Flight Disruption Prediction For Kuala Lumpur International Airport

Yi-Hong Leong, Choo-Yee Ting

Faculty of Computing and Informatics, Multimedia University, 63000 Cyberjaya, Selangor

Abstract - Flight disruption has been a major issue for both the airlines and passengers. Various methods have been employed by researchers to overcome the challenge. Among the attempts were the statistical and machine learning approaches. In this study, we attempted a Feature Selection approach 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. The data consists of 17 features and 60 thousand rows. The dataset features include the airline info, departure info, arrival info, flight duration, and delayed information. We supplemented the data with extra information such as weather, flight model and foot traffic. 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 with an accuracy of 93.1%.

Keywords—Flight disruption, Prediction, Machine Learning Algorithm, Data Analysis, Risk Management

I. INTRODUCTION

Flight disruption happens when a flight takes off or arrives later than the scheduled time that occurs in most airlines around the world causing enormous economic losses for an airline company, and inconveniences for passengers (Gui et al., 2020). Flight cancellations, departure and arrival delays can occur (Hassan et al., 2020). According to Toolkit (2021), flight disruption is the collective term for the event that prevents passengers from completing the itineraries on time such as flight and tarmac delays, flight cancellations and denials of boarding. The immediate impact of flight disruption on a passenger is that the original travel plan is altered and often ends up with a rearrangement of the travel itinerary. Flight disruption will often ruin passengers' travel plans and end up causing disappointment to them. The occurrence of Flight disruption will be costing enormous economic losses for an airline company, and bringing huge inconvenience for passengers (Gui et al., 2020). The main factors that cause passenger disturbance are analyzed, including the length of flight delay time, the flight departure time in the schedule, current time of flight delay, crowd density of boarding gate, airline service quality and management ability (Gu & Yang, 2019). According to Chakrabarty (2019), delays make passengers lose their trust in such a famous and internationally recognized airline. Flight disruptions are also disliked by airlines. The occurrence of flight disruption has a significant financial and reputational impact on the airlines, such as fuel usage, airport fees, passenger compensation, overtime pay, and so on. According to Yazdi, Kamel, Chabok, & Kheirabadi (2020), flight disruption is inevitable, and it plays an important role in both profits and losses of the airlines. Flight delays and other disruptions result in additional expenses such as extra parking time for the plane, fuel charges, and personnel overtime pay. The annual cost of flight delays to the global economy was estimated to be $50 billion in 2019 (Cai et al., 2022).

The main objectives of this study are focused on understanding and predicting flight disruption at KLIA. Firstly, the aim is to identify the variables contributing to airport flight disruptions. We can gain insights into the key drivers behind flight disruptions by analyzing various factors such as weather conditions, operational issues, and other relevant variables. This deeper understanding will provide valuable insights into the complex dynamics that lead to flight disruptions at KLIA. The final objective is to construct a predictive model that can accurately forecast flight disruptions at KLIA. By leveraging machine learning algorithms and utilizing the identified variables and their relationships, we aim to develop a reliable model capable of predicting the likelihood of flight disruptions. Such a model can serve as a proactive tool for airlines and airport authorities, enabling them to take preventive measures, allocate resources effectively, and mitigate the impact of disruptions on passengers and operations.

II. LITERATURE REVIEW (10-Font size, Times New Roman)

A. Factors Causing Flight Disruption

A variety of factors can cause flight disruptions. These factors include air traffic, weather, aircraft issues, airport infrastructure, etc. Most of these occurrences are unexpected and will occur on occasion. The factors causing flight disruption to have included the length of flight delay time, the flight departure time in the schedule, current time of flight delay, crowd density of boarding gate, service quality of airline service and management ability (Gu & Yang, 2019). Flight disruptions are a major concern for all those involved in the aviation industry - from airport terminals to airline companies and even passengers. Flight disruptions can also have a ripple effect on flight operations and airport resources. If there is no reasonable method to control the propagation of delays, the spread of delays will continue to expand (Zhou et al., 2022).

Table 1. Summary of findings of researcher's factors

Factors **Author**	Weather	Airport	Airline	Network	Security	Aviation Control	Passenger	Time	Passenger Load	Aircraft
(Li & Jing, 2021)	✓			✓						
(Zhou, Li, Jiang, Cai, & Xue, 2022)	✓	✓	✓							
(Wang Y., Li Y., 2020)	✓		✓		✓	✓	✓			
(Zhou, Jiang, Lu, & Wang, 2022)				✓						
(Yazdi, M.F., Kamel, S.R., Chabok & S.J.M. et al. , 2020)		✓						✓	✓	
(Cai, Li, Fang & Zhu, 2022)				✓						
(Jiang, Liu, Liu & Song, 2020)	✓									
(Kalyani, Jeshmitha, Sai, Samanvitha, Mahesh & Kiranmayee, 2020)	✓									
(Khaksar & Sheikholeslami, 2019)	✓									✓
(Shu, 2021)	✓		✓		✓	✓				
(Yanying, Mo, & Haifeng, 2019)		✓	✓					✓		
(Jiang, Miao, Zhang, & Le, 2020)	✓	✓						✓		✓
(Wang & Pan, 2022)								✓		
(Hu, Zhang, & Li, 2021)								✓		
(Liu, Sun, Liu, Yang, & Gui, 2020)	✓	✓		✓				✓		
(T. Wang, Lin, & Gao, 2021)	✓									
(K. Wang, Li, & Tian, 2019)	✓									
(Kalyani et al., 2020)	✓									
(Mang & Chen, 2020)	✓									
(Gui et al., 2020)	✓	✓						✓		
(Anees & Huang, 2021)	✓		✓		✓	✓				
(Almaameri & Mohammed, 2022)	✓									
(H. Wang, 2022)	✓		✓					✓		
(Wu, Cai, Yan, & Li, 2019)		✓	✓							✓
(Yiu, Ng, Kwok, Tung Lee, & Mo, 2021)	✓									

Weather conditions are among the primary factors contributing to flight disruptions due to their inherent variability. Weather conditions significantly impact flight operations, causing delays and disruptions. Strong winds, such as crosswinds, can deflect planes off course and affect landing. In fact, the weather has been identified as a crucial parameter influencing flight delays, making it an essential aspect for predicting and managing flight dispruptions (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 pose risks with electrical system damage and reduced visibility. As a precautionary measure, flight crews actively avoid areas affected by thunderstorms, which can inevitably lead to flight disruptions. Fog hampers pilots' ability to see runways clearly. 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. These weather elements require proper planning and coordination to ensure safety and productivity.

