Available online at
https://jcrinn.com/
https://crinn.conferencehunter.com/

Journal of Computing Research and Innovation 9(2) 2024

Journal of Computing Research and Innovation

Machine Learning Approach of Predicting Airline Flight Delay using Naïve Bayes Algorithm

Ahmad Adib Baihaqi Shukri[1], Syarifah Adilah Mohamed Yusoff[2*], Saiful Nizam Warris[3], Mohd Saifulnizam Abu Bakar[4], Rozita Kadar[5]

[1]*College of Computing, Informatics, and Mathematics, UiTM Cawangan Terengganu Branch, Kuala Terengganu Campus, 21080 Kuala Terengganu, Malaysia*
[2,3,4,5]*Department of Computer Sciences and Mathematics, UiTM Pulau Pinang Branch, Permatang Pauh Campus, 13500 Permatang Pauh, Penang, Malaysia*

ARTICLE INFO

Article history:
Received 26 June 2024
Revised 20 July 2024
Accepted 22 July 2024
Online first
Published 1 September 2024

Keywords:
Classification
Flight Delay
Algorithm
Naïve Bayes
Prediction
Machine Learning

DOI:
10.24191/jcrinn.v9i2.460

ABSTRACT

The aviation industry plays a critical role in global transportation, facilitating economic growth and revolutionizing travel. However, flight delays have become a growing concern, impacting both airlines and passengers. This study aims to study the Naïve Bayes algorithm for flight delay prediction. The objective is to develop a reliable flight delay prediction model using the Naïve Bayes algorithm and evaluate its performance. The data set that records flight delay and cancellation data from U.S Department of Transportation's (DOT) was used for the prediction. This study has modified the parameter tuning for Gaussian Naïve Bayes to identify optimum values specifically to construct model for this flight delay dataset. The performance of parameters tuning Gaussian Naïve Bayes model was compared with another two well-known algorithms which are K-Nearest Neighbors (KNN) and Support Vector Machine (SVM)). The KNN and SVM algorithms were also trained and tested to complete the binary classification of flight delays for benchmarking purposes. The evaluation of algorithms was fulfilled by comparing the values of accuracy, specificity and ROC AUC score. The comparative analysis showed that the Gaussian Naïve Bayes has the best performance with an accuracy of 93% and KNN has the worst performance with ROC AUC score 63%.

1. INTRODUCTION

The creation, marketing, and use of airplanes are part of the worldwide aviation industry. It is an essential component of the global economy and helps many other businesses develop. The aviation industry has been

[2*] Corresponding author. *E-mail address*: syarifah.adilah@uitm.edu.my
https://dx.doi.org/10.24191/jcrinn.v9i2.460

a crucial element of global transportation for decades, revolutionizing the way people and goods travel across the world. This industry has grown rapidly over the past several decades as a result of people increasingly choose to travel by air. Subsequently, the number of flights that fail to take off on time increases (Tang, 2021). Instead of that, the sector has also been impacted by a few other factors, including fuel costs, environmental issues, and technological advancements.

When talking about the aviation industry, the first thing that comes to mind is the airport and airplanes. Airports are essential components of the aviation sector, functioning as vital air transport hubs (Freestone, 2009). They offer the infrastructure and facilities required to support the transportation of people, goods, and aircraft. Terminals, runways, taxiways, and control towers are all components of airports that work together to guarantee safe and effective operations. They act as gates, linking various areas and nations, and play an important role in encouraging tourism and generating economic progress. Airplanes are the aviation industry's backbone, acting as the major form of air travel (Buhalis, 2004). They are meticulously designed to guarantee safe and efficient flying. Commercial airliners, cargo planes, luxury jets, and military aircraft are all available in a variety of sizes and kinds to suit a variety of needs. They are rigorously tested and certified to assure their dependability and performance. Technological advances have resulted in the production of more fuel-efficient and environmentally friendly aircraft, lowering environmental impact and boosting sustainability. Airports, airplanes and the aviation industry are all interrelated components of the worldwide air transportation system. Collaborative efforts, innovation, and sustainability measures are required to overcome obstacles and ensure the aviation industry's future is safe, efficient, and connected.

