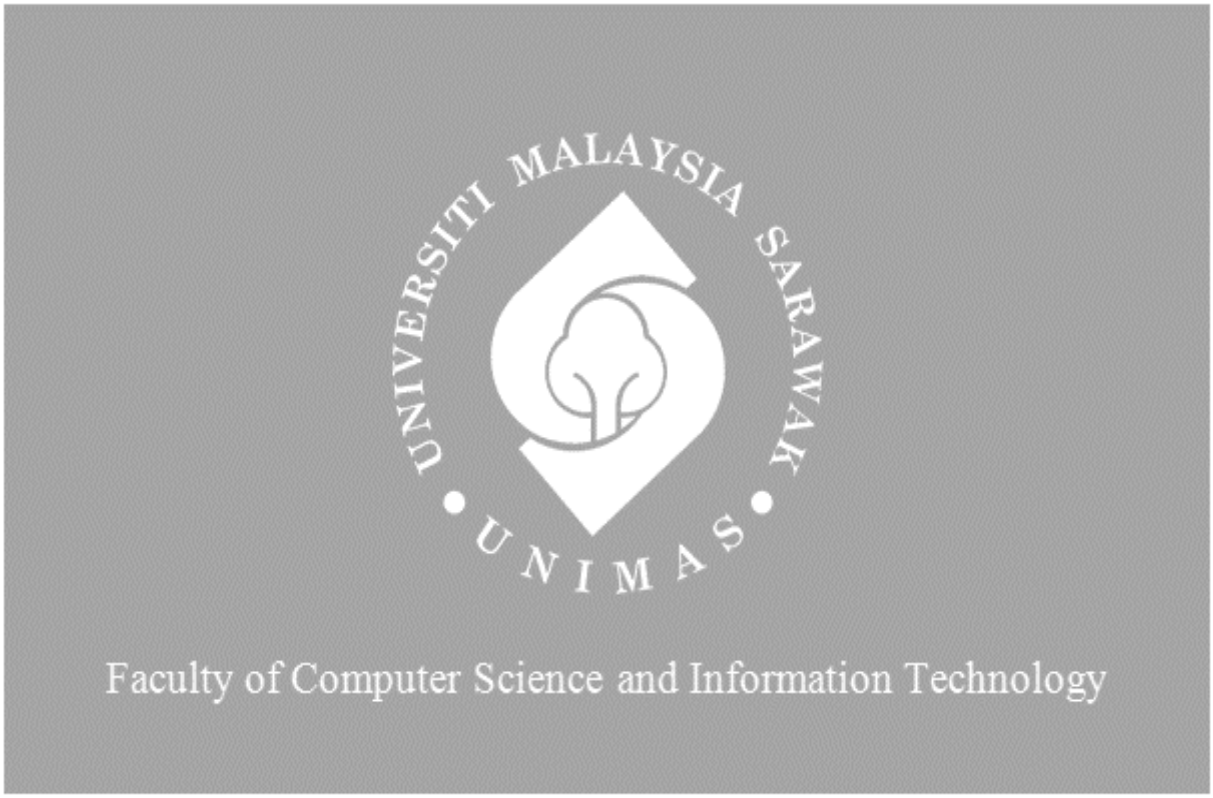

FLIGHT STATUS PREDICTION

Mohamad Aizad bin Radi

Bachelor of Computer Science with Honours (Information System)

2023

UNIVERSITI MALAYSIA SARAWAK

THESIS STATUS ENDORSEMENT FORM

TITLE FLIGHT STATUS PREDICTION

ACADEMIC SESSION: 2022/2023

MOHAMAD AIZAD BIN RADI
(CAPITAL LETTERS)

hereby agree that this Thesis* shall be kept at the Centre for Academic Information Services, Universiti Malaysia Sarawak, subject to the following terms and conditions:

1. The Thesis is solely owned by Universiti Malaysia Sarawak
2. The Centre for Academic Information Services is given full rights to produce copies for educational purposes only
3. The Centre for Academic Information Services is given full rights to do digitization in order to develop local content database
4. The Centre for Academic Information Services is given full rights to produce copies of this Thesis as part of its exchange item program between Higher Learning Institutions [or for the purpose of interlibrary loan between HLI]
5. ** Please tick (√)

