used for constructing predictive models to predict flight disruptions at Kuala Lumpur International Airport. The models were built using data mining techniques and machine learning algorithms, including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbours, and Logistics Regression. The process involved dividing the dataset into training and testing sets, training the models on the training set to learn the patterns and relationships between the features and the target variable (flight disruptions), and evaluating the models' performance using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score.

Initially, certain columns that are deemed not useful for building the prediction model are dropped from the merged dataset. These columns include *dep_Lat*, *dep_Lon*, *arr_Lat*, *arr_Lon*, *delayed*, and *flight_iata*.

Next, label encoding is applied to the dataset using the `LabelEncoder().fit_transform` function. This step is performed to convert categorical variables into numerical form, allowing the models to process the data effectively.

The Synthetic Minority Over-sampling Technique (SMOTE) was used to address the issue of class imbalance in the dataset. The original dataset was im-

Algorithm 8 : Modelling
<ol style="list-style-type: none"> 1. Read the data from an Excel file named "FYP2_Data_Final.xlsx" using 'pd.read_excel()' and store it in a pandas DataFrame called 'data'. 2. Drop specific columns from the 'data' DataFrame using the 'drop()' method. The columns being dropped are 'dep_Lat', 'dep_Lon', 'arr_Lat', 'arr_Lon', 'delayed', and 'flight_data'. 4. Apply label encoding to the 'data_encode' DataFrame using the 'LabelEncoder().fit_transform' function to convert categorical variables into numerical values. 6. Split the encoded data into input features (X) and the target variable (y). 7. Perform Synthetic Minority Over-sampling Technique (SMOTE) to balance the data using the 'SMOTE' class from the imbalanced-learn library. 8. Split the data into training and testing sets using the 'train_test_split()' function, with a test size of 30% and a random state of 10. 9. Fit the SMOTE model with the training data only, oversampling the minority class. 10. Convert the oversampled data ('os_data_X' and 'os_data_y') back to DataFrames and concatenate them using 'pd.concat()' to create the 'df_smote' DataFrame. 12. Perform a one-way ANOVA test ('f_oneway()') to compare the 'delayStatus' column between the original and oversampled data. 13. Define a 'ranking()' function that normalizes and ranks a list of feature importance scores. 14. Specify the input features ('X') and target variable ('y') for the models. 15. Create a Random Forest Classifier ('rf') with specified parameters (such as the number of estimators and maximum depth). 16. Use Recursive Feature Elimination with Cross-Validation (RFECV) to select the most important features and rank them using the 'rfe.ranking_' attribute. 17. Apply the 'ranking()' function to the feature rankings obtained from RFECV and store the results in the 'rfe_score' DataFrame. 18. Sort the 'rfe_score' DataFrame in descending order based on the "Score" column. 19. Define a function named 'model()' that trains and evaluates multiple classification models, including Naive Bayes, Support Vector Machine, Decision Tree, Random Forest, K Nearest Neighbours, and Logistic Regression. 20. For each model, fit the model with the training data, make predictions on the test data, calculate accuracy, precision, recall, and F1-score, and save the trained model using pickle. 21. Display the results of each model in a DataFrame. 22. Call the 'model()' function with different subsets of input features ('X') and the target variable ('y') to compare the performance of the models on different feature sets and data (original vs. oversampled).

Figure 3.9: Modelling Process

balanced, with significantly fewer samples of flight disruptions compared to non-disruptions. A count plot is created to examine the class distribution. The plot reveals that the dataset contains 50,213 instances with no delay (class 0) and 10,336 instances with a delay (class 1). To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied.

The SMOTE model is constructed by oversampling the minority class (delayed flights) to match the number of instances in the majority class (non-delayed flights). The dataset is split into training and testing sets using a 70:30 ratio, and the SMOTE model is fit only on the training data.

After performing SMOTE, another count plot is generated to examine the class

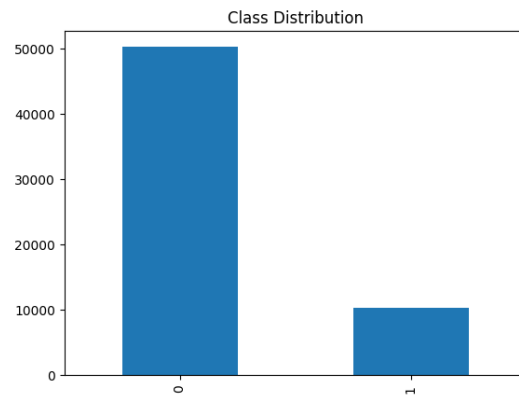


Figure 3.10: Count Plot before performing SMOTE

distribution. The plot now shows an equal distribution, with both class 0 and class 1 having 39,054 instances. By applying SMOTE, the class distribution was balanced, resulting in an equal number of samples for both classes.

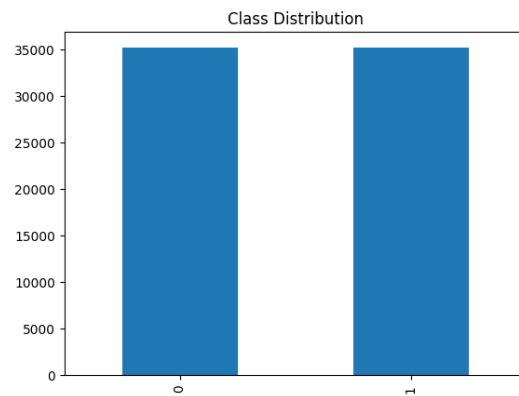


Figure 3.11: Count Plot after performing SMOTE

For feature selection, the Recursive Feature Elimination with Cross-Validation (RFECV) method is employed on the encoded dataset. A Random Forest classifier

(RF) is utilized as the base model for the selection process. The top 30 and bottom 30 features are identified based on their ranking scores obtained from RFECV.

With the preprocessed and feature-selected data, the model construction phase begins. The dataset is split into training and testing sets using a 70:30 ratio while ensuring the class distribution is maintained through stratification.

After that, the selected models were trained on two different databases which are Normal Dataset and SMOTE dataset. The six classification models, including Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression, Normal Data with 10 features, Normal Data with 30 features, SMOTE Data with 10 features, and SMOTE Data with 30 features.

At the end of the modeling phase, the constructed models will be saved using the pickle functions. This allows for easy retrieval and utilization of the trained models in future predictions or applications. Furthermore, storing the models enables efficient integration with other systems or applications for decision-making purposes later on.

After training the models, various evaluation metrics are calculated to assess their performance. These metrics include accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model's predictions. Precision

calculates the proportion of correctly predicted positive instances out of all instances predicted as positive, providing insights into the model's ability to avoid false positives. Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive instances out of all actual positive instances, indicating the model's ability to avoid false negatives. F1-score is the harmonic mean of precision and recall, offering a balanced assessment of the model's performance.

In summary, the methods used for predicting flight disruptions at Kuala Lumpur International Airport involved the construction of predictive models using data mining techniques. This included feature selection using RFE, addressing class imbalance using SMOTE, and training multiple classification models. The models' performance was evaluated using various metrics, leading to valuable insights for airlines and airport authorities to enhance their operational strategies and minimize disruptions.

3.5 Prediction Visualization

The prediction visualization component of this research methodology involves the development of an application using Streamlit, a user-friendly framework for creating interactive web applications. The application will allow users to upload their own datasets for visualisation and prediction purposes. Users can analyse and predict flight disruptions based on their specific data by allowing dataset uploads, increasing the applicability of the research findings.

The Prediction Visualization component utilizes Streamlit and Power BI to provide an interactive and visually appealing prediction experience. Streamlit is used as the initial platform for dataset processing and prediction. Its interactive nature allows users to easily upload their sample datasets and receive real-time results. Power BI is utilized to create visually appealing and informative dashboards. By linking Power BI to the generated prediction results, users can dynamically interact with the data and gain valuable insights.

To begin, the user is required to upload a dataset to the Streamlit application. This dataset is then processed and prepared for prediction. The previously generated prediction models, which have been saved and stored in a GitHub repository, are retrieved by Streamlit. The decision to save all the files, code, and models in

Algorithm 9 : Predictive Application
<ol style="list-style-type: none"> 1. Import the necessary libraries: 'streamlit', 'pandas', 'gsread', 'ServiceAccountCredentials', and 'pickle'. 2. Set up the Streamlit application and authenticate with Google Sheets API. 3. Create the file uploader component to allow users to upload a new dataset in CSV format. 4. If a dataset is uploaded, read the data into a pandas DataFrame and display it. 5. If no dataset is uploaded, retrieve the data from the Google Sheets spreadsheet and display it. 6. Define a function called 'model' to load pre-trained models, make predictions, and compute evaluation metrics. 7. Create a section to display the prediction result using pre-trained models. 8. Read the feature importance scores from a CSV file. 9. If the dataset is available, encode the categorical variables using label encoding. 10. Split the dataset into input features (X) and target variable (y). 11. Create four different DataFrames for different combinations of feature selection and data balancing techniques. 12. Train and evaluate six different machine learning models on each DataFrame. 13. Store the model performance metrics in separate DataFrames. 14. Merge all the DataFrames into a single DataFrame. 15. Identify the row with the highest F1-score and display the corresponding model information. 16. Make predictions using the selected model and the provided dataset. 17. Update the Google Sheets file with the predicted values. 18. Display a Power BI dashboard using an embedded iframe. 19. If no data is available in the spreadsheet, display a message indicating that no data is available. 20. End the algorithm.

Figure 3.12: Upload Data Streamlit

a GitHub repository provides several advantages. Firstly, it ensures version control and facilitates collaborative development. Additionally, GitHub allows for easy sharing and deployment of the prediction models. Storing the models in a central repository makes it convenient to retrieve and utilize them in the Streamlit application.

These models are crucial for generating accurate predictions on the uploaded dataset. Streamlit leverages these models to compute various results for the dataset, including metrics such as accuracy, recall, precision, and F1-score. The comparison of multiple models based on their F1 scores is a common practice in machine learning. The F1 score takes into account both precision and recall, providing a balanced assessment of the model's performance. By selecting the model with the highest F1 score, the Prediction Visualization component ensures the most accu-

rate and reliable predictions are generated.

Approximately 24 pre-trained models from the previous modelling phase are used for prediction. The prediction model will use the dataset uploaded by the user to make predictions for flight disruptions. To ensure the selection of the most accurate and reliable model, the algorithm will automatically choose the model with the highest F1 score, a metric that balances precision and recall. It ensures that the model can accurately identify disruptions (precision) while capturing a significant portion of actual disruptions (recall). This approach ensures that the selected model can effectively predict flight disruptions at Kuala Lumpur International Airport while considering both accuracy and completeness.

After determining the best model, the user's dataset will be used for prediction. The chosen model will use its learned patterns and relationships to make predictions. The output prediction data is then stored in Google Drive, ensuring that the most recent predictions are available. Power BI can retrieve the data from Google Drive using scheduled refresh, Power Automate, or manual refresh.