Nowadays, the demand for airline transportation is increasing significantly. People increasingly choose to travel by air, and it makes the amount of flights that fail to take off on time also increases and will cause delay (Tang, 2021). This growth makes airports even more crowded and harms the airline industry's finances. Delays in air travel are a sign of the aviation system's inefficiency. Both airline companies and their customers pay a hefty price for it. According to the estimation by the Total Delay Impact Study, the total cost of air transportation delay to air travelers and the airline industry in 2007 was $32.9 billion in the US, resulting in a $4 billion reduction in Gross Domestic Product (Khaksar & Sheikholeslami, 2019).

Next, with the increase of demands for flights, air load has been experiencing vigorous growth (Wang et al., 2022). Therefore, various airlines have an urgent need to predict the time-of-flight delays. According to the “Global Airports and Airlines Punctuality Report” the actual number of flights departing from airports around the world in 2019 was about 37.12 million, with a punctual departure rate of 75.58% and an average take off delay of 26.47 minutes (Tao et al., 2021). Flight delays not only disrupt passengers’ time planning, but also bring a lot of negative effects, and affect subsequent flight schedule and the image of the company.

Moreover, flight delays are also caused by adverse weather conditions. Adverse weather, such as storms, heavy snowfall, fog, or strong winds, poses a significant challenge to the aviation industry, leading to flight disruptions and delays. According to Gultepe et al. (2019), there were 8,657 aviation related accidents between 2003 and 2007, with weather being a factor in 1,740 of these accidents. Furthermore, weather was a primary contributing factor in 23% of all aviation accidents (Kulesa, 2003). The total national cost of the weather impact was estimated to be $3 billion due to accident damage and injuries, delays and unexpected operating costs.

To overcome this flight delay problem, a lot of research has been conducted and it has become a popular research area. Various researchers used different techniques of machine learning and data mining to conduct the investigation. For example, Esmaeilzadeh and Mokhtarimousavi (2020) used the support vector machine (SVM) to investigate the causes and patterns of air traffic delay at three major New York City airports. To determine their correlation with flight delay, airport operation, and flow management, a

https://dx.doi.org/10.24191/jcrinn.v9i2.460

few explanatory variables were evaluated. To better comprehend the reasons for departure delays, the odds of their causing the delay were estimated and compared.

This presentation commenced with a concise introduction to the background study and prevailing trends in the flight and aviation business. The second component of the study delved into the field of literature, specifically exploring the technological aspects of artificial intelligence with a particular emphasis on machine learning for data mining. The segment also provided an in-depth analysis of the Naïve Bayes method as it was proposed. The third element provided a justification for the framework used in the research process, ensuring that it may be conducted with rigorously, propriety, and reliability. The fourth section provides an analysis and evaluation of the results obtained from the developed model, as well as a comparison with state-of-the-art methods. The fifth component of the study is dedicated to concluding the discussion and presenting the findings.

2. LITERATURE REVIEW

The exponential rise in air travel also presents a number of difficulties. One major concern is the environmental impact of increased air traffic, leading to higher carbon emissions and contributing to climate change. Furthermore, logistical challenges may arise as a result of the increased workload on existing airport infrastructure and air traffic management systems which could lead to flight delays or airspace congestion. For example, in 2006, India experienced an increase in air passenger traffic. Significant air traffic operations are growing more quickly in Hyderabad (37%), Chennai (30.4%), Bangalore (31%), Mumbai (15.3%), and Delhi (19.6%). This has caused a substantial amount of congestion on the ground as well as in the air (Ramalingam, 2007). Concerns about security and safety measures have also arisen as a result of this increase in demand for air traffic, which require increased screening techniques and efficient risk management strategies. Moreover, increased numbers of tourists to popular destinations can place strain on local resources and infrastructures that lead to problems such as over tourism or degradation of natural and cultural heritage sites. Balancing the benefits of air travel with these challenges requires proactive measures and sustainable practices in the aviation industry.