	CONFIDENTIAL	(Contains classified information bounded by the OFFICIAL SECRETS ACT 1972)
	RESTRICTED	(Contains restricted information as dictated by the body or organization where the research was conducted)
√	UNRESTRICTED	

Validated by

aizad
(AUTHOR'S SIGNATURE)

(SUPERVISOR'S SIGNATURE)

Permanent Address

NO 348 LORONG 4
KAMPUNG TABUAN LOT
93450, KUCHING SARAWAK

Date: 27/6/2023

Date: 27/6/2023

Note * Thesis refers to PhD, Master, and Bachelor Degree
** For Confidential or Restricted materials, please attach relevant documents from relevant organizations / authorities

Declaration

I hereby declare that this report is an outcome of my own effort under the guidance of Dr Stephanie Chua Hui Li. This project is submitted to the University of Malaysia Sarawak for the fulfilment of the Bachelor of Computer Science with Honours (Information System).

(MOHAMAD AIZAD BIN RADI)

1/7/2023

(Date Submitted)

Acknowledgement

First and foremost, I would like to express my appreciation and gratitude to those who have supported me in completing this project. I extend special appreciation to my supervisor, Dr Stephanie Chua Hui Li, for her continuous support and motivation throughout the project. Without her guidance, it would have been impossible to achieve the desired outcomes.

I would also like to express my gratitude to my family and friends who have provided unwavering financial and emotional support. Their assistance has been invaluable in my decision-making process throughout this project.

Abstract

Air travel is one of the most widely used forms of transportation around the world, including in Malaysia. It originated in 1903 with the creation and first flight of the Wright Flyer by the Wright brothers. According to statistics, the number of flights worldwide is expected to reach up to 32.4 million. Although air travel is generally more expensive compared to other modes of transportation, it remains widely used due to its speed in reaching destinations. However, a common occurrence in the airline industry is flight delays or cancellations. Factors that can lead to these disruptions include staff shortages, adverse weather conditions, and technical problems with the aircraft. Such situations can leave passengers frustrated and disappointed, particularly when their travel plans are unexpectedly affected. Therefore, this project aims to predict flight statuses using a Machine Learning approach. Comparative studies will be conducted to evaluate and compare similar projects undertaken by others.

Abstrak

Penerbangan udara merupakan salah satu bentuk pengangkutan yang paling banyak digunakan di seluruh dunia, termasuk di Malaysia. Ia bermula pada tahun 1903 dengan penciptaan dan penerbangan pertama Wright Flyer oleh Wright bersaudara. Menurut statistik, jumlah penerbangan di seluruh dunia dijangka mencapai hingga 32.4 juta. Walaupun perjalanan udara biasanya lebih mahal berbanding dengan pengangkutan lain, ia tetap menjadi pilihan yang popular kerana kelajuan dalam mencapai destinasi. Namun, kejadian yang sering berlaku dalam industri penerbangan adalah kelewatan atau pembatalan penerbangan. Faktor-faktor yang menyebabkan gangguan ini termasuk kekurangan kakitangan, keadaan cuaca yang buruk, dan masalah teknikal pada pesawat. Keadaan seperti ini boleh membuat penumpang merasa kecewa, terutamanya apabila rancangan perjalanan mereka terjejas secara tiba-tiba. Oleh itu, projek ini bertujuan untuk meramalkan status penerbangan dengan menggunakan pendekatan *Machine Learning*. Kajian perbandingan akan dijalankan untuk menilai dan membandingkan projek-projek serupa yang telah dilakukan oleh pihak lain.

Table of Contents

List of Tables

Lists of Figures

Chapter 1 Introduction

1.1 Introduction

The first aeroplane was created in 1903, called the Wright Flyer, by the Wright Brothers (Hsu, 2019). Since then, many aeroplanes have been created, with Deutsche Luftschiffahrts-Aktiengesellschaft (DELAG) becoming the first world airline in 1909 (Oceansky Journal, 2022). The first airline in America is the St. Petersburg–Tampa Airboat Line in 1914 (Hardiman, 2022), while in Malaysia, the first airline is Malayan Airways Limited (now known as Malaysia Airlines) with its first flight in 1947 (Curran, 2021). Even though the price to use an aeroplane is expensive, air travel is still one of the most common modes of transportation due to faster travel compared to other modes (Bernal, 2018). In 2023, the total number of domestic and international airline flights is expected to be 32.4 million (Statista Research Department, 2023).