After that, Power BI will be employed to extract the latest data from Google Spreadsheets to visualize and analyze the prediction results. By connecting to Google Spreadsheets, Power BI will generate a dynamic and informative dashboard based on the most recent prediction data.

After inserting info into the Power BI dashboard, the dashboard will be embedded into the Streamlit application. The user will be able to explore and interact with the dashboard, gaining a comprehensive understanding of the predicted flight disruptions based on their uploaded dataset inside the applications.

In summary, the prediction visualization component involves the development of an interactive application using Streamlit. Users can upload their datasets to generate predictions for flight disruptions. The best-performing model will be chosen based on the F1 score, and the prediction output, along with the dataset, will be saved in Google Spreadsheets. Power BI will extract data from Google Spreadsheets and create a visually appealing dashboard that will be embedded in the application. This approach enables users to easily explore and analyse prediction results and make informed decisions based on the most recent data.

4 Analysis and Findings

4.1 The Results of Exploratory Data Analysis

In this section, the findings and insights obtained from the exploratory data analysis (EDA) conducted on the airline performance dataset are presented. The EDA provided us with valuable information regarding various aspects of airline operations, weather influence, busyness, and flight characteristics. The results are presented through interactive visualizations and descriptive statistics, offering a comprehensive understanding of the dataset.

4.1.1 Airline Performance

In the first section of the EDA, we focused on analyzing airline performance. The figure below shows the entire dashboard of the Airline Performance dashboard.

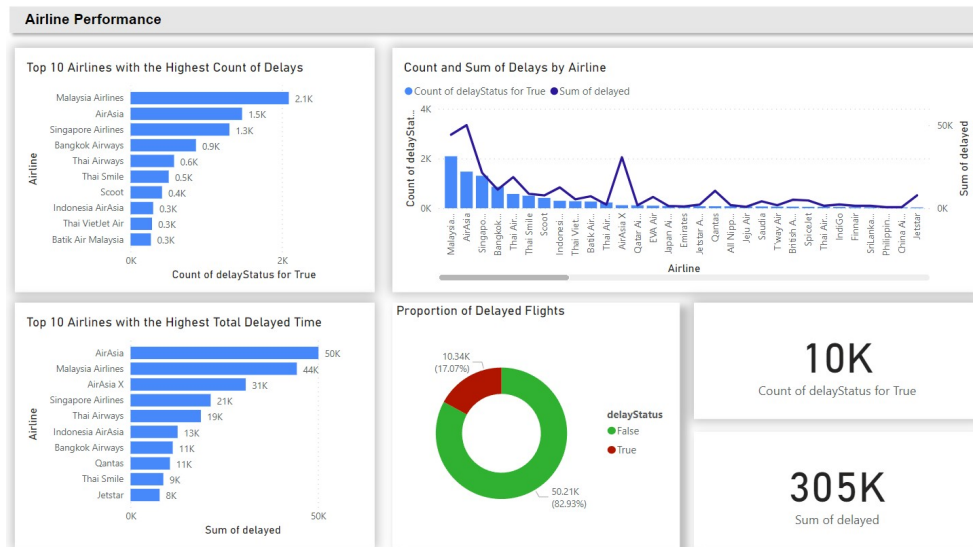


Figure 4.1: Airline Performance Dashboard

Chart 1: Top 10 Airlines with the Highest Count of Delays

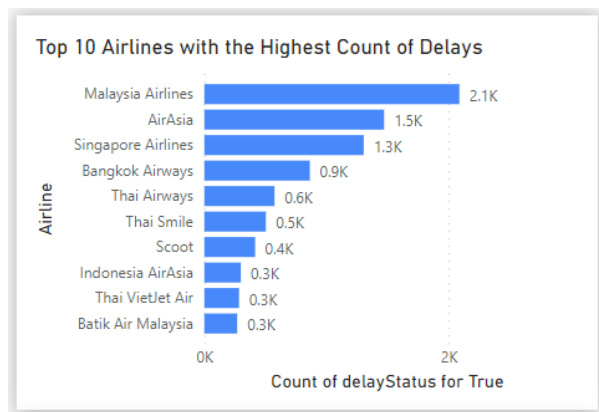


Figure 4.2: Top 10 Airlines with the Highest Count of Delays

The top-left chart of the dashboard is a bar chart titled "Top 10 Airlines with

the Highest Count of Delays.” It visualizes the count of delays on the x-axis and the top 10 airlines on the y-axis, showcasing which airlines experience the most delays.

The results of the charts show that Malaysia Airline had the highest count of delays with 2090 delayed flights, followed by AirAsia with 1473, Singapore Airline with 1306, Bangkok Airways with 864, and Thai Airways with 575, among others.

Chart 2: Top 10 Airlines with the Highest Total Delayed Time

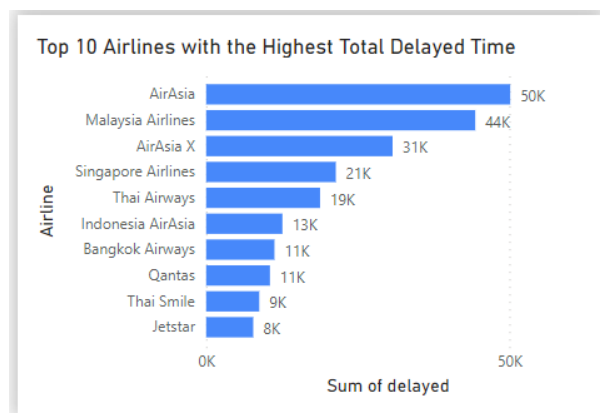


Figure 4.3: Top 10 Airlines with the Highest Total Delayed Time

The bottom-left bar chart is titled ”Top 10 Airlines with the Highest Total Delayed Time.” It displays the sum of delayed time in minutes on the x-axis and the top 10 airlines on the y-axis, highlighting the airlines with the greatest accumulated

delay time.

The result shows that AirAsia recorded the highest total delayed time with 50137 minutes, followed by Malaysia Airline with 44372 minutes, AirAsia X with 30742 minutes, and Singapore Airline with 21378 minutes, among others.

Chart 3: Count of Delay and Sum of Delayed Time by Airline

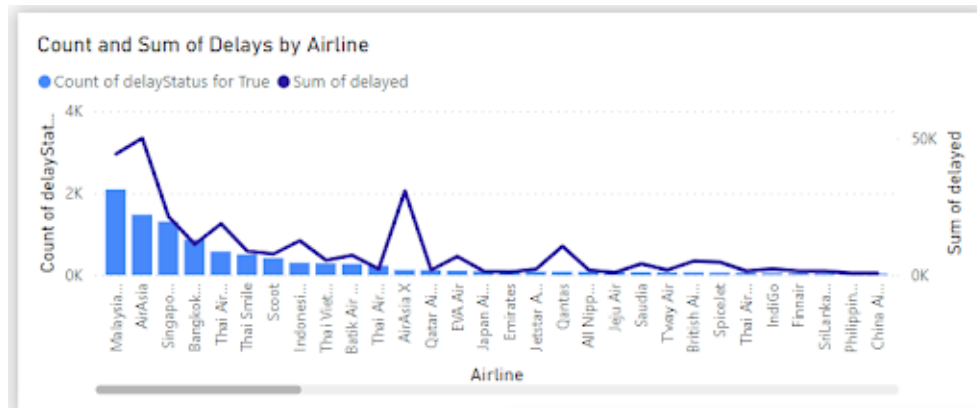


Figure 4.4: Count of Delay and Sum Of Delayed time by Airline

The top-right chart, titled "Count and Sum of Delays by Airline," combines a bar chart and a line chart. The bar chart represents the count of delays, while the line chart illustrates the sum of delay time. The x-axis shows all the airlines, providing a comprehensive overview of delay counts and durations for each airline.

The results revealed an interesting observation for AirAsia X, which had 122 flights that experienced delays but had a surprisingly high total delayed time of

30742 minutes. Other airlines displayed a relatively normal distribution of delays in terms of count and delay time. Malaysia Airline and AirAsia were the prominent airlines in terms of both delayed flights and total delayed time.

Chart 4: Proportion of Delayed Flights

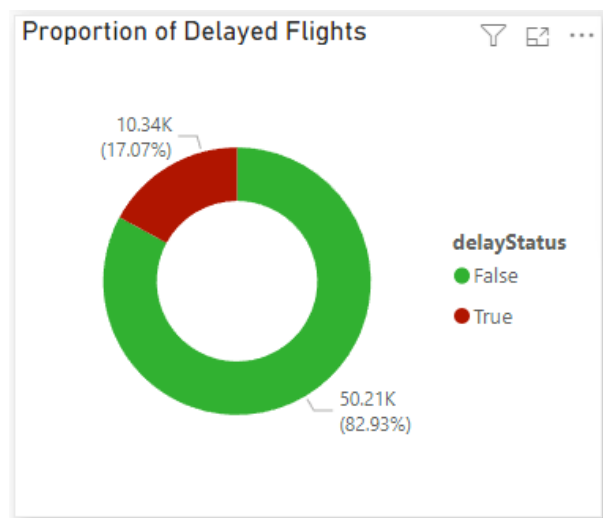


Figure 4.5: Proportion of Delay Flight

The fourth chart in this dashboard is a donut chart titled "Proportion of Delayed Flights." It represents the distribution of delay status in the dataset. The red colour indicates delays, while green indicates no delays.

The results show that the proportion of delayed flights was analyzed, with 17.07% (10336 flights) categorized as delayed and 82.93% (50213 flights) categorized as not delayed.

Card 1 & 2: Total Number of Delays Sum of the Delays Time

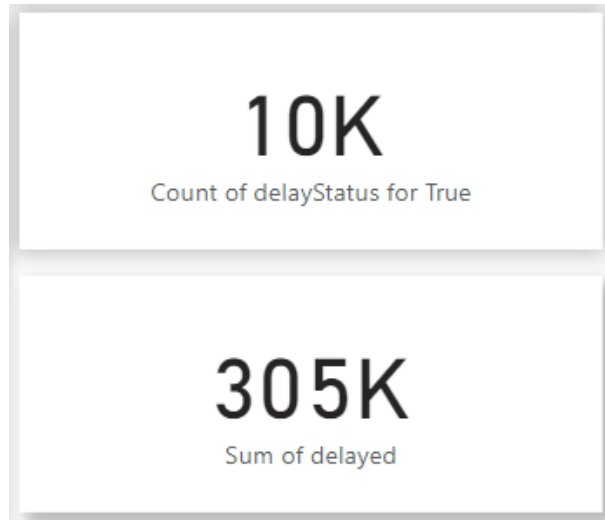


Figure 4.6: Delay information

The two cards on the bottom-right corner of the dashboard display essential metrics. The first card shows the total number of delays in the entire dataset, while the second card presents the sum of all delayed time. The result shows that in this dataset, around ten thousand flights were delayed, and the total number of times delayed was 305k minutes, which is around 5083 hours.

4.1.2 Weather Influence

Moving to the second dashboard, focuses on analyzing the influence of weather conditions on flight delays. The figure below shows the entire dashboard of Weather Influence dashboard.