2.1 Implementing Artificial Intelligence as Potential Solving Method

Artificial Intelligence plays an important role in addressing the issues posed by rapid growth of air travel. First, AI-powered algorithms can enhance air traffic control systems, enabling more effective routing and easing airspace congestion. It is in charge of directing and planning the movement of airplanes in airspace. Artificial Intelligence algorithms can create more efficient flight paths and suggest changes to maintain smooth and safe operations by analyzing real-time data, such as aircraft routes, weather conditions, and air traffic patterns. This optimization aids in reducing traffic and delays, improving on-time performance and the general flying experience for passengers.

Moreover, Artificial Intelligence can also help in forecasting and controlling weather patterns. Artificial Intelligence systems can produce precise forecasts by examining past weather data and present atmospheric conditions, enabling airlines to plan their flights intelligently. By doing so, it is possible to avoid bad weather, use less gasoline, and improve passenger safety and comfort. Nowadays, there is a lot of research that has been done to forecast weather using Artificial Intelligence. For an example, Biswas et al. (2018), have done a research on weather forecast prediction using Naïve Bayes algorithm.

By leveraging AI technologies, airlines can optimize fuel consumption and reduce carbon emissions. To identify opportunities for fuel savings, Artificial Intelligence algorithms can analyze large amounts of flight data such as aircraft performance, weather conditions and air traffic patterns. Optimization of climbing and descent rates, changes in flight speed or recommendations on alternate routes can be part of

https://dx.doi.org/10.24191/jcrinn.v9i2.460

this. By implementing these AI-driven recommendations, airlines can minimize fuel consumption and contribute to environmental sustainability.

A lot of research has been done to solve the problem of flight delays and it has become a popular research area. For this investigation, a range of researchers have used different techniques in machine learning and data mining. Tang (2021) has used dataset of flight leaving JFK airport for the whole year of 2020. The study was compared the effectiveness of various machine learning classification algorithms in the prediction of flight delays. There are many algorithms that have been used in this research such as Logistic Regression, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosted Tree. The research is conducted because of the increasing flight delays, which lead to financial difficulties and customer dissatisfaction for airline companies. The result of this research is Decision Tree model produces the highest values of accuracy, precision, recall and f1-score with the average value was 97.78%.

The need for accurate prediction of flight delays in the aviation industry is the reason why this research was conducted, (Tao et al., 2021). The aim of this research is to enhance the accuracy and efficiency of flight delay prediction by employing the LightGBM algorithm. The algorithms used is LightGBM algorithm, GBDT, and XGBoost algorithm. From this research, we can conclude that LightGBM algorithm is superior compared to GBDT and XGBoost algorithm. The R-squared, MAE and training time index show a higher prediction accuracy with 0.9748, 4.57(min) and 0.133(s) respectively.

Wang et al. (2022) has developed a prediction model that can suppose accurately estimate the exact arrival delay time for flights due to difficulty of accurately predicting flight delays leads to higher financial costs and increased passenger dissatisfaction due to the rising number of delays. The implementation of this study was using Random Forest Algorithm as prediction model. As the result, Random Forest Algorithm was claimed perform better than the other multi-classification methods and outperform other regression methods with the lowest Mean Square Error (MSE) rate by 3.23.

As for the flight delay prediction based on ARIMA by Wang and Pan (2022). The need to accurately predict flight delays in real-time was the factor why the study was conducted. The objective was to create a model that can accurately predict flight delays by using up-to-date flight data, offering airlines and airports valuable information to effectively handle and mitigate the impact of delays. A few algorithms used in this study such as K-Nearest Neighbor (KNN) algorithm, Grey Relational Analysis (GRA), Genetic Algorithm (GA) and ARIMA model. As the result of this study, ARIMA model was adopted to make flight delay time series stable by difference operation, and the flight delay time during holidays was predicted by parameter determination and residual test. The ARIMA model was claimed to consistence and stable accurately predict the changing trend of flight delay time and guarantee the prediction accuracy of flight delay time by showing minimum error reach0.4% and relative error less than 15%. It also can provide theoretical basis for airlines and airports to formulate response measures to flight delay.