1.2 Problem Statement

Flight delays and cancellations are common scenarios in air travel. Various reasons may lead to a flight being delayed or cancelled. Some of the reasons include weather conditions, security issues, and staff shortages (Refundor, 2022). Flight delays and cancellations usually occur at the last minute before a flight is scheduled to depart. It can be highly inconvenient for travellers to receive information about the flight status at such a late stage. Therefore, finding a solution that enables travellers to predict their flight status before their intended departure is desirable, helping them make informed decisions when booking flights and reducing the likelihood of delays or cancellations. This would also allow them to conveniently plan their journey to the airport.

1.3 Aims and Project Objectives

The objectives of this project are:

- To determine the best feature set for flight status classification
- To build classification models for flight status classification using the machine learning approach
- To compare the performance of the classification models for flight status
- To determine the best classification model for flight status prediction

1.4 Brief Methodology

In this project, Knowledge Discovery in Databases (KDD) will be used. KDD is finding relevant knowledge from a set of data (Techopedia, 2017). This popular data mining method involves several steps, including data selection and preparation, data cleaning, the incorporation of prior knowledge about the data sets, and the interpretation of precise answers from the observed results. The overall process is in the following steps:

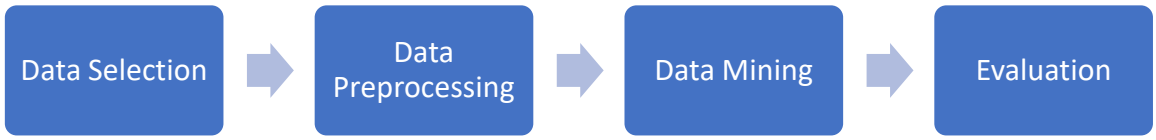

Figure 1.1 Knowledge Discovery in Databases

1. Data Selection

   During the data selection phase, the flight data obtained from Kaggle will be utilized (Mulla, 2022). This dataset encompasses flight data spanning from the year 2018 to 2022. The dataset comprises a total of 61 attributes, providing comprehensive information for analysis. With a significant volume, the dataset contains 27,838,495 entries, enabling a detailed exploration of the data. It is important to note that the dataset encompasses flight information from various airlines operating worldwide, ensuring a diverse and representative sample of flight data.

2. Data Pre-processing

During the data preprocessing phase, data cleaning will be performed to ensure the data is prepared appropriately for analysis. This includes addressing missing values, handling outliers, and resolving any inconsistencies within the dataset. Feature selection techniques will also be employed to reduce the number of attributes and identify the most relevant data for predicting flight status. Data analysis will play a crucial role in this process, enabling the exploration and examination of relationships, patterns, and distributions within the dataset. Through data analysis, the most important and influential attributes for predicting flight status can be identified, facilitating informed decisions for attribute selection in the final predictive model.

3. Data Mining

In data mining, the machine learning algorithms that will be used in this project are:

a. Logistic Regression Algorithm
b. Decision Tree Algorithm
c. Random Forest Algorithm
d. K-Nearest Neighbour Algorithm
e. Naïve Bayes Algorithm

4. Evaluation

In evaluation, evaluation metrics that will be used to evaluate the classification models are:

a. Accuracy
b. Precision
c. Recall
d. F1 Score
e. Area Under the ROC curve (AUC – ROC)

1.5 Scope

The scope of this project is limited to the analysis of flight data specifically from the Top 3 American Airlines, covering the period from 2018 to 2019. The dataset comprises 61 features and includes a total of 27,838,495 rows of data. The decision to exclude data from the year 2020 onwards is due to the extraordinary circumstances and atypical flight statuses resulting from the COVID-19 pandemic. By focusing on pre-pandemic data, the project aims to capture a more representative and reliable picture of flight patterns and statuses.

1.6 Significance of Project

The significance of this project lies in its development of a classification model for predicting flight status. The flight status can be accurately predicted using this model. These forecasts can help airlines and travellers plan ahead and lessen the effects of potential delays. The model also has the potential to boost overall operational effectiveness and enhance the passenger experience.