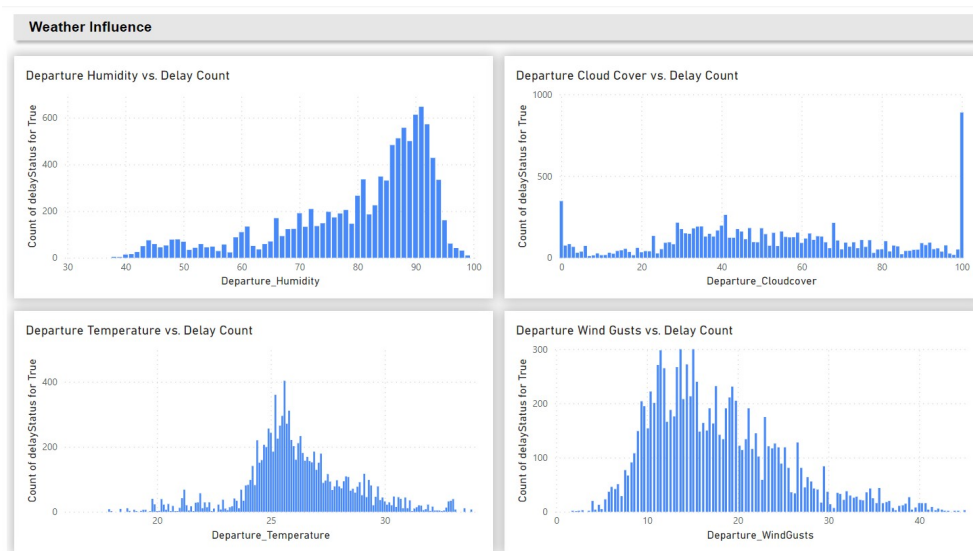


Figure 4.7: Weather Influence Dashboard

Chart 1: Departure Humidity vs. Delay Count

The top-left bar chart, titled "Departure Humidity vs. Delay Count," compares the departure humidity on the x-axis with the count of delay statuses on the y-axis. The chart revealed a significant increase in flight delays as humidity levels increased. The peak delay count of 647 flights was observed at a humidity level

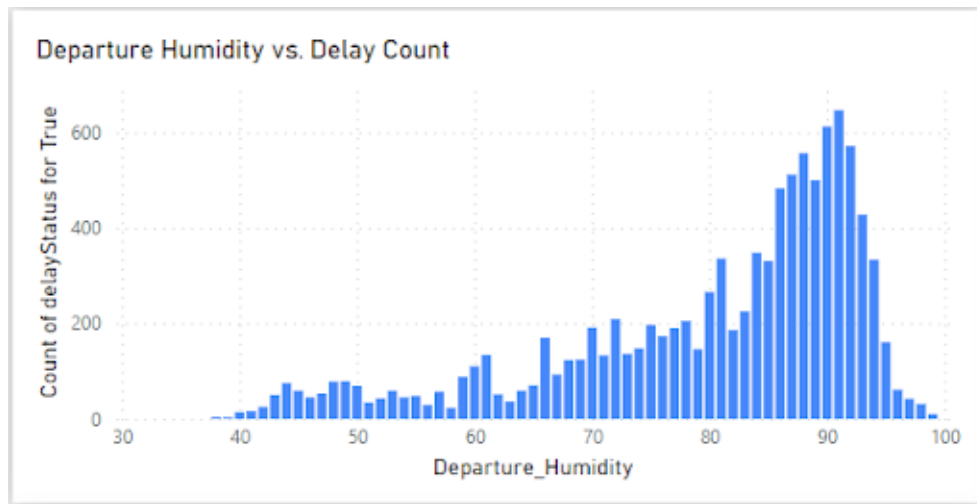


Figure 4.8: Departure Humidity vs. Delay Count

of 91.

Chart 2: Departure Cloud Cover vs. Delay Count

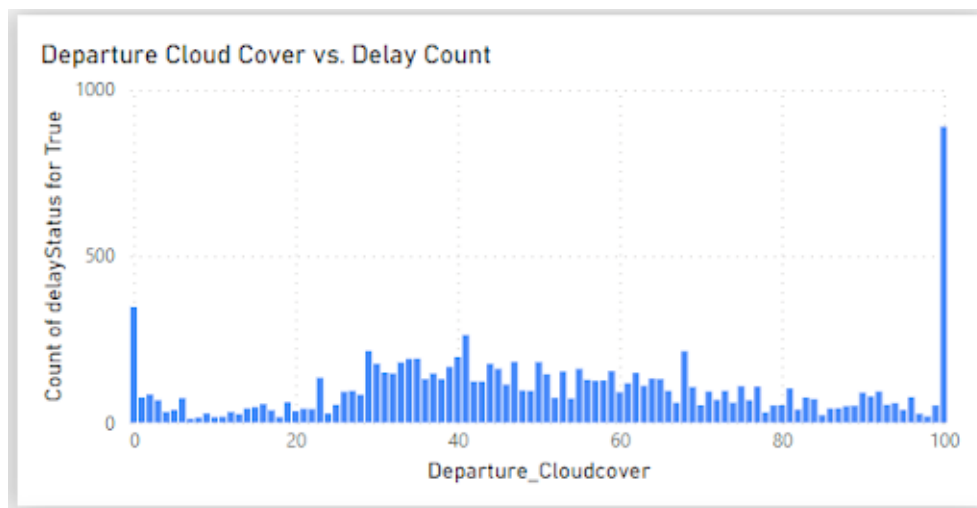


Figure 4.9: Departure Cloud Cover vs. Delay Count

The top-right bar chart, titled "Departure Cloud Cover vs. Delay Count," showcases the relationship between delay count and departure cloud cover. The x-axis represents the range of cloud cover values, while the y-axis represents the count of delay statuses.

Flight delays were distributed evenly across different cloud cover percentages, except for 100% cloud cover, which had a significantly higher number of delayed flights (889 flights). The range of delay counts for cloud cover percentages from 1% to 99% was relatively consistent, ranging from 20 to 70 flights delayed. At 0% cloud cover, there were 347 delayed flights.

Chart 3: Departure Temperature vs. Delay Count

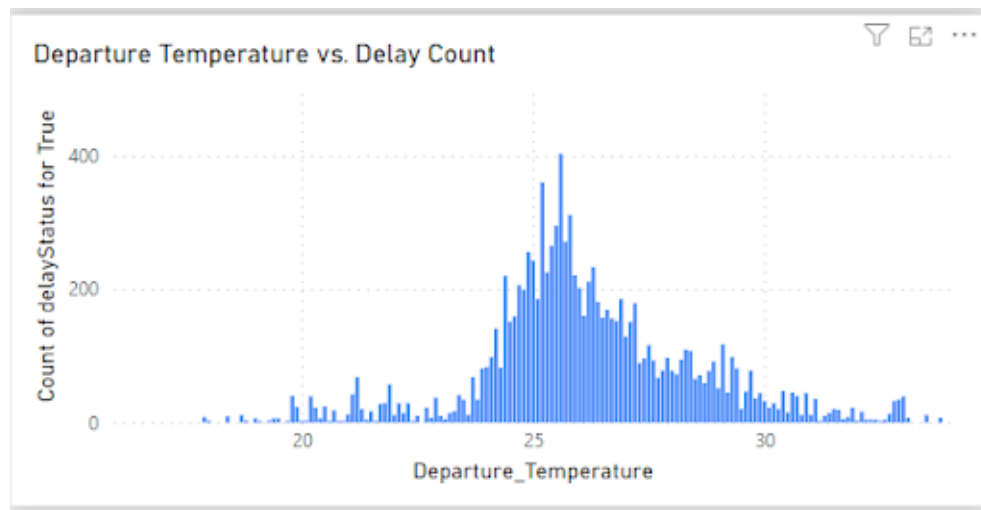


Figure 4.10: Departure Temperature vs. Delay Count

The bottom-left bar chart, titled "Departure Temperature vs. Delay Count," examines the impact of departure temperature on delay count. The x-axis displays the range of temperatures, while the y-axis represents the count of delay statuses. The chart indicated that most delays occurred within the temperature range of 24 to 26 degrees Celsius.

Chart 4: Departure Wind Gusts vs. Delay Count

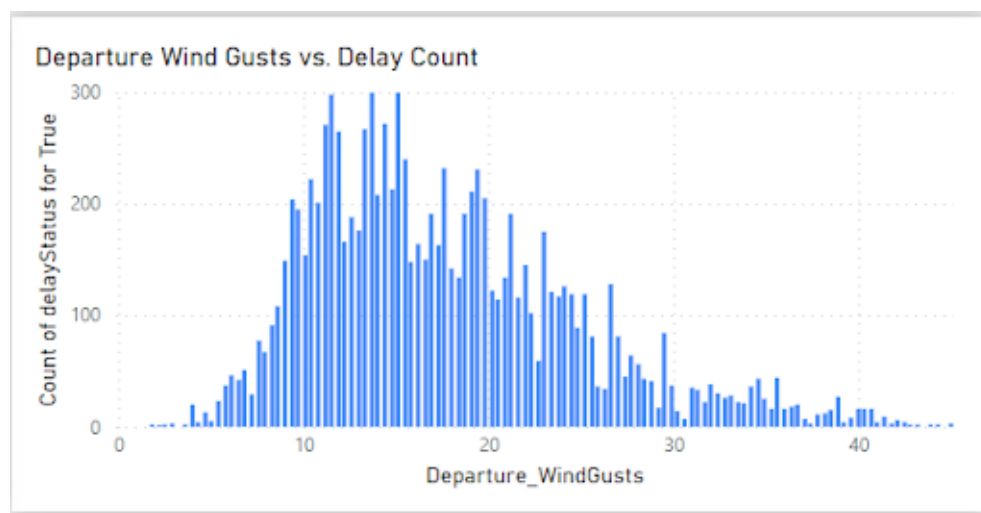


Figure 4.11: Departure Wind Gusts vs. Delay Count

Similarly, the bottom-right bar chart, titled "Departure Wind Gusts vs. Delay Count," explores the correlation between delay count and departure wind gusts. The x-axis represents the range of wind gusts, while the y-axis depicts the count of delay statuses.

Flight delays showed a distinct pattern with wind gusts. Delays were highest in the range of 10 to 15, with a decrease in delays beyond 15. Notably, there was a significant increase in delays from 0 to around 10 wind gusts.

4.1.3 Busyness

Moving to the third dashboard, focuses on analyzing the busyness factors on flight delays. The figure below shows the entire dashboard of Busyness dashboard.