Gui et al. (2019) has justified that the problem that lead to this study is the need for accurate flight delay prediction methods that are applicable across a wider range of factors and can overcome the issue of overfitting in limited datasets. The main objective was to analyze factors influencing flight delays, compare machine learning-based models for generalized prediction, and develop a robust model that achieves high accuracy and overcomes overfitting issues in limited datasets. The algorithm used in this study is Random Forest algorithm. The results of this study show that Random Forest based method can obtain good performance for the binary classification task which 90.2 %.

The study conducted by Nigam and Govinda (2017) inspired by some significant impact of aircraft delays on the aviation industry and passengers, leading to economic losses, inconvenience, and frustration. The main objective of this study is to accurately predict flight delays using logistic regression and machine

https://dx.doi.org/10.24191/jcrinn.v9i2.460

learning, incorporating weather and airport data, to enhance planning processes and minimize disruptions in commercial airlines. The algorithm used in this research is Logistic regression algorithm. The results of this study show that after testing the remaining 30% of data, the result they got around 80.6% accuracy using logistic regression.

2.2 The Prospective of Naïve Bayes as A Potential Classification Algorithm in Data Mining

Naïve Bayes is a simple but powerful algorithm that has been applied in the field of machine learning and natural language processing to classify and predict. It is based on Bayes' theorem, which describes the probability of an event given prior knowledge or evidence. According to Wickramasinghe and Kalutarage (2021), Naïve Bayes (NB) is a well-known probabilistic classification algorithm. It's an efficient but simple algorithm that can be used for a wide range of real-world applications, from product recommendations to medical diagnosis and the control of autonomous vehicles.

This algorithm is based on Bayes theorem with some assumptions. It presumes that the existence of a class is not dependent on other classes. The algorithm is based on conditional probability. The Bayes theorem defines the likelihood that an event will occur in relation to other events which have occurred before (Gnaneswar & Jebarani, 2017). Naïve Bayes classifiers are among the simplest Bayesian network models, but still, it can achieve high accuracy levels. It is easy and especially useful for very large sets of data to build an NB model. Naive Bayes, in combination with simplicity, can be regarded as superior to far more sophisticated methods of classification (Ray, 2023).

The concept of Naive Bayes is to develop a probabilistic model based on analysis of training data, which are labelled examples. It estimates the probability that the class identifier will be assigned to a new or unviewed instance, on the basis of probabilities for each feature. The algorithm calculates two probabilities which are Prior probability, and Conditional probability. According to Hayes (2023), Prior probability is the probability of an event happening prior to the collection of new data. In other words, it is the most appropriate assessment of the likelihood of a specific outcome based on current knowledge prior to an experiment. Conditional probability is the probability of a particular feature value given a class label. It is estimated based on the frequencies observed in the training dataset.

Eq. (1) is the Bayes theorem provides a way of computing posterior probability P (y| x) from P(y), P(x) and P (x| y).

$$P(\mathrm{y}|x) = P(x|\mathrm{y})P(\mathrm{y})P(x)$$
$$P(\mathrm{y}|X) = P(X1|\mathrm{y}) \times P(X2|\mathrm{y}) \times \ldots \times P(Xn|\mathrm{y}) \times P(\mathrm{y}) \quad (1)$$

- P(y| x) represents the posterior probability of class (c, target) given predictor (x, attributes).
- P(y) represents the prior probability of class.
- P(x| y) represents the likelihood which is the probability of the predictor given class.
- P(x) represents the prior probability of the predictor.