1.7 Project Schedule

ID	Task Name	Start	Finish
1	**Data Collection**	**17/10/2022**	**2/12/2022**
2	Selecting Flight Dataset	17/10/2022	2/12/2022
3	**Data Preprocessing**	**5/12/2022**	**27/1/2023**
4	Data Cleaning	5/12/2022	6/1/2023
5	Feature Selection	9/1/2023	27/1/2023
6	**Data Mining**	**30/1/2023**	**21/4/2023**
7	Choosing Algorithm	30/1/2023	3/2/2023
8	Training the data	6/2/2023	17/3/2023
9	**Evaluation**	**24/4/2023**	**30/6/2023**
10	Evaluate all model	24/4/2023	2/6/2023
11	Choose most suitable model	12/6/2023	30/6/2023

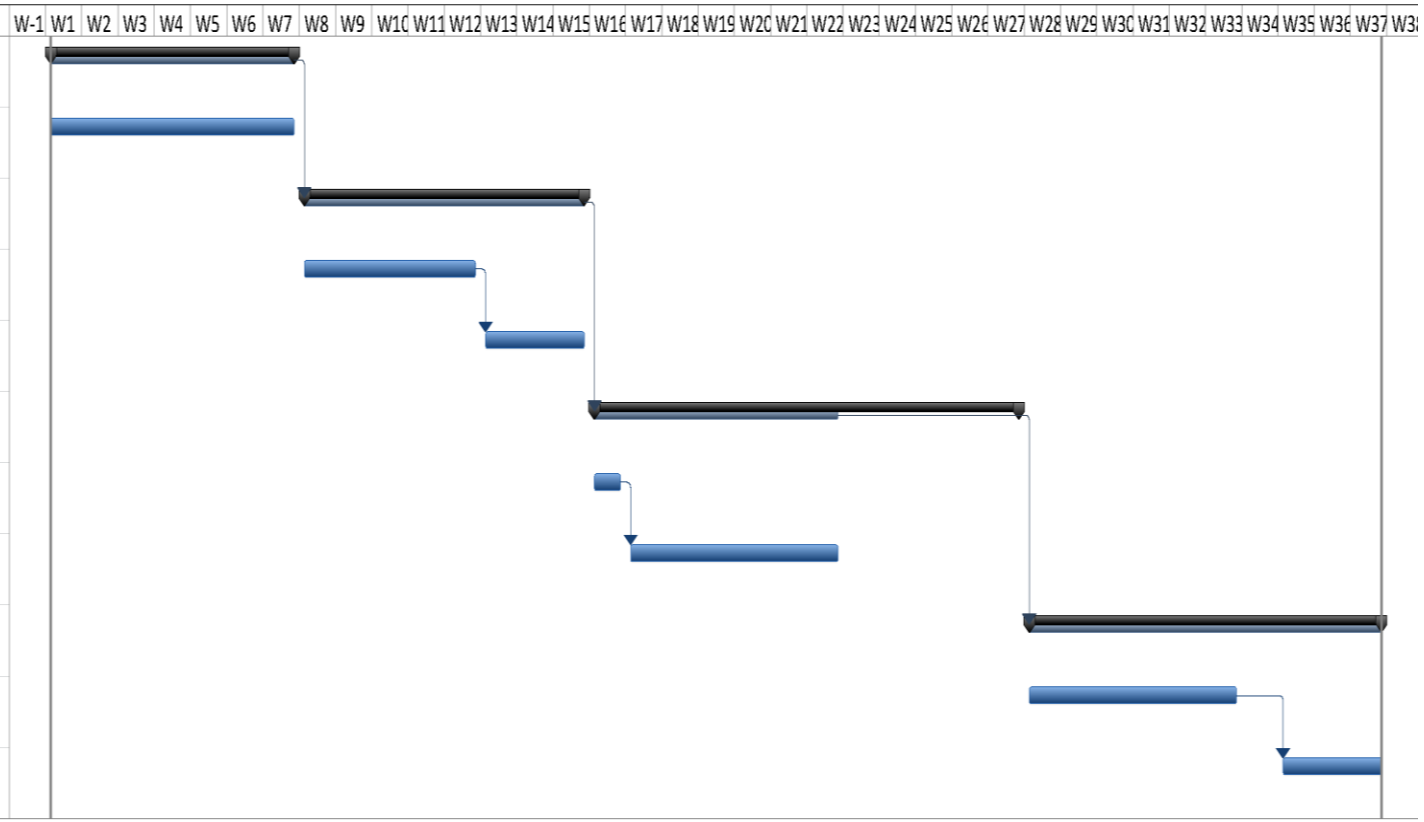

Figure 1.2 Project Schedule

1.8 Project Outcome

The outcome of this project is a comparative study of 5 machine learning models for flight status classification to determine the best classification model.