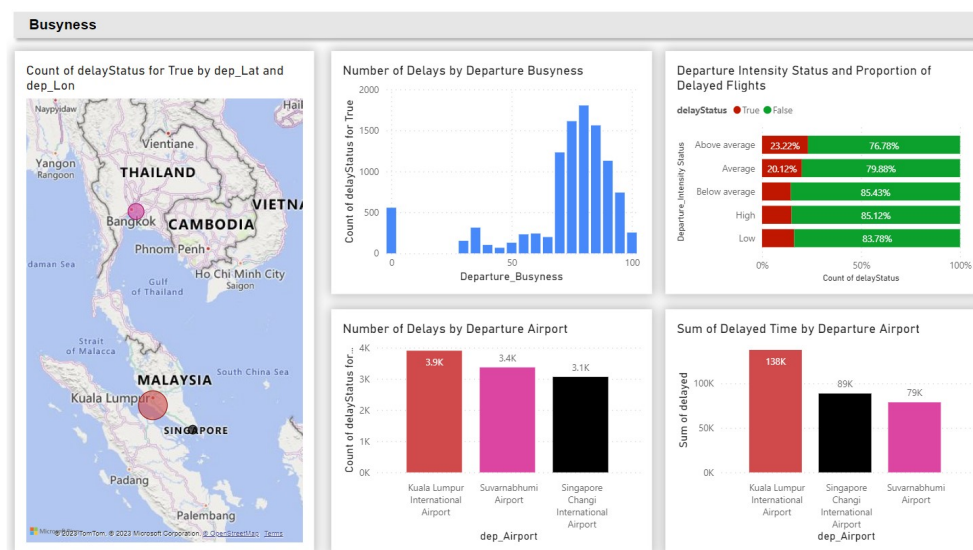


Figure 4.12: Busyness Dashboard

Chart 1 and 2 : Count of Delay Status on the Map and Number of Delays by Departure Airport

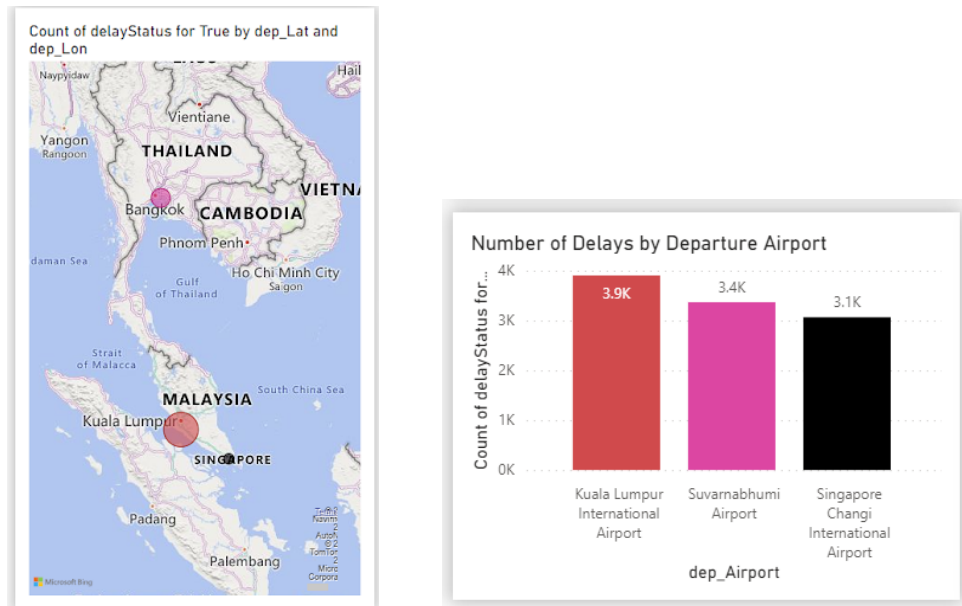


Figure 4.13: Map and Number of Delays by Departure Airport

The map on the left visualizes all the departure airports as dots, with the size of the bubble indicating the number of delays at each airport. The middle-bottom bar chart, titled "Number of Delays by Departure Airport," illustrates the number of delayed flights at each departure airport. The x-axis shows the airports, while the y-axis represents the count of delayed flights.

According to both of the charts, the results indicate that Kuala Lumpur International Airport had the highest number of delays, with approximately 3.9k delayed

flights. It was followed by Suvarnabhumi Airport with around 3.4k delays and Singapore Changi International Airport with around 3.1k delays.

Chart 3: Number of Delays by Departure Busyness

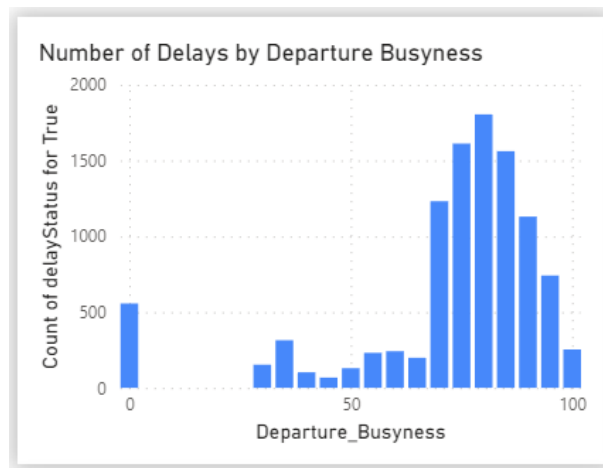


Figure 4.14: Number of Delays by Departure Busyness

The middle-top bar chart, titled "Number of Delays by Departure Busyness," represents the number of delayed flights against the departure busyness level. The x-axis displays the departure busyness ranging from 0 to 100, while the y-axis represents the number of delayed flights. The chart revealed that departure busyness levels of 80% had the highest number of delays, with 1805 delayed flights. Most delays occurred within the busyness range of 70% to 90%.

Chart 4: Departure Intensity Status and Proportion of Delayed Flights

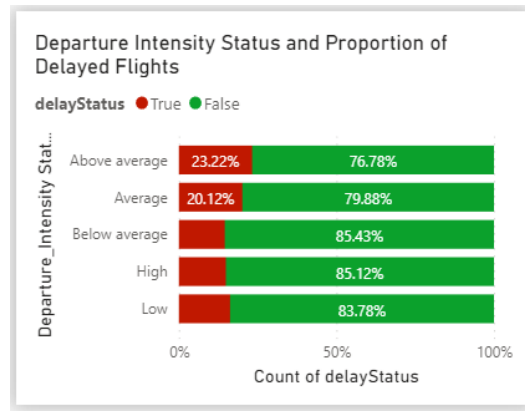


Figure 4.15: Departure Intensity Status and Proportion of Delayed Flights

The top-right chart is a 100% stacked bar chart titled "Departure Intensity Status and Proportion of Delayed Flights." It showcases the different intensity statuses and the percentage of delayed flights. The x-axis represents the percentage range from 0% to 100%, while the y-axis displays the intensity statuses. The red colour indicates the proportion of delayed flights, while the green represents flights without delays.

By looking at the charts, the proportion of delayed flights seems to occupy a larger percentage when the intensity status is above average with a percentage of 23.22%. The proportion of delayed flights is below average, with a proportion of delayed flights of 14.57%. Overall, there is no significant difference between the intensity and the proportion of delay for each category.

Chart 5: Sum of Delayed Time by Departure Airport

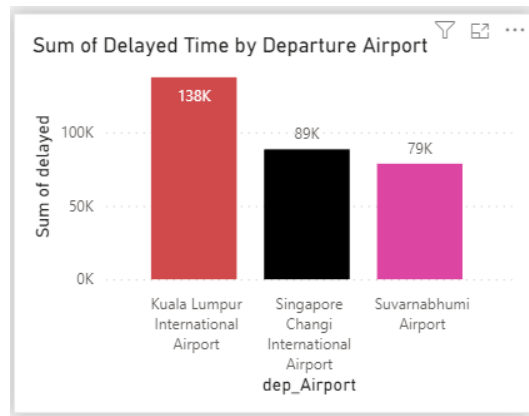


Figure 4.16: Departure Intensity Status and Proportion of Delayed Flights

The bottom-right chart, titled "Sum of Delayed Time by Departure Airport," displays the total time delayed at each departure airport. The x-axis represents the airports, and the y-axis represents the total time delayed. According to the charts, Kuala Lumpur International Airport recorded the highest sum of delayed time, with approximately 138k minutes, followed by Singapore Changi Airport with around 89k minutes and Suvarnabhumi Airport with around 79k minutes.

4.1.4 Flight Characteristics

The fourth section of the EDA focused on analyzing flight characteristics and their impact on delays. The following figures shows the dashboard of Flight Characteristics dashboard.



Figure 4.17: Flight Dashboard

Chart 1: Count of Delayed Flights by Flight Type

The top-left bar chart, titled "Count of Delayed Flights by Flight Type," compares the count of delayed flights for international and domestic flights. The chart revealed that international flights had a higher count of delayed flights (8.2k) compared to domestic flights (2.1k).

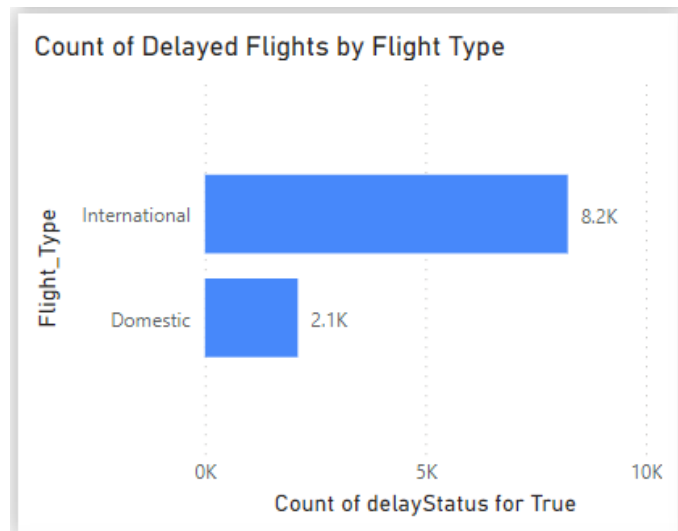


Figure 4.18: Count of Delayed Flights by Flight Types

Chart 2: Average Delayed Time by Flight Type

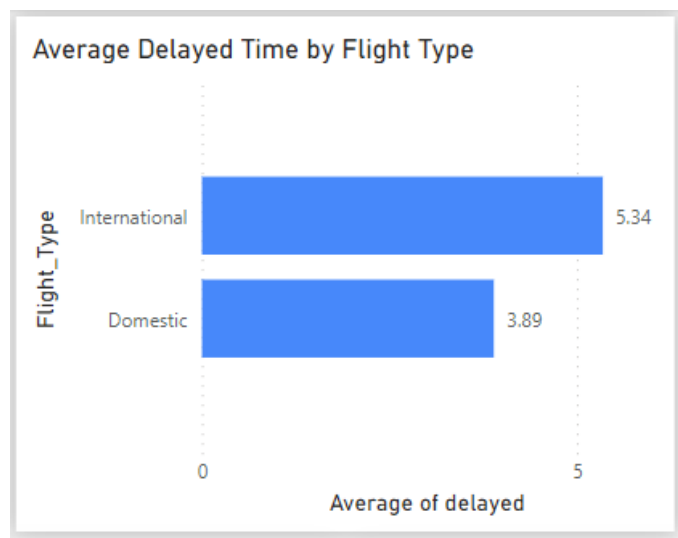


Figure 4.19: Average Delayed Time by Flight Type

The top-middle chart, titled "Average Delayed Time by Flight Type," displays the average delayed time for different flight types. The x-axis represents the average time, while the y-axis represents the flight types (domestic and international). On average, international flights experienced a delay of 5.34 minutes, while domestic flights had an average delay of 3.89 minutes.