To classify a new instance, the algorithm calculates the posterior probability of each class label for that instance using Bayes' theorem. The class label with the highest posterior probability is assigned as the predicted class. If a space has many variables with continuous values, then the Gaussian Naive Bayes (GNB) equation from Gaussian distribution will be utilized. The equation is adjusted to Eq. (2).

https://dx.doi.org/10.24191/jcrinn.v9i2.460

Where, parameters σ_y^2 and μ_y are estimated using maximum likelihood.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

Naive Bayes is known for being easy to use, flexible and efficient especially when it comes to text classification tasks such as spam detection or sentiment analysis. However, in situations where there is a significant correlation between features the assumption of feature independence may limit its accuracy. Despite its "naive" assumptions, Naive Bayes can often provide reliable results and serve as a good baseline algorithm for classification problems, particularly when the dataset is large, and the features are relatively independent.

The architecture overview for the implementation of Naïve Bayes in classification as supervised learning is depicted in Fig. 1. First, the dataset will be loaded. Data shall be divided into training and tests data as soon as the dataset has been loaded. In the Training phase, the Likelihood Probability, and the probability of the Class Label with respect to the feature given, the Probability for the given data is computed using the Naïve Bayesian computation.

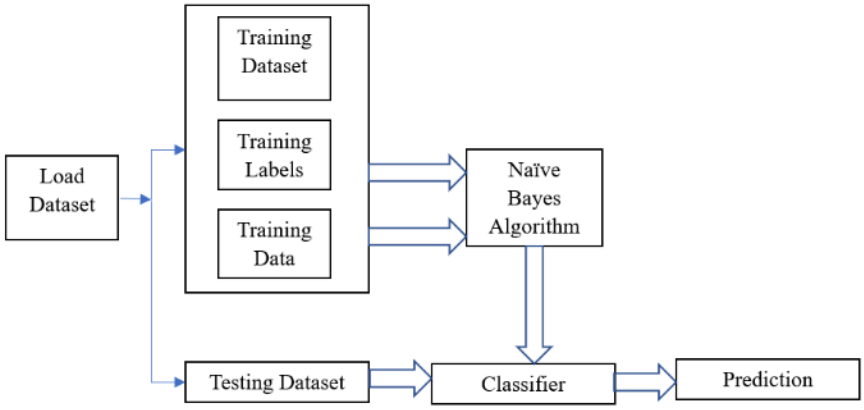

Fig. 1. Supervise learning architecture overview for Naïve Bayes implementation

Source: Venkatesh et al. (2020)

During the test phase, a feature with a maximum value shall be calculated by using a Naïve Bayes classifier and assigned to that class label. The classifier is based on finding functions for the probability that it belongs to a certain class label due to this feature. The classifiers could consist of Gaussian Nave Bayes, Bernoulli Naïve Bayes, or Multinomial Nave Bayes. The performance is predicted based on the Conditional Independence.

3. RESEARCH METHODOLOGY

The research methodology of this study provides the details implementation and methods involved in developing the Flight Delay Prediction Model. The project's flow and progress are also covered, along with the methodology that will be use.

https://dx.doi.org/10.24191/jcrinn.v9i2.460

3.1 The Architecture of the Propose Study

Fig. 2 illustrates the system architecture specifically for the flight delay prediction model. The process begins with the loading of the selected dataset, which then undergoes data pre-processing. The data preprocessing consists several important steps such as handling missing data, outliers, nominal attributes, data transformation and feature selection to ensure the data is prominent format for analysis.

The prominent data were divided into training data and testing data prior classification process. Classification procedure, which was used to create the prediction model, depends on this classification. In this instance, a classification model was used, more specifically the Naïve Bayes Model. The Naïve Bayes Model processes the training data to identify patterns and connections between features and flight delays. The model was then put to the test using the testing data in order to gauge its effectiveness and confirm its predicting abilities.