1.9 Project Outline

Chapter 2 provides an overview of five distinct projects concerning flight prediction, while also delving into the discussion of machine learning algorithms. Chapter 3 outlines the methodology that will be employed in this project, detailing the specific approach to be followed. In Chapter 4, the implementation phase takes place, encompassing all the necessary development steps. Chapter 5 focuses on presenting the experimental results and analyzing the obtained outcomes. Lastly, Chapter 6 serves as the project's conclusion, evaluating the findings and deriving new insights for future endeavours.

Chapter 2 Literature Review

2.1 Introduction

In this chapter, extensive coverage of the machine learning algorithms to be employed will be provided. Additionally, the chapter will explore and discuss five existing systems that bear similarities or close resemblance to the proposed system. Through dedicated research, the strengths and weaknesses of these systems will be identified and analyzed. A comparative analysis will be presented in the form of a table, facilitating a comprehensive evaluation. This literature review holds significant importance as it serves as a valuable reference and guide for conducting the project, ensuring a well-informed and guided approach.

2.2 Flight and Airline

The first aeroplane was created in 1903 called The Wright Flyer by The Wright Brothers (Hsu, 2019). Since then, many aeroplanes have been created with Deutsche Luftschiffahrts-Aktiengesellschaft (DELAG) becoming the first world airline in 1909 (Oceansky Journal, 2022). The first airline in America is St. Petersburg–Tampa Airboat Line in 1914 (Hardiman, 2022) while in Malaysia, the first airline is Malayan Airways Limited (now known as Malaysia Airlines) with its first flight in 1947 (Curran, 2021). In 2023, the total number of domestic and international airline flights is expected to be 32.4 million (Statista Research Department, 2023). However, there is one common scenario in air travel that can make many people infuriated which is a flight being delayed or cancel. This situation usually happens at the very last minute. Thus, this flight status prediction is developed to help travellers to predict their flight status and help them in making their decision as to which flight to book to reduce the chance of delay or cancellation.

2.3 Machine Learning

Machine learning is a subfield of artificial intelligence (AI) and computer science that is centred on using data and algorithms to mimic human learning processes and progressively increase accuracy. To make a prediction or classification, machine learning will be used. Data will be input, and the algorithm will generate a pattern in the data. In this project, there are 6 machine-learning algorithms will be used (IBM Cloud Education, 2020).

2.3.1 Logistic Regression Algorithm

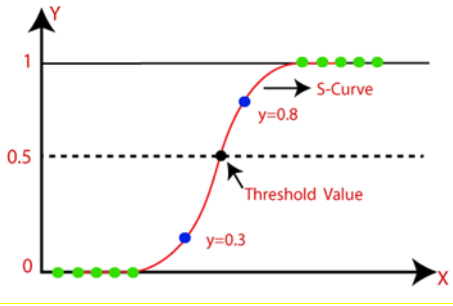

Figure 2.1 Logistic Regression Algorithm

Logistic regression is a machine learning algorithm that is used to predict the categorical dependent variable from a set of independent variables. It will predict the output of a categorical dependent variable. Thus, the outcome must be a discrete or categorical value. It can be 0 or 1, Yes or No but rather than providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. “S” Shaped logistic function is fitted which predicts two maximum values (0 or 1) (Javatpoint, n.d.).

2.3.2 Decision Tree Algorithm

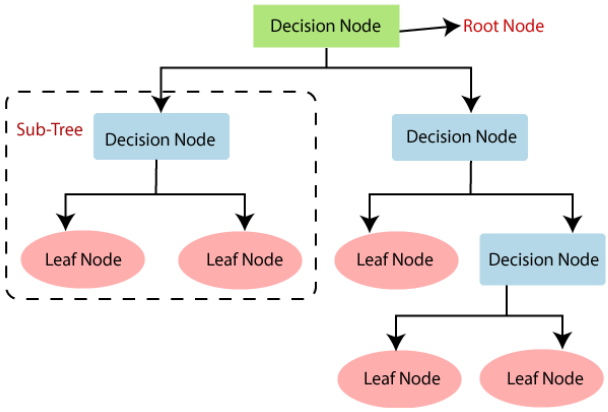

Figure 2.2 Decision Tree Algorithm

Decision Tree Algorithm is an algorithm that can be used to solve classification problems. It is a tree-structured classification, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. There is a decision node that is used to make a decision and has branches and there is also a leaf node that represents the output of the decision and no branches. Based on the characteristics of the available dataset, decisions or tests are run (Javatpoint, n.d.).