High passenger traffic density at airports leads to challenges in security checks, check-in procedures, and immigration processes. According to Y. Wang & Li (2020),Urban airport areas, characterized by complex flows of people, materials, and aircraft, present various risk factors that can contribute to flight delays during operations. Security screening issues and excessive foot traffic can disrupt operations. Security breaches may lead to terminal clearance and subsequent re-boarding delays (Anees & Huang, 2021). Delays in boarding and long waiting times at check-in counters, immigration, and customs checkpoints contribute to disruptions. Effective management of airport demand and capacity is crucial to reduce congestion and delays. Implementing effective strategic management of airport demand and capacity becomes essential in reducing congestion and delays (Xu, Wang, Wang, & Delahaye, 2022).

Flight disruptions can be influenced by factors related to airport operations. Air traffic control delays, runway maintenance or closure, and inefficient ground handling services can affect efficiency. Infrastructure problems and construction activities also impact flight operations. Airlines' factors include aircraft maintenance issues, crew availability, inefficient scheduling, and planning. Disruptions in airline networks and passenger-related factors like overbooking or security concerns further contribute to disruptions.

In summary, weather conditions, high passenger traffic, airport operations, airline-related aspects, network dynamics, and passenger loads significantly impact flight operations and cause disruptions. Effective management strategies addressing airport demand, capacity, and passenger flow are crucial to mitigate congestion, reduce delays, and minimize the impact on airlines, airports, and passengers.

B. Machine Learning Techniques for Flight Disruption Prediction

Flight disruptions and cancellations are a significant source of concern for airlines, passengers, and airport operators. These disruptions can have a significant economic impact as well as cause inconvenience for all parties involved. There has been an increase in interest in developing methods to predict, prevent, and mitigate flight disruptions in recent years.

To address flight disruptions, a wide range of methods have been proposed in the literature, including statistical modelling, machine learning, and optimization techniques. Researchers developed models that can predict flight disruptions using data from various sources, such as flight and weather data. Furthermore, many studies have focused on identifying the key factors that contribute to flight disruptions and developing mitigation strategies. This literature review aims to provide an overview of the current state of research on methods for addressing flight disruptions and identify areas for future research.

Table 2. Summary of findings of researcher's methods

Author	Visualization	K-Nearest Neighours	Bayesian Modelling	XGBoost	Logistic Regression	Linear Regression	Gradient Boosting Classifier	Multilayer Perceptron	Decision Tree	Support Vector Machine	Random Forest	Arima Model	Neural Network	Causal Analysis	Levenburg-Marquat algorithm	Social Network Analysis
(Li & Jing, 2021)											✓					
(Zhou, Li, Jiang, Cai, & Xue, 2022)													✓			
(Wang Y., Li Y., 2020)																✓
(Zhou, Jiang, Lu, & Wang, 2022)														✓		
(Yazdi, M.F., Kamel, S.R., Chabok & S.J.M. et al. , 2020)															✓	
(Cai, Li, Fang & Zhu, 2022)													✓			
(Jiang, Liu, Liu & Song, 2020)								✓	✓	✓	✓					
(Chakrabarty, 2019)							✓									
(Kalyani,Jeshmitha, Sai, Samanvitha, Mahesh & Kiranmayee, 2020)				✓		✓										
(Khaksar & Sheikholeslami, 2019)		✓	✓						✓		✓					
(Shu, 2021)			✓		✓	✓										
(Yanying, Mo, & Haifeng, 2019)			✓		✓				✓	✓						
(Jiang, Miao, Zhang, & Le, 2020)													✓			
(Tao, Man, & Yanling, 2021)									✓							
(Wang & Pan, 2022)												✓				
(Hu, Zhang, & Li, 2021)											✓					
(Liu, Sun, Liu, Yang, & Gui, 2020)									✓							
(T. Wang, Lin, & Gao, 2021)				✓		✓										
(K. Wang, Li, & Tian, 2019)											✓					
(Kalyani et al., 2020)				✓		✓										
(Mang & Chen, 2020)				✓							✓					
(Balamurugan, Maria, Baranidaran, MaryGladence, & Revathy, 2022)			✓		✓				✓		✓					
(Gui et al., 2020)													✓			

(Anees & Huang, 2021)											✓					
(Almaameri & Mohammed, 2022)									✓	✓			✓			
(H. Wang, 2022)	✓															
(Ballakur & Arya, 2020)													✓			
(Wu, Cai, Yan, & Li, 2019)										✓						
(Meel, Singhal, Tanwar, & Saini, 2020)			✓		✓		✓		✓		✓					
(Yiu, Ng, Kwok, Tung Lee, & Mo, 2021)		✓	✓						✓		✓		✓			
(Huo, Keung, Lee, Ng, & Li, 2020		✓	✓		✓				✓		✓					
(Hopane & Gatsheni, 2019)		✓						✓	✓	✓						
(Pamplona, Weigang, de Barros, Shiguemori, & Alves, 2018)													✓			
(Chakrabarty, 2019)							✓									

The authors mentioned in the table have employed various techniques to address flight disruption problems at KLIA. One of the method is the development of prediction models. These models analyze a wide range of data to proactively identify potential disruptions and allow for timely interventions. By understanding the factors contributing to disruptions and their interrelationships, prediction models assist in identifying root causes and devising appropriate solutions. They optimize resource allocation, such as staff and equipment, to minimize the impact of disruptions. The prediction models provide valuable insights and forecasts, enabling informed decision-making in managing disruptions. By adopting data-driven approaches, airport authorities can allocate resources effectively, minimize disruption consequences, and improve operational efficiency and reliability.

Numerous studies have been conducted by researchers to predict flight disruptions using various modelling approaches. Jiang 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 & Sheikholeslami (2019) 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. Yanying et al. (2019) predicted flight cancellations using logistic regression, support vector machine, naive Bayes, and decision tree algorithms. 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. 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 tree, random forest, 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. Almaameri & 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. Chakrabarty (2019) focused on analyzing flight information of US domestic flights operated by American Airlines and predicting possible arrival delays using data mining and machine learning approaches. These studies contribute to the advancement of predictive models for flight disruptions and provide insights into improving operational efficiency and passenger experience.

In summary, the authors leverage techniques such as prediction models ande exporatory data analysis to address flight disruption issues at KLIA. These approaches contribute to the enhancement of airport operations through optimized resource allocation, informed decision-making, and the mitigation of disruption risks.

III. DATASET DESCRIPTION

A. Flight Dataset

The dataset included all the information on flights that operated mainly at KLIA for some time. The flight dataset contained 60511 rows of data, each representing a unique flight. The dataset included 17 features or variables. 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), codesharing 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 flight data set was a key source of information for our analysis. It contained information about each flight that mainly operated at Kuala Lumpur International Airport (KLIA), including the airlines, flight number, departure times, arrival times, destination, flight durations and flight delayed information. The purpose of the flight dataset was to gather information on flights that operated at Kuala Lumpur International Airport (KLIA) to analyze flight disruptions at the airport. One potential application of the flight dataset is to identify factors that contribute to flight delays or disruptions at KLIA.