Once the classification process was completed, the prototype of the flight delay prediction model was implemented. Users can interact with the prototype, leveraging the insights and predictions generated by the model. This architecture showcases the systematic flow from data loading and preparation to classification model development, culminating in the implementation of a practical tool for flight delay prediction.

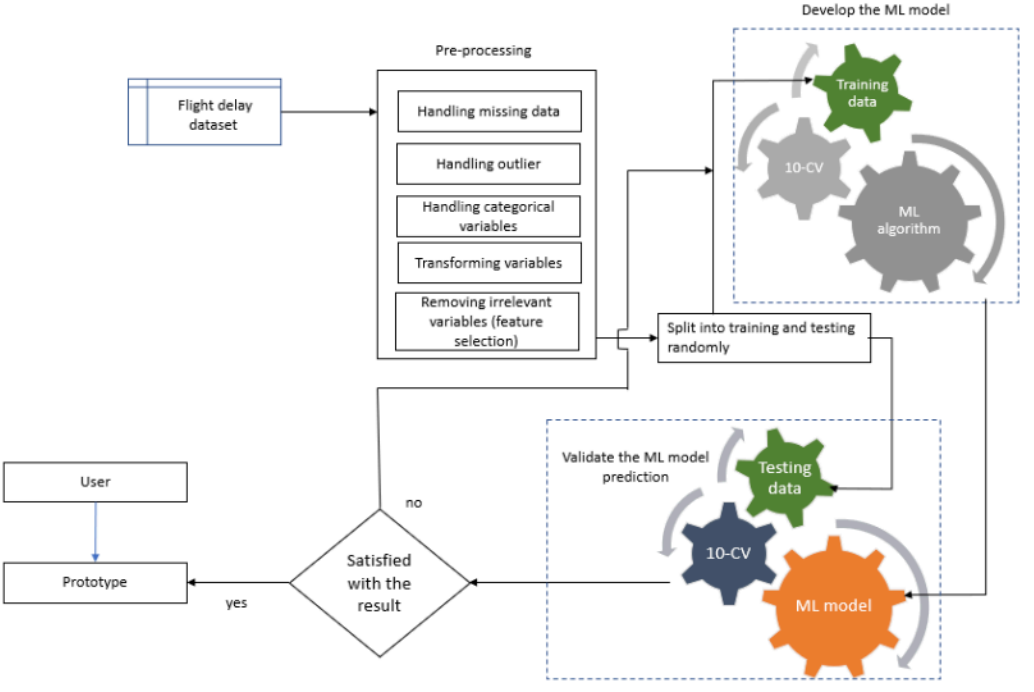

Fig. 2. Architecture overview of airline flight delay prediction prototype development

https://dx.doi.org/10.24191/jcrinn.v9i2.460

3.2 Data Description

The study utilised secondary data referred to as "The Flight Take Off Data from JFK Airport." The dataset is available for download on the Kaggle website (https://www.kaggle.com/datasets/deepankurk/flight-take-off-data-jfk-airport). It was posted by Deepankur Kansal in 2021 and can be accessed by the public. The dataset's size is about 2.6MB and in csv file format. This file contains data about flights leaving from JKF airport between November 2019 until December 2020. The dataset contains 28,821 samples and 23 attributes. By using the explorative data analysis technique data was analyzed in a the Jupyter Notebook. K-fold cross-validation technique, where K = 10, was used for splitting the dataset into training and testing data. Then, using the Naïve Bayes algorithm, the dataset will go through classification process. Table 1 shows the details of various attributes equivalent for 28,821 samples.