Chart 3: Number of Delays by Aircraft Model

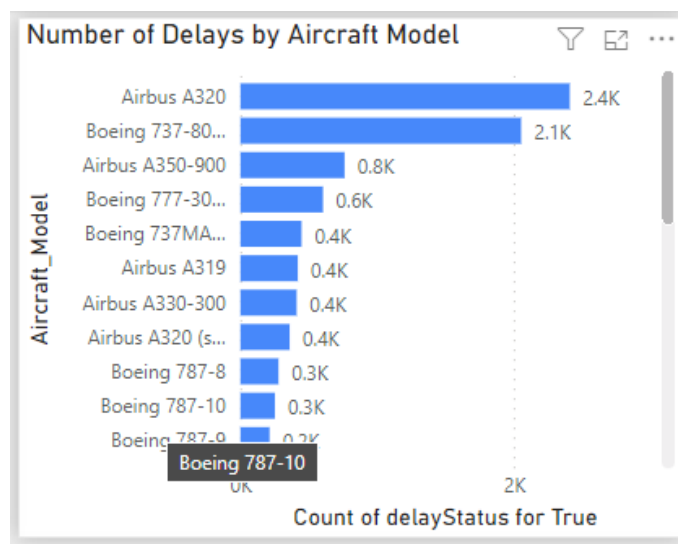


Figure 4.20: Number of Delays by Aircraft Model

The top-right chart, titled "Number of Delays by Aircraft Model," presents the count of delayed flights for different aircraft models. The y-axis shows the different aircraft models, while the x-axis represents the number of flight delays.

The chart showed that Airbus A320 had the highest number of delayed flights (2.4k), followed by Boeing 737-800 (2.1k) and Airbus A350-900 (762).

Chart 4: Count of Delayed Flights by Duration of Flight

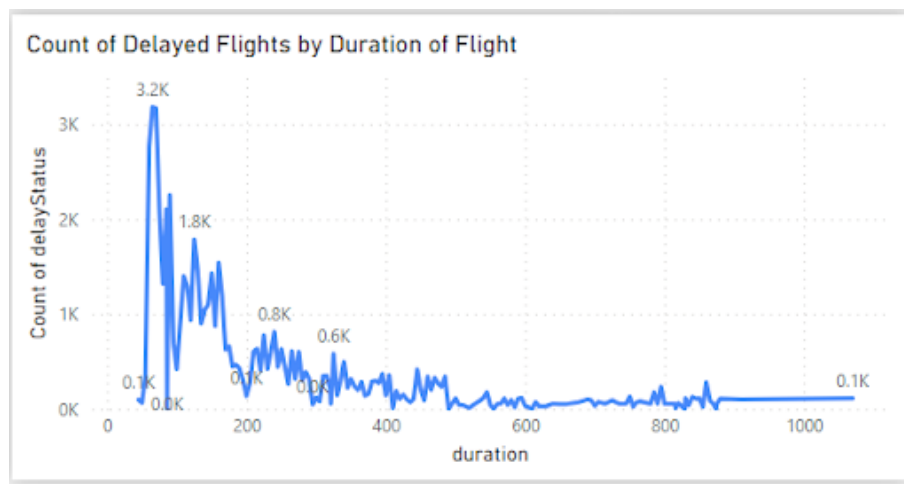


Figure 4.21: Count of Delayed Flights by Duration of Flight

The bottom-left chart, titled "Count of Delayed Flights by Duration of Flight," visualizes the count of delayed flights based on the duration of the flights. The x-axis displays the duration of flights, and the y-axis represents the number of flight delays. The chart indicated that flights with shorter durations tended to experience more delays. For example, at a flight duration of 65 minutes, there were 3182 delayed flights. As the duration of the flight increased, the number of delayed flights decreased.

Chart 5: Number of Delays by Departure Time

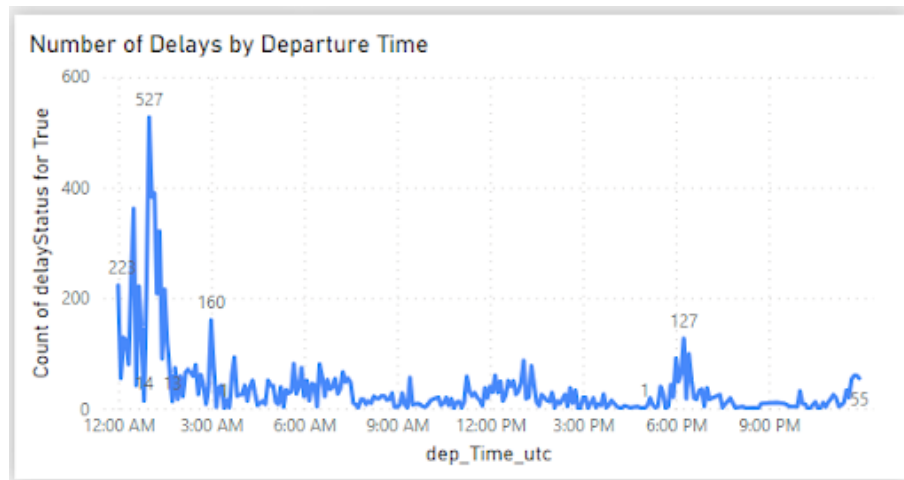


Figure 4.22: Number of Delays by Departure Time

The bottom-right chart, titled "Number of Delays by Departure Time," represents the count of delayed flights based on the departure time. The x-axis shows the time of the flight, and the y-axis displays the number of flight delays. The chart highlighted that flight delays were more prevalent during midnight hours, particularly from 12:00 am to 2:00 am. The highest number of delays (527 flights) occurred at 1 am.

4.1.5 Evaluation of Findings

The EDA provided valuable insights into various aspects of airline performance, weather influence, busyness, and flight characteristics. The findings revealed several key points worth highlighting:

- Malaysia Airline and AirAsia were prominent in terms of both delayed flights and total delayed time, indicating potential areas for improvement in their operations.
- Weather conditions such as humidity, cloud cover, temperature, and wind gusts showed varying degrees of influence on flight delays. Understanding these relationships can assist airlines in better managing their operations during adverse weather conditions.
- Airport busyness, as indicated by departure intensity and busyness levels, had an impact on flight delays. Identifying periods of high busyness can help airports and airlines optimize their resources to minimize delays.
- Flight characteristics, including flight type, aircraft model, duration, and departure time, exhibited associations with delays. International flights tended to experience more delays than domestic flights, certain aircraft models showed higher delay counts, shorter flights had more delays, and specific departure times had increased delay frequencies.

These findings provide valuable insights for airlines, airports, and stakeholders

to enhance their operational strategies, optimize resources, and improve customer experiences. By understanding the factors contributing to delays, proactive measures can be taken to minimize disruptions and ensure smoother travel experiences for passengers.

4.2 Feature Selection

Feature selection is a crucial step in the modelling process as it helps identify the most relevant features that contribute significantly to predicting flight delays. In this section, we present the results of the feature selection analysis.

The top 30 features selected based on their scores are as follows:

Table 4.1: Top 30 Features

1. dep_Time_utc	16. Departure_SeaLevelPressure
2. arr_Time_utc	17. dep_Airport
3. Airline	18. Departure_WindGusts
4. dep_PartofHour_local	19. Arrival_Temperature
5. Departure_Temperature	20. Departure_Intensity Level
6. Departure_Busyness	21. arr_City
7. duration	22. Arrival_Intensity Level
8. Departure_Humidity	23. Departure_WindDirection
9. dep_Date_utc	24. Departure_WindSpeed
10. dep_Day_utc	25. Arrival_Humidity
11. arr_Date_utc	26. arr_Day_utc
12. arr_PartofHour_local	27. dep_City
13. Aircraft_Model	28. Aircraft_isWidebody
14. arr_Airport	29. Arrival_SeaLevelPressure
15. Arrival_Busyness	30. Departure_Cloudcover

These features have been ranked based on their scores, with higher scores indicating greater importance in predicting flight delays. The top-ranked features encompass various aspects, including departure and arrival times, airline information, weather conditions (such as temperature, humidity, wind gusts, and cloud

cover), flight duration, airport busyness, aircraft model, and more. These features collectively provide valuable information for predicting delays accurately. On the other hand, the bottom 30 features, with relatively lower scores, may have less influence on flight delays. These features include factors like precipitation, snow-fall, wind direction, and weather code

4.3 Modelling Results

After performing feature selection, the selected features were used to train several machine learning models. Here, we present the modelling results using both normal data and SMOTE data.

4.3.1 Modelling Results with Normal Data

The modelling results using the top 10 and top 30 features with normal data are as follows:

Table 4.2: Result using Top 10 Features and Normal Data

Model	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	83.52%	0.807	0.835	0.811
Support Vector Machine	84.61%	0.822	0.846	0.814
Decision Tree	84.58%	0.846	0.846	0.846
Random Forest	88.35%	0.875	0.884	0.877
K Nearest Neighbors	85.52%	0.842	0.855	0.845
Logistic Regression	82.86%	0.777	0.829	0.771

These results demonstrate the performance of different models in predicting flight delays using the selected features. The Random Forest model achieved the highest accuracy of 88.35% among the top 10 features.

Table 4.3: Result using Top 30 Features and Normal Data

Model	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	81.9%	0.806	0.819	0.812
Support Vector Machine	82.93%	0.688	0.829	0.752
Decision Tree	83.5%	0.837	0.835	0.836
Random Forest	87.94%	0.869	0.879	0.870
K Nearest Neighbors	84.42%	0.826	0.844	0.831
Logistic Regression	83.13%	0.789	0.831	0.782

The Random Forest model achieved the highest accuracy of 87.94% among the top 30 features, indicating its effectiveness in predicting flight delays.

4.3.2 Modelling Results with SMOTE Data

SMOTE data augmentation technique was applied to address the imbalanced class distribution. The modelling results using the top 10 and top 30 features with SMOTE data are as follows:

Table 4.4: Result using Top 10 Features and SMOTE data

Model	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	66.75%	0.668	0.668	0.667
Support Vector Machine	72.36%	0.733	0.724	0.721
Decision Tree	85.18%	0.852	0.852	0.852
Random Forest	91.07%	0.911	0.911	0.911
K Nearest Neighbors	83.25%	0.834	0.832	0.832
Logistic Regression	67.32%	0.677	0.673	0.672

The modelling results with SMOTE data indicate that the Random Forest model

achieved the highest accuracy of 91.07% among the top 10 features.