By analyzing the data, it may be possible to identify patterns or trends that could be addressed to improve the reliability of flights at the airport. Another use of the flight dataset could be to develop predictive models that forecast future flight disruptions or delays based on past data. This could be useful for airport management and airlines in planning and preparing for potential disruptions, as well as for passengers in making travel plans. The flight dataset could also be used to analyze the performance of different aircraft types or airlines operating at KLIA. By comparing flight durations, disruptions, and other metrics, it may be possible to identify which aircraft or airlines are performing best and to identify areas for improvement.

The flight dataset was cleaned by identifying and correcting errors or inconsistencies in the data, and by handling missing values. The data in the dataset was cleaned and preprocessed to correct errors, handle missing values, and ensure consistent formatting of timestamps and encoding of categorical variables. The cleaned and preprocessed flight dataset provided a detailed and reliable source of information for the analysis of flight disruptions at KLIA. Overall, the flight data set provided a wealth of information that was crucial to our analysis of flight disruptions at KLIA.

B. Weather Dataset

According to Kalyani et al. (2020), the weather was found to affect the delay to a great extent after considering all the parameters that are the cause for the delay and hence used as a contributing aspect to predict the delay of the flight. To collect the weather dataset, the weather datasets were crawled using the online API called Open Meteo.

All the inbound and outbound airport weather data were crawled using the online API. This airport weather dataset included hourly weather observations for the airports. In total, the weather dataset contained 244368 rows of data, each representing a unique weather observation. The weather dataset includes data from 180 airports. Each airport will have a range of 7 days and 24 hours of observation of weather data. The dataset included 14 features or variables, such as airport location, date and time, temperature, humidity, sea level pressure, precipitation, rain, snowfall, cloud cover, wind speed, wind direction and wind gusts. The weather data set was another important source of information for our analysis.

By analyzing the data, it may be possible to identify specific weather conditions or patterns that are associated with higher rates of disruptions or cancellations and to assess the relative importance of different weather variables in predicting disruptions. This type of analysis could be useful for airport management, airlines, and other stakeholders in the aviation industry in planning and preparing for potential disruptions, as well as for passengers in making travel plans.

The weather dataset could also be used in combination with other datasets, such as the flight dataset, to explore more complex relationships between weather and flight disruptions. For example, it may be possible to identify specific weather conditions that are most likely to lead to disruptions or cancellations or to assess the relative importance of different weather variables in predicting disruptions. The weather dataset could be used to develop predictive models that forecast weather conditions at KLIA in the future. This could be useful for airport management, airlines, and other stakeholders in planning and preparing for potential disruptions, as well as for passengers in making travel plans. The weather dataset was cleaned by identifying and correcting errors or inconsistencies in the data, and by handling missing values. Basic preprocessing was also performed, such as converting temperature values to a consistent scale and encoding categorical variables. Overall, the weather dataset provided valuable information about the weather conditions at KLIA, which was important for understanding the impact of weather on flight disruptions.

C. Foot Traffic Dataset

According to (Xu, Wang, Wang, & Delahaye, 2022), effective strategical management for airport demand and capacity to reduce congestion and delays is necessary. To collect the foot traffic dataset, the datasets were crawled using an online API called BestTimes. The foot traffic data from all the airports were crawled using the API. This dataset included hourly observations of foot traffic from all the outbound and inbound airports over weeks. In total, the foot traffic datasets contained 19656 rows of data each representing unique observations of foot traffic per hour. The datasets only include a total of 7 days and 24 hours of data for all the airports. On top of that, the foot traffic dataset contains 7 features or variables such as Airports name, Longitude, Latitude, Day (Monday, Tuesday, Wednesday, ...), Hour, Intensity Status, and Busyness. The foot traffic dataset was another important source of information for our analysis. It contained information about the passenger demand at Kuala Lumpur International Airport (KLIA) over one year, including the number of passengers arriving and departing, and the number of vehicles entering and exiting the airport.

The purpose of using the foot traffic dataset for analyzing flight disruptions at Kuala Lumpur International Airport (KLIA) is to understand how passenger demand may impact the reliability of flights at the airport. By analyzing the data, it may be possible to identify trends or patterns in passenger demand that are associated with higher rates of disruptions or cancellations and to assess the relative importance of passenger demand compared to other factors such as weather or aircraft performance. This type of analysis could be useful for airport management, airlines, and other stakeholders in the aviation industry in planning and preparing for potential disruptions, as well as for passengers in making travel plans. In addition to analyzing the impact of passenger demand on flight disruptions at KLIA, the foot traffic dataset could also be used to study trends and patterns in travel to and from the airport. For example, the dataset could be used to identify the busiest times of the day or years for passengers or to understand the types of vehicles being used to access the airport.

The foot traffic dataset could also be used in combination with other datasets, such as the flight dataset, to explore more complex relationships between passenger demand and flight disruptions. For example, it may be possible to identify the impact of different demand levels on the likelihood of disruptions or cancellations or to assess the relative importance of passenger demand compared to other factors such as weather or aircraft performance. The foot traffic dataset was cleaned by identifying and correcting errors or inconsistencies in the data, and by handling missing values. Some basic preprocessing, such as aggregating the data to a daily level and encoding categorical variables is performed. Overall, the foot traffic dataset provided valuable information about passenger demand at KLIA, which was important for understanding the impact of passenger demand on flight disruptions.

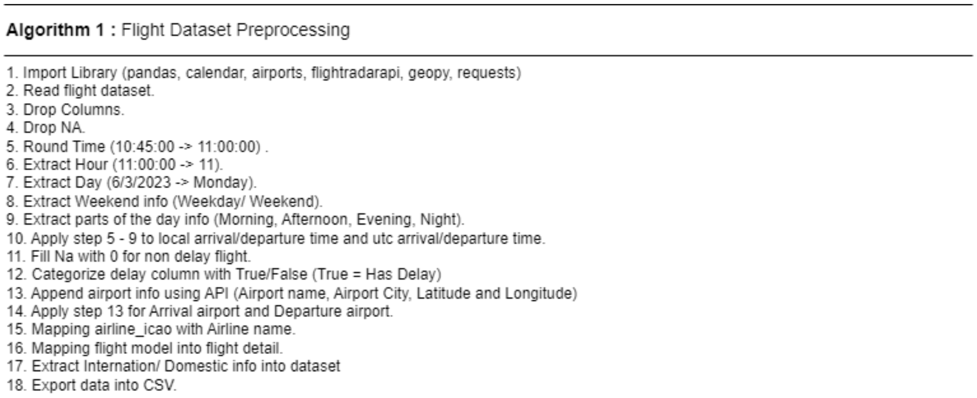
D.Flight Dataset Data Pre-Processing

The first datasets that will be performing the data preprocessing are the flight datasets. The flight datasets contain a lot of information about flight schedules, cancellations, or disruptions. The data first will need to be pre-processed for analysis and modelling later.