2.3.3 Random Forest Algorithm

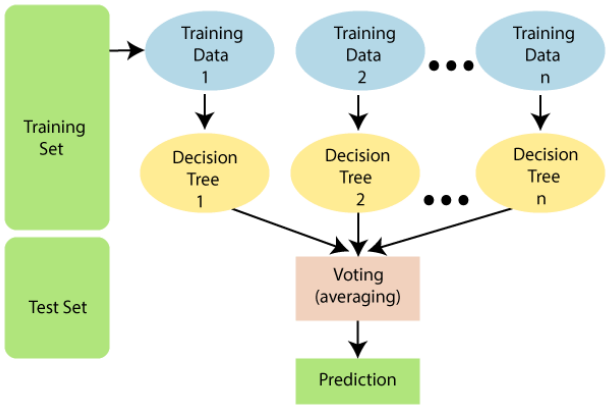

Figure 2.3 Random Forest Algorithm

Random Forest Algorithm is machine learning that can be used for both classification and regression problems. It is based on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance. As implied by the name, random forest lists several decision trees on different subsets of the provided dataset and averages them to increase the dataset's predictive accuracy. Rather than only using one decision tree, the random forest takes predictions from each tree and predicts the final output based on the majority (Javatpoint, n.d.).

2.3.4 K-Nearest Neighbour Algorithm

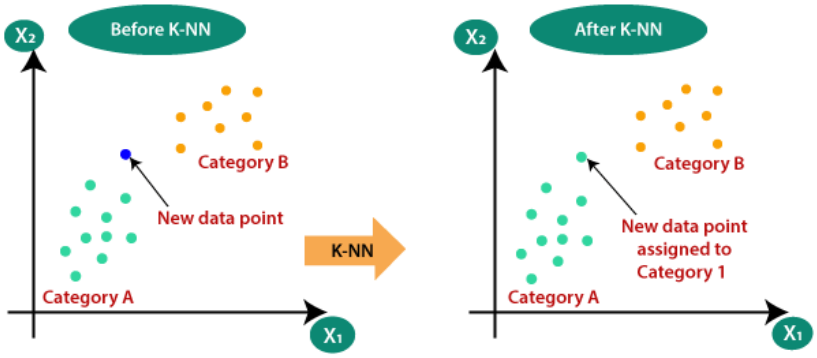

Figure 2.4 K-Nearest Neighbour Algorithm

One of the simplest algorithms is K-Nearest Neighbour. The K-NN algorithm assumes that the new case and the existing cases are similar, and it places the new case in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that using the K-NN algorithm, new data can be quickly and accurately classified into a suitable category. K-NN is also known as a lazy learner algorithm because instead of learning from the training set immediately, it stores the dataset and only performs an action at the time of classification (Javatpoint, n.d.).

2.3.5 Naïve Bayes Algorithm

The Naïve Bayes algorithm is a machine learning algorithm based on the Bayes theorem. It is one of the most effective classification algorithms which helps in the development of quick machine learning models capable of making accurate predictions. Bayes' theorem is used to calculate a hypothesis's probability using existing information. The conditional probability determines this (Javatpoint, n.d.). The formula for Bayes' theorem is below.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 2.5 Bayes' Theorem Formula

2.4 Review of Similar Work

This section will review the previous related works in the last few years. There are five related works chosen. The first project is done by Adrian Alexander Arteche Simmons from Universidad del Pais Vasco, Spain for his Computer Engineering Degree in 2015 (Simmons, 2015). The main objective of this project is to predict flight delays due to weather. The dataset used is flight data in 2013 obtained from the Bureau of Transportation Statistics. The dataset has 22 variables with over 800 thousand entries. During the data preprocessing, data is cleaned, transformed, and merged with weather data. Along with flight data, weather data is also used to get information about weather observation. The algorithms that have been used are linear discriminant analysis classifier with over 74% accuracy and Naïve Bayes Algorithm with 69% accuracy. The confusion matrix is used to test the accuracy.