Table 4.5: Result using Top 30 Features and SMOTE data

Model	Accuracy (%)	Precision	Recall	F1-score
Naive Bayes	67.44%	0.675	0.674	0.674
Support Vector Machine	71.56%	0.727	0.716	0.712
Decision Tree	86.96%	0.870	0.870	0.870
Random Forest	93.1%	0.931	0.931	0.931
K Nearest Neighbors	83.84%	0.856	0.838	0.836
Logistic Regression	67.35%	0.677	0.673	0.672

Among the top 30 features with SMOTE data, the Random Forest model again achieved the highest accuracy of 93.1%, indicating its robustness in handling imbalanced class distribution.

4.3.3 Modelling Result Summary

Overall, the modelling results demonstrate that the Random Forest model consistently performed well across different feature selections and data augmentation techniques. It showcased the highest accuracy in both the normal data and SMOTE data scenarios, suggesting its effectiveness in predicting flight delays based on the selected features.

These findings highlight the importance of feature selection and the impact it can have on model performance. The selected features provide valuable insights into the factors influencing flight delays, enabling airlines and airports to take

proactive measures for better operational efficiency and passenger satisfaction.

4.4 Predictive Visualization

The Prediction Visualization component utilizes Streamlit and Power BI to provide an interactive and visually appealing prediction experience. Streamlit is used as the initial platform for dataset processing and prediction. Its interactive nature allows users to easily upload their sample datasets and receive real-time results. Power BI is utilized to create visually appealing and informative dashboards. By linking Power BI to the generated prediction results, users can dynamically interact with the data and gain valuable insights.

To begin, the user is required to upload a dataset to the Streamlit application. This dataset is then processed and prepared for prediction.

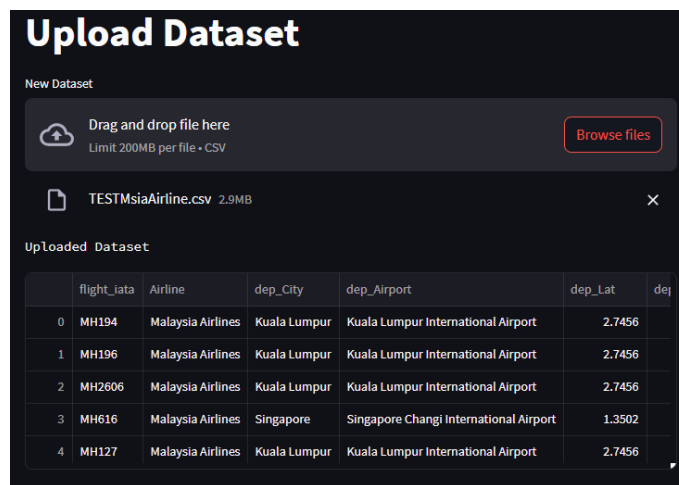


Figure 4.23: Upload Data Streamlit

Streamlit leverages the pre-trained models to compute various results for the dataset, including metrics such as accuracy, recall, precision, and F1-score. By selecting the model with the highest F1 score, the Prediction Visualization component ensures the most accurate and reliable predictions are generated. 24 pre-trained models are used for prediction, and their performance is compared using the F1-score metric. The model with the highest F1 score is selected as the best model for generating the final output.

Prediction Result Using Pre-Trained Model

Model Trained Using top 10 Features and Normal Data

	Model	Accuracy	Precision	Recall	F1-score
0	Naive Bayes	51.96	0.6541	0.5196	0.5376
1	Support Vector Machine	71.38	0.7167	0.7138	0.5984
2	Decision Tree	52.89	0.5662	0.5289	0.5442
3	Random Forest	68.93	0.5607	0.6893	0.5928
4	K Nearest Neighbours	54.9	0.568	0.549	0.5577
5	Logistic Regression	58.61	0.5877	0.5861	0.5869

Model Trained Using top 30 Features and Normal Data

	Model	Accuracy	Precision	Recall	F1-score
0	Naive Bayes	69.82	0.6605	0.6982	0.6664
1	Support Vector Machine	71.2	0.6069	0.712	0.5922

Figure 4.24: Prediction Process Streamlit

The final output consists of the dataset along with the generated prediction results. This updated data is then stored in Google Drive, ensuring that the most recent predictions are available. Power BI can retrieve the data from Google Drive

```

Naive Bayes using Normal Data with 30 Features pre-trained model has the
highest F1-Score of 0.6643046366530525 compare to other model.
Therefore, Naive Bayes is used in the process afterwards.

{
  "spreadsheetId" : "1NrCe0Guyqxax2AnvieooR3bHL58LVe-GFgEuUgHwh18"
  "updatedRange" : "Sheet1!A1:BD7258"
  "updatedRows" : 7258
  "updatedColumns" : 56
  "updatedCells" : 406448
}

```

Figure 4.25: Selection of Best Model Streamlit

using scheduled refresh, Power Automate, or manual refresh.

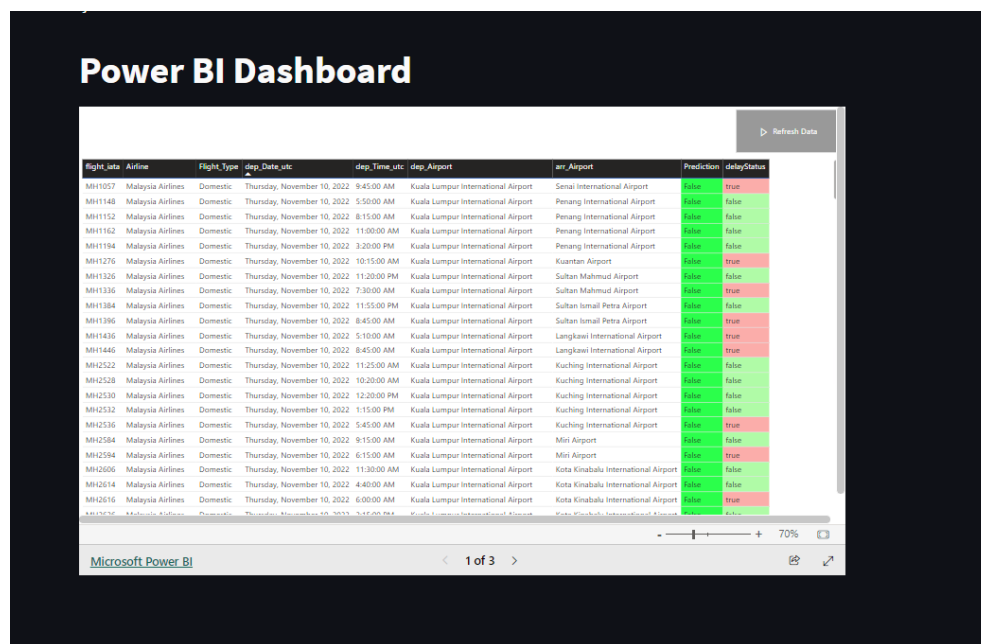
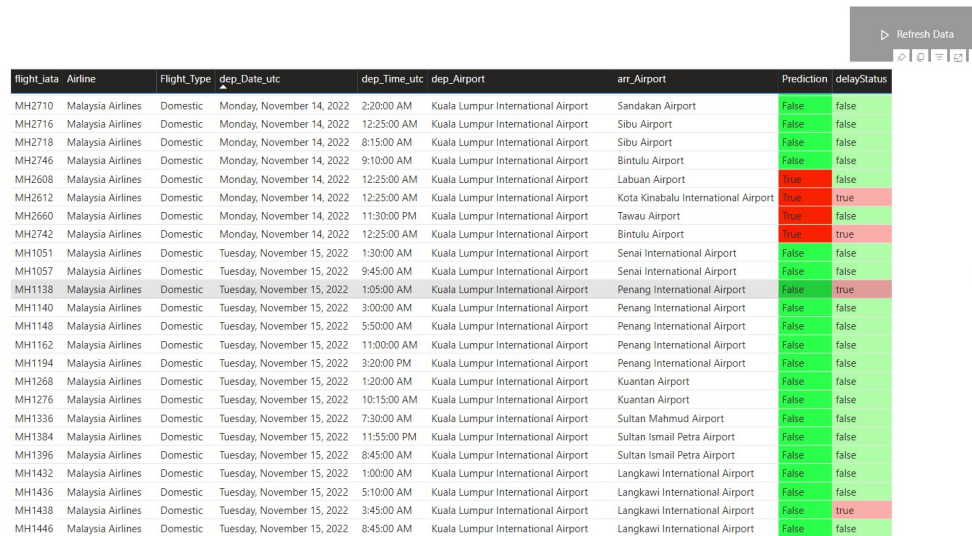


Figure 4.26: Power BI in Streamlit

Power BI plays a crucial role in displaying the results of the prediction process through an interactive and visually appealing dashboard. The dashboard consists

of three pages, each providing insightful information related to the predicted flight delays.



The screenshot shows a web application interface for a flight prediction dashboard. At the top right, there is a button labeled 'Refresh Data' with a circular arrow icon. Below the button is a table with 9 columns: flight_iata, Airline, Flight_Type, dep_Date_utc, dep_Time_utc, dep_Airport, arr_Airport, Prediction, and delayStatus. The table contains 24 rows of flight data. The 'Prediction' column uses green for 'false' and red for 'true'. The 'delayStatus' column uses green for 'false' and red for 'true'. The table is styled with alternating row colors and a dark header.

flight_iata	Airline	Flight_Type	dep_Date_utc	dep_Time_utc	dep_Airport	arr_Airport	Prediction	delayStatus
MH2710	Malaysia Airlines	Domestic	Monday, November 14, 2022	2:20:00 AM	Kuala Lumpur International Airport	Sandakan Airport	false	false
MH2716	Malaysia Airlines	Domestic	Monday, November 14, 2022	12:25:00 AM	Kuala Lumpur International Airport	Sibu Airport	false	false
MH2718	Malaysia Airlines	Domestic	Monday, November 14, 2022	8:15:00 AM	Kuala Lumpur International Airport	Sibu Airport	false	false
MH2746	Malaysia Airlines	Domestic	Monday, November 14, 2022	9:10:00 AM	Kuala Lumpur International Airport	Bintulu Airport	false	false
MH2608	Malaysia Airlines	Domestic	Monday, November 14, 2022	12:25:00 AM	Kuala Lumpur International Airport	Labuan Airport	true	false
MH2612	Malaysia Airlines	Domestic	Monday, November 14, 2022	12:25:00 AM	Kuala Lumpur International Airport	Kota Kinabalu International Airport	true	true
MH2660	Malaysia Airlines	Domestic	Monday, November 14, 2022	11:30:00 PM	Kuala Lumpur International Airport	Tawau Airport	true	false
MH2742	Malaysia Airlines	Domestic	Monday, November 14, 2022	12:25:00 AM	Kuala Lumpur International Airport	Bintulu Airport	true	true
MH1051	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	1:30:00 AM	Kuala Lumpur International Airport	Senai International Airport	false	false
MH1057	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	9:45:00 AM	Kuala Lumpur International Airport	Senai International Airport	false	false
MH1138	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	1:05:00 AM	Kuala Lumpur International Airport	Penang International Airport	false	true
MH1140	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	3:00:00 AM	Kuala Lumpur International Airport	Penang International Airport	false	false
MH1148	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	5:50:00 AM	Kuala Lumpur International Airport	Penang International Airport	false	false
MH1162	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	11:00:00 AM	Kuala Lumpur International Airport	Penang International Airport	false	false
MH1194	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	3:20:00 PM	Kuala Lumpur International Airport	Penang International Airport	false	false
MH1268	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	1:20:00 AM	Kuala Lumpur International Airport	Kuantan Airport	false	false
MH1276	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	10:15:00 AM	Kuala Lumpur International Airport	Kuantan Airport	false	false
MH1336	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	7:30:00 AM	Kuala Lumpur International Airport	Sultan Mahmud Airport	false	false
MH1384	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	11:55:00 PM	Kuala Lumpur International Airport	Sultan Ismail Petra Airport	false	false
MH1396	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	8:45:00 AM	Kuala Lumpur International Airport	Sultan Ismail Petra Airport	false	false
MH1432	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	1:00:00 AM	Kuala Lumpur International Airport	Langkawi International Airport	false	false
MH1436	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	5:10:00 AM	Kuala Lumpur International Airport	Langkawi International Airport	false	false
MH1438	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	3:45:00 AM	Kuala Lumpur International Airport	Langkawi International Airport	false	true
MH1446	Malaysia Airlines	Domestic	Tuesday, November 15, 2022	8:45:00 AM	Kuala Lumpur International Airport	Langkawi International Airport	false	false