The first step involved in the data preprocessing in the flight datasets is the dropping of columns. It includes a lot of irrelevant data columns inside the flight datasets and these columns will not be useful for analysis and modelling later. The columns that will be removed from the current dataset are *airline_iata*, *flight_icao*, *flight_number*, *cs_airline_iata*, *cs_flight_number*, *cs_flight_iata*, and *schedule*. The *airline_iata* and *flight_icao* will be removed from the dataset because the column *flight_iata* has included the information of both columns and in a better way of representation. On top of that, the *flight_iata* column will be removed from the dataset due to is exists another column called *flight_icao*, both columns have the same information but in a way of different representation. Therefore, only retain the *flight_icao* column and remove the *flight_iata* column. The next column that will be removed is *Status*. The *status* column only consists of flights scheduled, diverted, cancelled, landed and active. This will not be helpful for analysis and modelling later. The last section that will be removed is the column with Code Share. Codesharing is done by airlines to expand their reach, by working with other airlines and increasing their route network. Therefore, there will be two different airline codes applied to the same flights. This data will not be relevant for the datasets since only some flights have codeshare and these columns will need to be removed. These code-sharing columns include *cs_airline_iata*, *cs_flight_number* and *cs_flight_iata*.

The next step for the data preprocessing for the flight dataset is removing NA. The lack of departure time and arrival time data will be removed from the dataset. The next part of data pre-processing will be related to the time in the dataset. The first part will be to change the data format of the time column from String to DateTime format for easier processing later. The columns involve changing the format Date column and Time column for the departure and arrival columns. For a better merging with another dataset, time rounding by the hour was performed on the time data. On top of that, the hour of the day and the current day of the data were extracted for merging purposes. Next, the null value of the delayed column inside the dataset was filled with 0, assuming no delays. Subsequently, a delay status column was created, these columns discretize the delayed data into True and False.

To further enhance the dataset, some data discretization was performed on the dataset as well. The parts of the day's data were extracted from the time column. Parts of the day will be useful for the analysis of the flight disruptions later. The parts of the day will be divided into Morning, Afternoon, Evening and Night. On the other hand, the weekend data were extracted from the Day data as well. The data includes Monday, Tuesday, Wednesday, etc..... Other than that, extra info about the airport was inserted into the dataset as well. This data includes the name of the airports, the city of the airport, latitude, and longitude of the airport. The airport data used are data provided by the Airportsdata library. Apart from that, the next info inserted into the dataset is the Airline name, the current *airline_icao* data were converted into the Airline name for a better understanding of the data. The data used was provided by FlightRadar24 API. The next info inserted into the dataset is the Aircraft model. The aircraft model was extracted using FlightStats API. The API will be providing the aircraft model data by just inserting the airline and flight number into the API. Lastly, the domestic and international data of the flight were inserted into the data as well, and the country of the arrival and departure airport were extracted to provide the information on the domestic and international flight data.



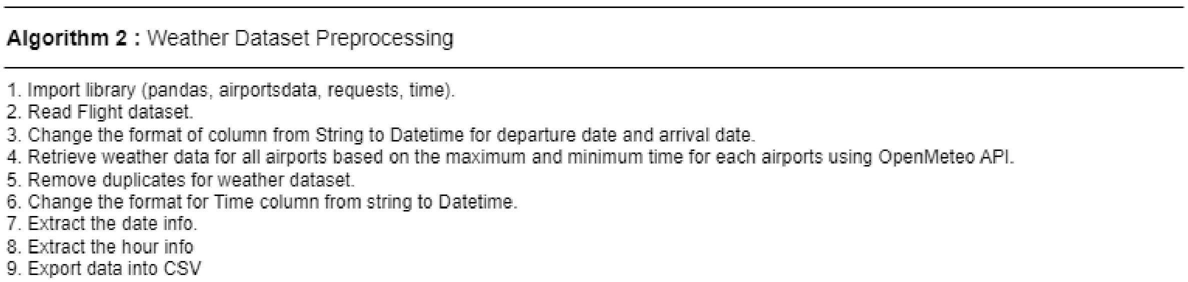
Algorithm 1 : Flight Dataset Preprocessing

```
1. Import Library (pandas, calendar, airports, flightradarapi, geopy, requests)
2. Read flight dataset.
3. Drop Columns.
4. Drop NA.
5. Round Time (10:45:00 -> 11:00:00) .
6. Extract Hour (11:00:00 -> 11).
7. Extract Day (6/3/2023 -> Monday).
8. Extract Weekend info (Weekday/ Weekend).
9. Extract parts of the day info (Morning, Afternoon, Evening, Night).
10. Apply step 5 - 9 to local arrival/departure time and utc arrival/departure time.
11. Fill Na with 0 for non delay flight.
12. Categorize delay column with True/False (True = Has Delay)
13. Append airport info using API (Airport name, Airport City, Latitude and Longitude)
14. Apply step 13 for Arrival airport and Departure airport.
15. Mapping airline_icao with Airline name.
16. Mapping flight model into flight detail.
17. Extract Internation/ Domestic info into dataset
18. Export data into CSV.
```

Figure 1. Flight Dataset Pre-Processing Algorithm

E.Weather Dataset Data Pre-Processing

First and foremost, the dataset is downloaded using OpenMeteo API. The second step performed on the dataset is data formatting. The datetime column data is converted from string to DateTime format for easier processing. On the other hand, the Date data were extracted from the time column. Lastly, the time info was extracted from the datetime column as well. The extraction of the Date and Time data is to set up the dataset for merging with the main dataset later on in the steps.



Algorithm 2 : Weather Dataset Preprocessing

```
1. Import library (pandas, airportsdata, requests, time).
2. Read Flight dataset.
3. Change the format of column from String to Datetime for departure date and arrival date.
4. Retrieve weather data for all airports based on the maximum and minimum time for each airports using OpenMeteo API.
5. Remove duplicates for weather dataset.
6. Change the format for Time column from string to Datetime.
7. Extract the date info.
8. Extract the hour info
9. Export data into CSV
```