Next, Flight Delay Prediction is a project done by Bhuvan Bhatia in 2018 as part of his master's degree (Bhatia, 2018). The dataset he uses is the flight data from the years 2007 and 2008, taken from the Bureau of Transportation Statics. According to the report, 2008 on-time performance data contains 7 million records. Other than Flight Data, he uses weather data

which is the Aviation System Performance Metrics (ASPM) data from the FAA Operations & Performance Data. The first step of his project was to do data collection and then data exploration on the flight data. The machine learning algorithm he chooses is Logistic Regression, Random Forest, and Support Vector Machine (SVM). He repeats the same process for the weather data with other algorithms. The conclusion of his project is Random Forest method has the best performance compared to the SVM model. The model correctly predicts 91% of the non-delayed flights. However, for the delayed flights, the model only correctly predicts 41% of the time.

After that, a project was done by Navoneel Chakrabarty from Jalpaiguri Government Engineering College, India (Chakrabarty, 2019). The dataset is obtained from the US Department of Transportation's Bureau of Transportation Statistics (BTS). The dataset contains flight data in the years 2016 and 2017 with 97,360 samples, 12 attributes, and 1 label. Feature selection is done using a correlation matrix and several data preprocessing technique is employed before training the model. There are two strategies used for data preprocessing which skip Data Imbalance Removal in the first strategy while in the second strategy, all technique is used. The algorithm chosen is Gradient Boosting Classifier. For the result, the validation accuracy for strategy 2 is 85.73%.

The next project is done by several authors from the University of Posts and Telecommunications, Nanjing, China in 2019 (Gui, et al., 2020). Automatic dependent surveillance-broadcast (ADS-B) messages are collected, pre-processed, and combined with other data, such as weather conditions, flight schedules, and airport details, to create a dataset for the suggested scheme. Initially, Long Short-term Memory is used but there is an overfitting problem. Then, Random Forest is used to overcome the overfitting problem. The model obtains 90.2% accuracy.

The last project is done by Mingdao Lu, Peng Wei, Mingshu He, and Yingli Teng from the University of Posts and Telecommunications, Beijing, China in 2021 (Lu, Wei, He, & Teng, 2021). The dataset contains flights in China from 2015 to 2017 from the public flight information with 19 columns and 328291 rows. The algorithms used are Gradient Boosted Decision Trees and Extreme Gradient Boosting. After feature selection, GBDT obtain 80.44% accuracy while XGBoost obtains 79.91% accuracy. Then feature extraction is conducted with GBDT obtaining 82.87% accuracy and XGBoost obtaining 82.48% accuracy. To further verify the prediction, GBDT is used to test actual data and it obtains 88.11% accuracy.

2.5 Comparison of Similar Work

Table 2.1 Comparison of Related Work

No.	References	Brief Description	Dataset	Algorithms	Result
1.	Simmons, A. (2015). *Flight Delay Forecast due to Weather.*	This project is done by Adrian Alexander Arteche Simmons from Universidad del Pais Vasco, Spain for his Computer Engineering Degree in 2015. The main objective of this project is to predict flight delays due to weather.	• The year 2013 • 22 Variable • Over 800 thousand entries • Obtain from the Bureau of Transportation Statistics	• Linear discriminant analysis classifier • Naïve Bayes algorithm	In the end, the Linear discriminant analysis classifier obtain 74% accuracy while Naïve Bayes obtain 69% accuracy
2.	Bhatia, B. (2018). *Flight Delay Prediction.*	This project is done by Bhuvan Bhatia in 2018 as part of his master's degree. The objective is to examine the methods for creating models that forecast flight delays brought on by poor weather.	• The year 2007 and 2008 • 7 million records • Taken from Bureau of Transportation Statics	• Logistic Regression • Random Forest • Support Vector Machine (SVM).	In the end, the Random Forest method has the best performance with 91% accuracy on non-delayed flights and 41% accuracy on delayed flights.
3.	Chakrabarty, N. (2019). A Data	This project is done by Navoneel Chakrabarty from Jalpaiguri Government	• The year 2016 and 2017	Gradient Boosting Classifier	There are two strategies used for data