Figure 4.27: Prediction Dashboard Page 1

On the first dashboard page, a comprehensive table is presented, showcasing the flight information of the predicted dataset. The table includes columns such as *Flight_IATA*, *Airline*, *Flight Type*, *Dep_Date_Utc*, *dep_Airport*, *arr_Airport*, *Prediction Result*, and Original Delay Information. By examining this table, users can easily compare the predicted delay status with the original delay status, gaining valuable insights into the accuracy of the predictions. Additionally, a Power Automate button is included in the top-right corner, allowing users to reset the dataset when needed.

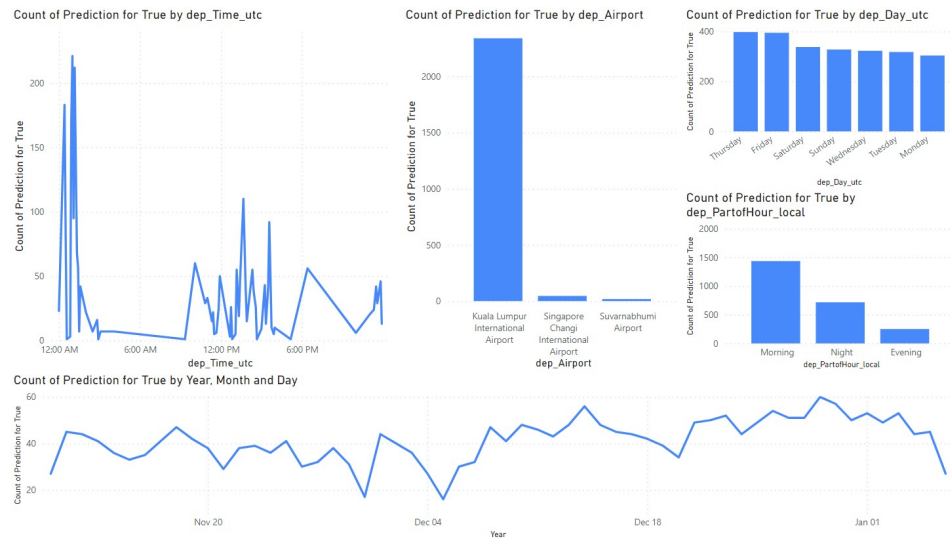


Figure 4.28: Prediction Dashboard Page 2

Moving to the second dashboard page, the focus shifts to analyzing delay information within the predicted dataset. The first chart, a line chart, displays the count of all predicted delayed flights against the departure time. This chart provides an overview of how the predicted delays are distributed throughout different times of the day. The second chart, a bar chart, depicts the count of predicted delay flights against the departure airport, enabling users to identify airports with a higher frequency of predicted delays. The third chart, another bar chart, presents the count of predicted delay flights categorized by the departure day of the week (Monday, Tuesday, etc.), allowing users to observe any patterns or trends. The fourth chart visualizes the count of predicted delay flights based on the departure

part of the day (Morning, Night, Evening, Afternoon), offering insights into the temporal distribution of delays. Lastly, a chart displays the count of all predicted delay flights over time, with the x-axis representing the count of predicted delays and the y-axis representing the dates of the flights.

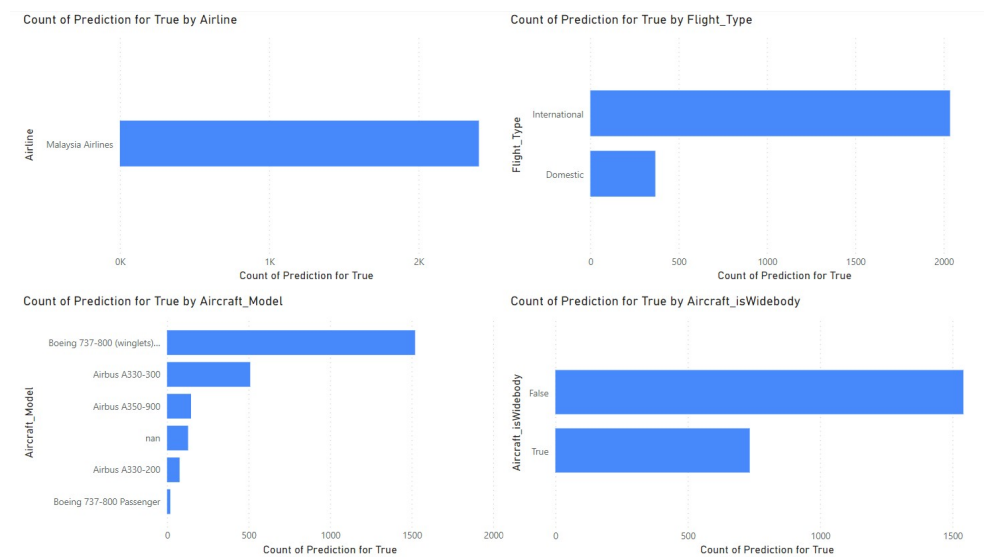


Figure 4.29: Prediction Dashboard Page 3

The third and final dashboard page focuses on flight information. The first bar chart showcases the count of all predicted delay flights grouped by airline. This chart enables users to identify airlines with a higher occurrence of predicted delays. The second bar chart presents the count of predicted delay flights categorized by flight type, providing insights into the impact of flight types on delays. The third bar chart visualizes the count of predicted delay flights based on the aircraft model,

allowing users to identify specific models associated with a higher likelihood of delays. Lastly, a bar chart illustrates the count of predicted delay flights based on aircraft widebody information, offering insights into the impact of widebody aircraft on delays.

The selection of these specific chart types and visualizations is driven by their ability to effectively convey the underlying data patterns and facilitate meaningful insights. Each chart is carefully designed with a title and clearly labeled x-axis and y-axis to provide clarity and context for interpretation. By utilizing these charts, users can easily explore and analyze the predicted delay information, uncovering valuable trends and patterns related to flight delays.

5 Conclusion

By achieving three primary objectives, this project aimed to understand and predict flight disruptions at Kuala Lumpur International Airport. First and foremost, the first objective to identify the variables contributing to flight disruptions at Kuala Lumpur International Airport was achieved. The project successfully uncovered the specific factors that have an impact on flight disruption through the analysis of a comprehensive dataset. With the use of the feature selection algorithm in the dataset, the main variables contributing to flight disruption were able to identify. By utilising this knowledge, the models were able to focus on the features with the most relevant aspects and increase the accuracy of models and minimise flight disruptions.

The second objective of the project was to uncover the hidden relationships between the identified variables. The project revealed patterns, correlations, and dependencies among the variables by using Exploratory Data Analysis (EDA). This understanding of the relationships between various elements provides critical insights into the complex patterns of flight disruptions. Airlines can use this data to gain a better understanding of the root causes of disruptions and performance measures mitigation plans.

The final objective of the project was to construct a predictive model that accurately forecasts flight disruptions at Kuala Lumpur International Airport. A reliable predictive model that can forecast the likelihood of disruptions was successfully developed throughout the project. The model enables the airlines to manage the disruptions by creating effective plans to reduce flight disruptions such as rescheduling flights, adjusting resources, and implementing contingency plans.

Overall, this project has made significant advances in understanding and predicting flight delays at Kuala Lumpur International Airport. The project provides valuable insights and tools for airlines to effectively manage and mitigate flight disruptions by identifying key variables, uncovering relationships, and constructing a predictive model. These insights enable airlines to make more informed decisions, optimise their operations, improve punctuality, reduce costs, and increase customer satisfaction.

I've learned a lot of important things throughout the project's duration. The first and most important thing I've learned is how to use Power BI. Many Power BI functions were unfamiliar to me; at first, I knew nothing about the Power BI's functionality; however, as the project progressed, the Power BI's functionality was gradually discovered and more functions were used. Finally, I was able to fully comprehend the flow or structure of Power BI. The API is the second most

important thing I've learned; many applications rely on it to communicate with one another. This project made extensive use of APIs such as OpenMeteo API (for weather), BestTime API (for foot traffic), Google Console Cloud API (for linking with Google Sheets), and Azure Directory API (embed Power BI in jupyter notebook). Because I had never worked with APIs before, it took a lot of time and effort to understand them. Aside from that, this project allowed me to hone my data mining skills in areas such as data pre-processing, feature selection, and exploratory data analysis.

While the project's main goals have been met, there are other areas where future improvement and expansion might be implemented. The prediction model can be integrated with more data sources in the future. With more advanced sources of data, more factors can be investigated providing the model with more information and increasing the efficiency of the current model. The model's ability to predict can be improved by enhancing the dataset, allowing airlines to better anticipate and manage delays. On top of that, as new data is available, the models should be fed with the newer data to make sure the predictions are always up to date. This continuous approach will allow the model to adapt to changing patterns and dynamics in flight disruptions, ensuring its long-term usefulness.

To further enhance the system, It is important to have a more user-friendly

interface or dashboard. The current dashboard makes use of the Streamlit Application, which is not quite user-friendly to all users. In the future, maybe a more user-friendly phone application or web application can be created to enhance the user experience. The interface can be made using Android Studio and Jhtml to provide a better experience to the user. This interface would enable airlines to access and interpret the insights provided by the model with ease. Visualizations, interactive features, and real-time updates can facilitate effective decision-making and prompt actions in response to disruptions. Moreover, in the future there should be more collaboration with airport authorities, air traffic control, and other stakeholders, as they can provide additional data sources and valuable insights to refine the model and improve its predictions.