Figure 2. Weather Dataset Pre-Processing Algorithm

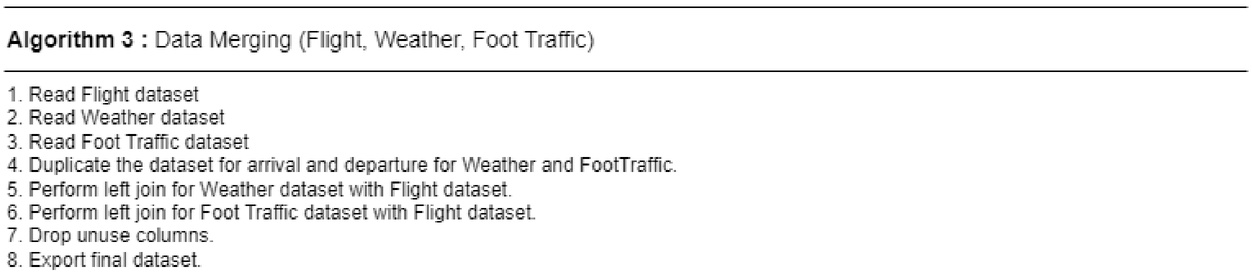
F.Data Merging

The last part of the pre-processing for the whole project will be the merging of the Flight dataset, Weather dataset and Foot Traffic dataset. The first datasets to undergo the merging will be the weather datasets. As the weather datasets involved the arrival airport and departure airport together, the datasets will have to be loaded twice for merging of departure airport's weather data and the arrival's weather data. The cells will be renamed to *dep_XXXX* for the departure weather data and *arr_XXXX* for the arrival weather data. The merging process involved will be left joined. The columns on the flight dataset, *dep_iata, dep_date_utc and dep_hour_utc* will be used as indexes to perform joining *with departure_airport, departure_date, and departure_hour* from weather datasets. This process will be applied to the arrival's weather data as well, *arr_iata, arr_date_utc and arr_hour_utc* will be used as indexes to perform joining with *arrival_iata, arrival_date_utc, and arrival_hour* from weather datasets.

The joined datasets from previous steps will need to perform a merging process with foot traffic data to complete the steps. The foot traffic datasets will be loaded twice for merging of departure airport foot traffic data and arrival airport foot traffic data. The columns on the main datasets, *dep_iata, dep_Day_local and dep_Hour_local* will be used as indexes to perform a left join with the foot traffic column of *Departure_Airport, Departure_Day, Departure_Hour*. The same process will be applied to arrival foot traffic data, *arr_iata, arr_Day_local and arr_Hour_local* will use as indexes to perform a left join with *Arrival_Airport, Arrival_Day and Arrival_Hour*.

After finishing performing the left joined on the main dataset, there will be extra columns with the same information that exists in the datasets. This information will need to be removed from the dataset to avoid confusion. These column includes *dep_iata, arr_iata, dep_Hour_utc, dep_Day_local, dep_Hour_local, arr_Hour_utc, arr_Day_local, arr_Hour_local, Departure_Airport_x, Departure_Date, Departure_Hour_x, Arrival_Airport_x, Arrival_Date ,Arrival_Hour_x, Departure_Airport_y, Departure_Day, Departure_Hour_y, Arrival_Airport_y, Arrival_Day, and Arrival_Hour_y*.

In the end, a final dataset consisting of flight dataset, departure airport weather data, arrival airport weather data, departure airport foot traffic data and arrival foot traffic data will be compiled from the processing steps. The datasets are now prepared for the analysis and modelling steps.



Algorithm 3 : Data Merging (Flight, Weather, Foot Traffic)

```
1. Read Flight dataset
2. Read Weather dataset
3. Read Foot Traffic dataset
4. Duplicate the dataset for arrival and departure for Weather and FootTraffic.
5. Perform left join for Weather dataset with Flight dataset.
6. Perform left join for Foot Traffic dataset with Flight dataset.
7. Drop unuse columns.
8. Export final dataset.
```

Figure 3. Data Merging Pre-Processing Algorithm

IV. PREDICTIVE MODEL CONSTRUCTION

This section aims to introduce the approach used for, feature selection, SMOTE, and model selection in the context of predicting flight disruptions at Kuala Lumpur International Airport.

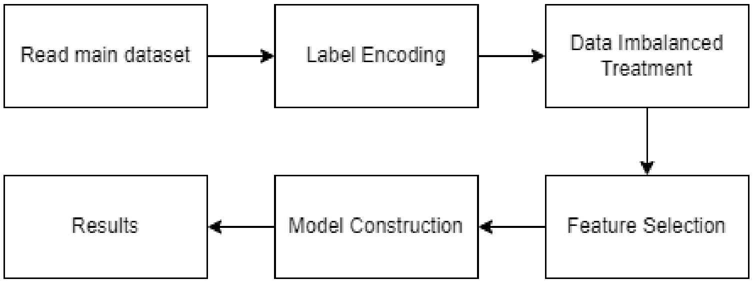


Figure 4. Model Construction Chart

A. Feature Selection

In the context of predicting flight disruptions at Kuala Lumpur International Airport, the feature selection process involved identifying the most relevant features that can impact the performance of the model. To accomplish this, Recursive Feature Elimination (RFE) was utilized, which is a backward elimination method that recursively eliminates features with the least impact on the model until the desired number of features is obtained.

First, The data is split into predictor variables and the target variable, with the predictor variables contained in X and the target variable in y. Random Forest Classifier model is defined with specific hyperparameters, such as 'n_estimators', 'max_depth', and 'class_weight', and fits it on the predictor variables and target variable. Next, Random Forest Classifier model was passed into the RFE function with a parameter of 'min_features_to_select=1' and 'cv=2' to determine the optimal number of features for our model.

Once the RFE algorithm was completed, the features were ranked based on their importance scores using the 'ranking' function. A Pandas DataFrame was created to store the features' scores and sorted them in descending order to identify the top 10 most relevant features for our model. The top 10 features included *'dep_Time_utc*', '*Airline*', *'dep_PartofHour_local*', *'arr_Time_utc*', '*duration*', *'Departure_Temperature*', *'Departure_Busyness*', *'Aircraft_Model*', *'Departure_Humidity*', and '*dep_Date_utc*.' In contrast, the bottom 10 features, which had the least impact on the model, included *'Arrival_WindDirection*', *'Aircraft_isWidebody*', *'Arrival_WeatherCode*', '*dep_Weekend_local*', '*Aircraft_isTurboProp*', *'arr_Weekend_local*', '*Arrival_Rain*', '*Departure_Snowfall*', '*Arrival_Precipitation*', and '*Arrival_Snowfall*.'

By using RFE for the feature selection, the number of features in our model can significantly reduce, making it more efficient and less prone to overfitting. Additionally, the selected features are deemed the most relevant for predicting flight disruptions at Kuala Lumpur International Airport based on their importance scores.

Table 3. Feature Selection Scores

Top 10 Feature		**Bottom 10 Feature**	
Features	**Score**	**Features**	**Score**
dep_Time_utc	1	*Arrival_Cloudcover*	0.4
Airline	0.98	*Arrival_Humidity*	0.38
dep_PartofHour_local	0.96	*dep_City*	0.36
arr_Time_utc	0.94	*Departure_Rain*	0.34
duration	0.91	*Arrival_WindSpeed*	0.32
Departure_Temperature	0.89	*Arrival_Intensity Status*	0.3
Departure_Busyness	0.87	*Departure_Intensity Status*	0.28
Aircraft_Model	0.85	*Departure_Intensity Level*	0.26
Departure_Humidity	0.83	*Departure_Precipitation*	0.23
dep_Date_utc	0.81	*Departure_WeatherCode*	0.21

B. Data Imbalance Treatement

To address the class imbalance issue in the dataset, SMOTE is used, a popular oversampling method. First, the original dataset was checked to see how imbalanced it was. The count graph of the classes showed that the flight disruption class was the minority class, with only 10,336 samples, while the non-disruption class had 55,673 samples. This severe imbalance could negatively affect the performance of the predictive model, as the algorithm could simply predict the majority class for every instance and still achieve high accuracy. Therefore, SMOTE was used to balance the class distribution by generating synthetic samples of the minority class to match the number of samples in the majority class.