There are bright prospects for this software package in the future. The system can be expanded to include more airports or airline companies, providing a comprehensive solution for controlling flight interruptions in a variety of locations. This would require modifying the model to the specific characteristics and variables of each airport or airline, hence improving its applicability and usefulness in the aviation industry. Furthermore, forming partnerships and working together with businesses can make it easier to obtain extra data from airlines and airports. Collaborating closely with these partners will not only improve the predictive

model's accuracy but will additionally promote knowledge sharing and develop a collaborative atmosphere for managing flight disruptions on a larger scale.

Finally, this study was successful in terms of understanding and predicting flight problems at Kuala Lumpur International Airport. This project's insights and developments help airlines more effectively manage disruptions, optimise operations, and improve passenger satisfaction. The project utilised program-specific data analysis, and modelling, along with evaluation abilities and knowledge to illustrate the actual execution of these techniques in the aviation sector. Moving forward, there is the possibility of additional improvements and enhancements to the application programme to benefit additional airports and airline partners.

References

- Almaameri, I. M., & Mohammed, A. (2022, 05). *Predicting airplane flight delays using neural networks*. Retrieved from <https://ieeexplore.ieee.org/document/9888363> doi: 10.1109/IICETA54559.2022.9888363
- Anees, A., & Huang, W. (2021, 09). *Flight delay prediction: Data analysis and model development*. Retrieved from <https://ieeexplore.ieee.org/abstract/document/9594260> doi: 10.23919/ICAC50006.2021.9594260
- Balamurugan, R., Maria, A. V., Baranidaran, G., MaryGladence, L., & Revathy, S. (2022, 04). *Error calculation for prediction of flight delays using machine learning classifiers*. Retrieved from <https://ieeexplore.ieee.org/document/9776709> doi: 10.1109/ICOEI53556.2022.9776709
- Ballakur, A. A., & Arya, A. (2020, 10). Empirical evaluation of gated recurrent neural network architectures in aviation delay prediction. doi: 10.1109/icccs49678.2020.9276855
- Beth. (2022, 02). Storm eunice: Flights and train services cancelled. *BBC*. Retrieved from <https://www.bbc.com/news/business-60430197>
- Borsky, S., & Unterberger, C. (2019, 06). Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation*, 18, 10-26. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2212012218300753> doi: 10.1016/j.ecotra.2019.02.002
- Cai, K., Li, Y., Fang, Y.-P., & Zhu, Y. (2021). A deep learning approach for flight delay prediction through time-evolving graphs. *IEEE Transactions on Intelligent Transportation Systems*, 1-11. doi: 10.1109/tits.2021.3103502
- Chakrabarty, N. (2019, 03). *A data mining approach to flight arrival delay prediction for american airlines*. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8876970> doi: 10.1109/IEMCONX.2019.8876970
- Czerby, A. I., Fu, X., Zheng, L., & Tae H., O. (2021, 01). Post pandemic aviation market recovery: Experience and lessons from china. *Journal of Air Transport Management*, 90, 101971. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0969699720305548> doi: 10.1016/j.jairtraman.2020.101971
- Dhanawade, R., Deo, M., Khanna, N., & Deolekar, R. V. (2019, 03). *Analyzing factors influencing flight delay prediction*. Retrieved 2023-07-04, from <https://ieeexplore.ieee.org/document/8991208>

- Dietz, S. J., Kneringer, P., Mayr, G. J., & Zeileis, A. (2018, 11). Correction to: Forecasting low-visibility procedure states with tree-based statistical methods. *Pure and Applied Geophysics*, 176, 2645-2658. doi: 10.1007/s00024-018-1993-8
- Gu, Y., & Yang, J. (2019, 03). *Research on cause and governance path of passenger disturbance in flight delay in terminal*. Atlantis Press. Retrieved from <https://www.atlantis-press.com/proceedings/iafsm-18/55915239> doi: 10.2991/iafsm-18.2019.18
- Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020, 01). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69, 140-150. doi: 10.1109/tvt.2019.2954094
- Hopane, J., & Gatsheni, B. (2019, 12). *A computational intelligence-based prediction model for flight departure delays*. Retrieved from <https://ieeexplore.ieee.org/document/9071393> doi: 10.1109/CSCI49370.2019.00107
- Hu, P., Zhang, J.-P., & Li, N. (2021, 10). Research on flight delay prediction based on random forest. doi: 10.1109/iccasit53235.2021.9633476
- Huo, J., Keung, K. L., Lee, C. K. M., Ng, K. K. H., & Li, K. (2020, 12). The prediction of flight delay: Big data-driven machine learning approach. *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. doi: 10.1109/ieem45057.2020.9309919
- Jiang, Y., Liu, Y., Liu, D., & Song, H. (2020, 08). Applying machine learning to aviation big data for flight delay prediction. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. doi: 10.1109/dasc-picom-cbdcom-cybercitech49142.2020.00114
- Jiang, Y., Miao, J., Zhang, X., & Le, N. (2020, 10). *A multi-index prediction method for flight delay based on long short-term memory network model*. Retrieved from <https://ieeexplore.ieee.org/document/9368554> doi: 10.1109/ICCASIT50869.2020.9368554
- Kalyani, N. L., Jeshmitha, G., Sai U., B. S., Samanvitha, M., Mahesh, J., & Kiranmayee, B. (2020, 10). *Machine learning model - based prediction of flight delay*. Retrieved from <https://ieeexplore.ieee.org/document/9243339/> doi: 10.1109/I-SMAC49090.2020.9243339

- Keoni. (2023, 05). *Video shows crosswinds force taiwan starlux plane to abort narita landing | taiwan news | 2023-05-08 10:48:00*. Retrieved from <https://www.taiwannews.com.tw/en/news/4885071>
- Khaksar, H., & Sheikholeslami, A. (2017, 12). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, 0, 0-0. doi: 10.24200/sci.2017.20020
- Kulkarni, R., Jenamani, R. K., Pithani, P., Konwar, M., Nigam, N., & Ghude, S. D. (2019, 04). Loss to aviation economy due to winter fog in new delhi during the winter of 2011–2016. *Atmosphere*, 10, 198. doi: 10.3390/atmos10040198
- Li, Q., & Jing, R. (2021, 06). Generation and prediction of flight delays in air transport. , 15, 740-753. doi: 10.1049/itr2.12057
- Liu, F., Sun, J., Liu, M., Yang, J., & Gui, G. (2020, 05). Generalized flight delay prediction method using gradient boosting decision tree. *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. doi: 10.1109/vtc2020-spring48590.2020.9129110
- Mang, C., & Chen, Y. (2020, 06). Research on flight delay prediction based on multi-model fusion. *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. doi: 10.1109/itoec49072.2020.9141816
- Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020, 02). *Predicting flight delays with error calculation using machine learned classifiers*. Retrieved from https://ieeexplore.ieee.org/abstract/document/9071159?casa_token=Ytr4rHcMGyUAAAAA:WAJibYubaslGAhmhhEMX0I688eva1F4eqrDRv_W0wZs4HMoM-YJCEsmgWW22nxQvBaWc3mBurHwv doi: 10.1109/SPIN48934.2020.9071159
- Pamplona, D. A., Weigang, L., de Barros, A. G., Shiguemori, E. H., & Alves, C. J. P. (2018, 07). *Supervised neural network with multilevel input layers for predicting of air traffic delays*. Retrieved from <https://ieeexplore.ieee.org/document/8489511> doi: 10.1109/IJCNN.2018.8489511
- Schuldt, S. J., Nicholson, M. R., Adams, Y. A., & Delorit, J. D. (2021, 01). Weather-related construction delays in a changing climate: A systematic state-of-the-art review. *Sustainability*, 13, 2861. Retrieved from <https://www.mdpi.com/2071-1050/13/5/2861> doi: 10.3390/su13052861
- Shu, Z. (2021, 12). *Analysis of flight delay and cancellation prediction based on machine learning models*. Retrieved from <https://ieeexplore.ieee.org/abstract/document/9731090> doi: 10.1109/MLBDBI54094.2021.000056

- Tao, J., Man, H., & Yanling, L. (2021, 10). Flight delay prediction based on lightgbm. *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*. doi: 10.1109/iccasit53235.2021.9633431
- Wang, H. (2022, 01). *Big data visualization and analysis of various factors contributing to airline delay in the united states*. Retrieved from <https://ieeexplore.ieee.org/document/9758536> doi: 10.1109/BDICN55575.2022.00042
- Wang, J., & Pan, W. (2022, 07). *Flight delay prediction based on arima*. Retrieved from <https://ieeexplore.ieee.org/document/9853327> doi: 10.1109/ICCEAI55464.2022.00047
- Wang, K., Li, J., & Tian, Y. (2019, 10). *Airport delay prediction method based on improved weather impacted traffic index*. Retrieved from <https://ieeexplore.ieee.org/document/8973213> doi: 10.1109/ICCASIT48058.2019.8973213
- Wang, T., Lin, L., & Gao, J. (2021, 09). *Explainable multi-task flight arrival delay prediction*. Retrieved from <https://ieeexplore.ieee.org/document/9564930> doi: 10.1109/ITSC48978.2021.9564930
- Wang, Y., & Li, Y. (2020). Complexity analysis on the influence factors of the flight delay risk based on sna. *Open Journal of Social Sciences*, 08, 54-71. doi: 10.4236/jss.2020.85005
- Wu, W., Cai, K., Yan, Y., & Li, Y. (2019, 09). *An improved svm model for flight delay prediction*. Retrieved from <https://ieeexplore.ieee.org/document/9081611> doi: 10.1109/DASC43569.2019.9081611
- Xu, M., Wang, M., Wang, Y., & Delahaye, D. (2022, 10). *Robust estimation of airport declared capacity*. Retrieved from <https://ieeexplore.ieee.org/document/9922312> doi: 10.1109/ITSC55140.2022.9922312
- Yanying, Y., Mo, H., & Haifeng, L. (2019, 01). A classification prediction analysis of flight cancellation based on spark. *Procedia Computer Science*, 162, 480–486. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1877050919320241> doi: 10.1016/j.procs.2019.12.014
- Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020, 11). Flight delay prediction based on deep learning and levenberg-marquart algorithm. *Journal of Big Data*, 7. doi: 10.1186/s40537-020-00380-z
- Yiu, C. Y., Ng, K. K. H., Kwok, K. C., Tung Lee, W., & Mo, H. T. (2021, 10). *Flight delay predictions and the study of its causal factors using machine learning algorithms*. Retrieved from <https://ieeexplore.ieee.org/document/9633571> doi: 10.1109/ICCASIT53235.2021.9633571

- Zhou, F., Jiang, G., Lu, Z., & Wang, Q. (2022, 02). Evaluation and analysis of the impact of airport delays. *Scientific Programming*, 2022, e7102267. Retrieved from <https://www.hindawi.com/journals/sp/2022/7102267/> doi: 10.1155/2022/7102267
- Zhou, H., Li, W., Jiang, Z., Cai, F., & Xue, Y. (2022, 07). Flight departure time prediction based on deep learning. *Aerospace*, 9, 394. doi: 10.3390/aerospace9070394