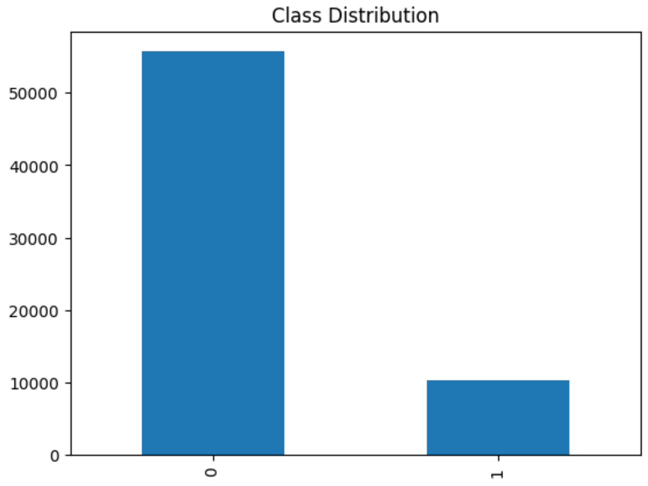


Figure 5. Class Distribution Before SMOTE

The SMOTE model was constructed using the SMOTE function from the imblearn library and then applied to the dataset. First, the dataset was split into training and testing sets, with a test size of 30% and a random state of 10. Then, the SMOTE model was fit with the training data only, and the oversampled data was obtained by calling the fit_resample function on the training set. The resulting oversampled dataset had a balanced class distribution, with 39,054 samples in each class.

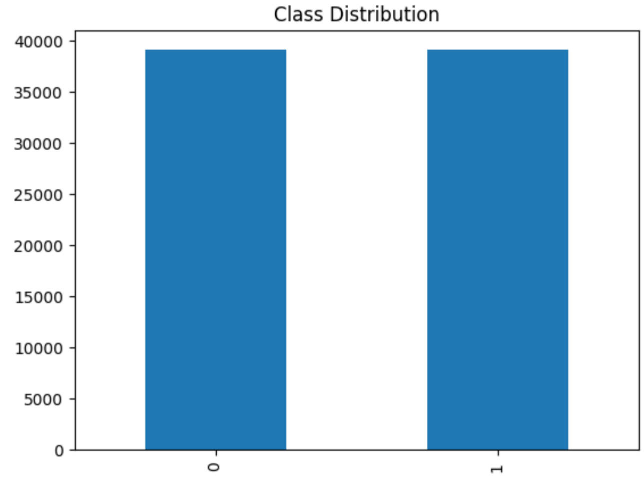


Figure 5. Class Distribution After SMOTE

To check if the SMOTE treatment had a significant effect on the class distribution, an analysis of variance (ANOVA) test was performed. Specifically, a one-way ANOVA was conducted on the original and SMOTE datasets, using the delayStatus column as the categorical variable. The F-statistic was found to be 21,527.69, with a p-value of 0. This indicates that the means of the delayStatus column in the two datasets are significantly different and that the SMOTE treatment was effective in balancing the class distribution.

In summary, SMOTE is used to balance the class distribution of the imbalanced dataset. The SMOTE model was fit with the training data only, and the resulting oversampled dataset had a balanced class distribution. The effectiveness of the SMOTE treatment was confirmed by a one-way ANOVA test, which showed a significant difference between the means of the delayStatus column in the original and SMOTE datasets. By balancing the class distribution, I improved the quality of the dataset and increased the chances of building an accurate predictive model.

C. Model

In the Predictive Model Construction section, the performance of six classification models: Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression, were compared using two datasets. These models were chosen based on their ability to handle imbalanced datasets, as well as their widespread use in classification tasks.

To start, the dataset is split into training and testing sets using the train_test_split method from the scikit-learn library. 70% of the data is assigned for training and the remaining 30% is for testing. The stratify parameter to y is set to ensure that both training and testing sets had a balanced distribution of the target variable. The six classification models proceeded to train on the pre-processed dataset with the selected features, which were determined using Recursive Feature Elimination (RFE) method. The first dataset used was the original dataset with imbalanced classes, while the second dataset was the resampled dataset using the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance issue.

For the Naive Bayes model, the GaussianNB method from the scikit-learn library was used to fit the training data and predicted the target variable for the testing set. Similarly, the SVM model is trained using the svm.SVC method and predicted the target variable for the testing set. The Decision Tree, Random Forest, KNN, and Logistic Regression models are trained using the DecisionTreeClassifier, RandomForestClassifier, K-Nearest Neighbours, and LogisticRegression methods, respectively. For each model, the training data is fitted and predicted as the target variable for the testing set. The performance of each model is evaluated using four metrics: accuracy, precision, recall, and F1-score.

After evaluating the six classification models on the first dataset, the same process is repeated on the second dataset, which was the resampled dataset using SMOTE to address the class imbalance issue. The performance of each model is compared on both datasets. In summary, six classification models were used to train two datasets and compared their performance using four metrics: accuracy, precision, recall, and F1-score. By using RFE for feature selection and SMOTE for resampling the dataset, the classification models' performance is aimed to improve the imbalanced dataset.

V. FINDINGS

In this section, the results of the comparative analysis of six classification models—Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression—are presented. The performance of each model was evaluated on two datasets: the original dataset and a dataset created using SMOTE to address the issue of data imbalance. The evaluation metrics used include accuracy, precision, recall, and F1-score.

When considering the top 10 features of the original dataset, the Random Forest model achieved the highest accuracy of 88.35%. The F1-score of the Random Forest model was 0.877, suggesting a good balance between precision and recall. Notably, the K-Nearest Neighbour model also demonstrated competitive performance with an accuracy of 85.52% and an F1-score of 0.845.

Table 4. 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 Neighbours	85.52	0.842	0.855	0.845
Logistic Regression	82.86	0.777	0.829	0.771

Expanding the feature set to the top 30 features of the original dataset, the Random Forest model outperformed the other models, achieving an accuracy of 87.94% and an F1-score of 0.870. To add on, the Decision Tree model also demonstrated competitive performance with an accuracy of 83.5% and an F1-score of 0.836.

Table 5. Result using Top 30 Features and Normal Data

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-score**
Naive Bayes	81.69	0.826	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 Neighbours	84.42	0.826	0.844	0.831
Logistic Regression	83.13	0.789	0.831	0.782

Considering the SMOTE dataset with the top 10 features, the Random Forest model achieved the highest accuracy of 91.07% and an impressive F1-score of 0.911. This indicates that the model successfully captured the imbalanced nature of the dataset and improved its predictive performance. The Decision Tree model also demonstrated excellent performance with an accuracy of 85.18% and an F1-score of 0.852.

Table 6. 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 Neighbours	83.25	0.834	0.832	0.832
Logistic Regression	67.32	0.677	0.673	0.672

Expanding the feature set to the top 30 features of the SMOTE dataset, the Random Forest model continued to exhibit exceptional performance, achieving an accuracy of 93.1% and an F1-score of 0.931. The Decision Tree model showed substantial improvement with an accuracy of 86.96% and an F1-score of 0.870 indicating that the additional features provided more discriminatory power for the model.

Table 7. 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 Neighbours	83.84	0.856	0.838	0.836
Logistic Regression	67.35	0.677	0.673	0.672

These results highlight the importance of feature selection and addressing data imbalance in achieving accurate predictions. The Random Forest model consistently outperformed other models across different feature sets and datasets, demonstrating its robustness and ability to handle imbalanced datasets effectively.

In the discussion, several points were addressed that shed light on the results. Firstly, the analysis emphasized that not all features contribute equally to the prediction of flight disruptions, highlighting the importance of feature selection. The results showed that the Random Forest model performed well when trained on the top features, both in terms of accuracy and F1 score.

Furthermore, the discussion highlighted the impact of using the SMOTE technique to address the class imbalance. The models trained on the SMOTE dataset generally exhibited improved performance compared to those trained on the normal dataset. The Random Forest model consistently demonstrated high accuracy and F1 scores in the SMOTE dataset scenarios, indicating the effectiveness of the technique in capturing the imbalanced nature of the data and improving predictions.

Additionally, the discussion emphasized the practical implications of the developed models. Airlines and airport authorities can utilize the Random Forest model trained on the SMOTE dataset with the top features to anticipate and mitigate flight disruptions. This model can aid in decision-making processes such as resource allocation, scheduling adjustments, and proactive management of potential disruptions.

However, it is essential to acknowledge the limitations of the study. The predictive models are based on historical data and assume that future flight disruptions will follow similar patterns. External factors such as weather conditions, air traffic control issues, or unforeseen events are not accounted for in the models. Furthermore, the models can benefit from further refinement by incorporating additional relevant features or exploring more advanced machine-learning techniques.

In conclusion, this study successfully developed predictive models for flight disruption prediction using different combinations of features and datasets. The results demonstrate the potential of machine learning models, particularly Random Forest, Decision Tree, and Support Vector Machines, in predicting flight disruptions. The use of the SMOTE technique to address class imbalance has shown improvements in model performance. However, further refinement and consideration of external factors are necessary to enhance the accuracy and applicability of these models in real-world scenarios.

VI. CONCLUSION

In conclusion, the main objectives of this study were focused on understanding and predicting flight disruption at KLIA. Firstly, this project aimed to understand and predict flight disruptions at Kuala Lumpur International Airport by achieving three main objectives. Firstly, the project aimed to identify the variables that contribute to flight disruptions at Kuala Lumpur International Airport. Through the analysis of a comprehensive dataset provided by an airline company, including flight schedules, weather conditions, and passenger demand, the project successfully uncovered the specific factors that have a significant impact on flight disruptions. This knowledge will assist airlines in focusing their resources and strategies on addressing the most influential variables to reduce disruptions.

Apart from that, the goal was to construct a predictive model capable of accurately forecasting flight disruptions. By leveraging machine learning algorithms and utilizing the identified variables and their relationships, a reliable model was developed to predict the likelihood of flight disruptions. This proactive tool can assist airlines and airport authorities in taking preventive measures, allocating resources effectively, and mitigating the impact of disruptions on passengers and operations.

In conclusion, this study contributes to the field of flight disruption prediction at KLIA by providing valuable insights and a reliable predictive model. It is important to acknowledge the limitations of the study, such as the reliance on historical data and the potential influence of external factors. Future research should focus on incorporating real-time data integration and exploring advanced machine-learning techniques to further improve the accuracy and robustness of the predictive model.

Overall, this study serves as a stepping stone towards proactive management of flight disruptions at KLIA. By understanding the underlying causes, and relationships, and employing predictive models, stakeholders in the aviation industry can better allocate resources, make informed decisions, and minimize the impact of disruptions on both operations and passengers. It is hoped that this research will inspire further collaboration and innovation in the field, leading to even more effective strategies for managing flight disruptions and enhancing the overall efficiency of airport operations.

REFERENCES

[1] Almaameri, I. M., & Mohammed, A. (2022, May 1). *Predicting Airplane Flight Delays Using Neural Networks*. IEEE Xplore. https://doi.org/10.1109/IICETA54559.2022.9888363

[2] Amulya Arun Ballakur, & Arya, A. (2020). *Empirical Evaluation of Gated Recurrent Neural Network Architectures in Aviation Delay Prediction*. https://doi.org/10.1109/icccs49678.2020.9276855

[3] Anees, A., & Huang, W. (2021, September 1). *Flight Delay Prediction: Data Analysis and Model Development*. IEEE Xplore. https://doi.org/10.23919/ICAC50006.2021.9594260

[4] Balamurugan, R., Maria, A. V., Baranidaran, G., MaryGladence, L., & Revathy, S. (2022, April 1). *Error Calculation for Prediction of Flight Delays using Machine Learning Classifiers*. IEEE Xplore. https://doi.org/10.1109/ICOEI53556.2022.9776709

[5] Beth, T. (2022, February 18). Storm Eunice: Flights and train services cancelled. *BBC*. https://www.bbc.com/news/business-60430197

[6] Borsky, S., & Unterberger, C. (2019). Bad weather and flight delays: The impact of sudden and slow onset weather events. *Economics of Transportation*, *18*, 10–26. https://doi.org/10.1016/j.ecotra.2019.02.002

[7] 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. https://doi.org/10.1109/tits.2021.3103502

[8] Chakrabarty, N. (2019, March 1). *A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines*. IEEE Xplore. https://doi.org/10.1109/IEMECONX.2019.8876970

[9] Dietz, S. J., Kneringer, P., Mayr, G. J., & Zeileis, A. (2018). Correction to: Forecasting Low-Visibility Procedure States with Tree-Based Statistical Methods. *Pure and Applied Geophysics*, *176*(6), 2645–2658. https://doi.org/10.1007/s00024-018-1993-8

[10] Gu, Y., & Yang, J. (2019, March 1). *Research on Cause and Governance Path of Passenger Disturbance in Flight Delay in Terminal*. Www.atlantis-Press.com; Atlantis Press. https://doi.org/10.2991/iafsm-18.2019.18

[11] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology*, *69*(1), 140–150. https://doi.org/10.1109/tvt.2019.2954094

[12] Hassan, L. K., Santos, B. F., & Vink, J. (2020). Airline Disruption Management: A Literature Review and Practical Challenges. *Computers & Operations Research*, 105137. https://doi.org/10.1016/j.cor.2020.105137

[13] Hopane, J., & Gatsheni, B. (2019, December 1). *A Computational Intelligence-Based Prediction Model for Flight Departure Delays*. IEEE Xplore. https://doi.org/10.1109/CSCI49370.2019.00107

[14] Hu, P., Zhang, J.-P., & Li, N. (2021). *Research on Flight Delay Prediction Based on Random Forest*. https://doi.org/10.1109/iccasit53235.2021.9633476

[15] Huo, J., Keung, K. L., Lee, C. K. M., Ng, K. K. H., & Li, K. C. (2020). The Prediction of Flight Delay: Big Data-driven Machine Learning Approach. *2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. https://doi.org/10.1109/ieem45057.2020.9309919

[16] Jiang, Y., Liu, Y., Liu, D., & Song, H. (2020). 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)*. https://doi.org/10.1109/dasc-picom-cbdcom-cyberscitech49142.2020.00114

[17] Jiang, Y., Miao, J., Zhang, X., & Le, N. (2020, October 1). *A multi-index prediction method for flight delay based on long short-term memory network model*. IEEE Xplore. https://doi.org/10.1109/ICCASIT50869.2020.9368554

[18] Kalyani, N. L., Jeshmitha, G., Sai U., B. S., Samanvitha, M., Mahesh, J., & Kiranmayee, B. V. (2020, October 1). *Machine Learning Model - based Prediction of Flight Delay*. IEEE Xplore. https://doi.org/10.1109/I-SMAC49090.2020.9243339

[19] Keoni, E. (2023, May 8). *Video shows crosswinds force Taiwan Starlux plane to abort Narita landing | Taiwan News | 2023-05-08 10:48:00*. Taiwan News. https://www.taiwannews.com.tw/en/news/4885071

[20] Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, *0*(0). https://doi.org/10.24200/sci.2017.20020

[21] Kulkarni, R., Jenamani, R. K., Pithani, P., Konwar, M., Nigam, N., & Ghude, S. D. (2019). Loss to Aviation Economy Due to Winter Fog in New Delhi during the Winter of 2011–2016. *Atmosphere*, *10*(4), 198. https://doi.org/10.3390/atmos10040198

[22] Li, Q., & Jing, R. (2021). *Generation and prediction of flight delays in air transport*. *15*(6), 740–753. https://doi.org/10.1049/itr2.12057

[23] Liu, F., Sun, J., Liu, M., Yang, J., & Gui, G. (2020). Generalized Flight Delay Prediction Method Using Gradient Boosting Decision Tree. *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*. https://doi.org/10.1109/vtc2020-spring48590.2020.9129110

[24] Mang, C., & Chen, Y. (2020). Research on Flight delay Prediction based on Multi-Model Fusion. *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC)*. https://doi.org/10.1109/itoec49072.2020.9141816

[25] Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020, February 1). *Predicting Flight Delays with Error Calculation using Machine Learned Classifiers*. IEEE Xplore. https://doi.org/10.1109/SPIN48934.2020.9071159

[26] Pamplona, D. A., Weigang, L., de Barros, A. G., Shiguemori, E. H., & Alves, C. J. P. (2018, July 1). *Supervised Neural Network with multilevel input layers for predicting of air traffic delays*. IEEE Xplore. https://doi.org/10.1109/IJCNN.2018.8489511

[27] Schuldt, S. J., Nicholson, M. R., Adams, Y. A., & Delorit, J. D. (2021). Weather-Related Construction Delays in a Changing Climate: A Systematic State-of-the-Art Review. *Sustainability*, *13*(5), 2861. https://doi.org/10.3390/su13052861

[28] Shu, Z. (2021, December 1). *Analysis of Flight Delay and Cancellation Prediction Based on Machine Learning Models*. IEEE Xplore. https://doi.org/10.1109/MLBDBI54094.2021.00056

[29] Su, Y., Xie, K., Wang, H., Liang, Z., Art Chaovalitwongse, W., & Pardalos, P. M. (2021). Airline Disruption Management: A Review of Models and Solution Methods. *Engineering*, *7*(4). https://doi.org/10.1016/j.eng.2020.08.021

[30] Toolkit, W. E. (2021, January 22). *Types and Categories of Flight Disruption: A Guide*. Otc-Cta.gc.ca. https://otc-cta.gc.ca/eng/publication/types-and-categories-flight-disruption-a-guide

[31] Wang, H. (2022, January 1). *Big Data Visualization and Analysis of Various Factors Contributing to Airline Delay in the United States*. IEEE Xplore. https://doi.org/10.1109/BDICN55575.2022.00042

[32] Wang, J., & Pan, W. (2022, July 1). *Flight delay prediction based on ARIMA*. IEEE Xplore. https://doi.org/10.1109/ICCEAI55464.2022.00047

[33] Wang, K., Li, J., & Tian, Y. (2019, October 1). *Airport Delay Prediction Method based on Improved Weather Impacted Traffic Index*. IEEE Xplore. https://doi.org/10.1109/ICCASIT48058.2019.8973213

[34] Wang, T., Lin, L., & Gao, J. (2021, September 1). *Explainable Multi-task Flight Arrival Delay Prediction*. IEEE Xplore. https://doi.org/10.1109/ITSC48978.2021.9564930

[35] 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*(05), 54–71. https://doi.org/10.4236/jss.2020.85005

[36] Wu, W., Cai, K., Yan, Y., & Li, Y. (2019, September 1). *An Improved SVM Model for Flight Delay Prediction*. IEEE Xplore. https://doi.org/10.1109/DASC43569.2019.9081611

[37] Xu, M., Wang, M., Wang, Y., & Delahaye, D. (2022, October 1). *Robust estimation of airport declared capacity*. IEEE Xplore. https://doi.org/10.1109/ITSC55140.2022.9922312

[38] Yanying, Y., Mo, H., & Haifeng, L. (2019). A Classification Prediction Analysis of Flight Cancellation Based on Spark. *Procedia Computer Science*, *162*, 480–486. https://doi.org/10.1016/j.procs.2019.12.014

[39] Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00380-z

[40] Yiu, C. Y., Ng, K. K. H., Kwok, K. C., Tung Lee, W., & Mo, H. T. (2021, October 1). *Flight delay predictions and the study of its causal factors using machine learning algorithms*. IEEE Xplore. https://doi.org/10.1109/ICCASIT53235.2021.9633571

[41] Zhou, F., Jiang, G., Lu, Z., & Wang, Q. (2022). Evaluation and Analysis of the Impact of Airport Delays. *Scientific Programming*, *2022*, e7102267. https://doi.org/10.1155/2022/7102267

[42] Zhou, H., Li, W., Jiang, Z., Cai, F., & Xue, Y. (2022). Flight Departure Time Prediction Based on Deep Learning. *Aerospace*, *9*(7), 394. https://doi.org/10.3390/aerospace9070394