Table 1. Attributes of the dataset airline flight delay

No	Attributes	Description	Values
1	Month	Month	Numeric
2	DAY_OF_MONTH	Date of flight	Numeric
3	DAY_OF_WEEK	Day of the week	Numeric
4	OP_UNIQUE_CARRIER	Carrier Code (Should generally be carrier company)	Nominal
5	TAIL_NUM	Airflight Number	Nominal
6	DEST	Destination	Nominal
7	DEP_DELAY	Departure delay of the flight	Numeric
8	CRS_ELAPSED_TIME	Scheduled journey time of the flight	Numeric
9	DISTANCE	Distance of the flight.	Numeric
10	CRS_DEP_M	Scheduled Departure Time.	Numeric
11	DEP_TIME_M	Actual Departure Time (Gate checkout of	Numeric
12	CRS_ARR_M	Scheduled Arrival Time	Numeric
13	Temperature	Temp.	Numeric
14	Dew Point	Dew	String
15	Humidity	Hum	Numeric
16	Wind	Wind	Nominal
17	Wind Speed	Wind speed	Numeric
18	Wind Gust	Wind Gust	Numeric
19	Pressure	Pressure	Numeric
20	Condition	Condition of the climate	Nominal
21	sch_dep	No. of flights scheduled for arrival.	Numeric
22	sch_arr	NO. of flights scheduled for departure.	Numeric
23	TAXI_OUT	Taxi-out time (Run away time)	Numeric

3.3 Hyperparameters tuning

Parameter tuning, also known as hyperparameter tuning, refers to the process of adjusting the hyperparameters of a machine learning model to optimize its performance. This paper uniquely explores the hyperparameters tuning to gain the optimum performance of the naïve bayes model for classification specifically in flied delay dataset. In this parameter tuning evaluation, several adjustments were made to the parameters of the Gaussian Naïve Bayes (GNB) algorithm. Table 2 shows the details of parameters tuning, which the mechanism has been explained in previous section 2.2. The prior indicates prior probabilities of the classes specific to target attribute. By adjusting the prior parameters, it became possible to directly specify the value instead of relying on the frequencies in the data and having control over the computed probability of the GNB model. Meanwhile, var_smoothing was used to keep model from breaking down in the presence of missing data. The value was very small thus usually negligible. In this

https://dx.doi.org/10.24191/jcrinn.v9i2.460

study, the prior tuning values has been experimenting with several values closely to default values using flight delay data.

Table 2. Hyperparameter tuning for Gaussian Naïve Bayes algorithm

Parameter	Tuning
priors	[0.2, 0.8]
priors	[0.3, 0.7]
var_smoothing	1e-8

After run random prior values several times, the chosen value as shown in Table 2 will be further experimented for two different test splits, 80:20 and 70:30 of classification and the best result were recorded in the next section.

3.4 Evaluation Metrics

The evaluation was made based on the effectiveness and performance of the constructed model from Naïve Bayes algorithm. In this study, the performance for the model was evaluated using accuracy, precision and recall. To evaluate the performance, the output data from the predicted model was gathered using confusion matrix. Table 3 shows the confusion matrix constructed for flight delay prediction.

Table 3. Hyperparameter tuning for Gaussian Naïve Bayes algorithm

		Predicted	
		Delayed (positive)	**Not Delayed (negative)**
Actual	Delayed (positive)	True Positive (TP)	False Negative (FN)
	Not Delayed (negative)	False Positive (FP)	True Negative (TN)

True Positives (TP): This represents the cases where the model predicted a flight delay correctly, and the flight was actually delayed.

True Negatives (TN): This represents the cases where the model predicted a flight to be on time or not delayed, and the flight was indeed not delayed. The model made a correct negative prediction.

False Positives (FP): This represents the cases where the model predicted a flight delay, but the flight was not delayed. In other words, the model produced a false positive prediction.

False Negatives (FN): This represents the cases where the model predicted a flight to be on time or not delayed, but the flight was actually delayed. In other words, the model produced a false negative prediction.

3.4.1. Accuracy

Accuracy measures the rate at which model or system will make correct predictions relative to the total number of predicted results. It'll help to assess the full accuracy of these predictions. The accuracy expected from the analysis are True Positive (TP) and True Negative (TN) of flight delay prediction model. The system is more likely to make accurate predictions if accuracy is high, which is desirable. On the other side, if accuracy is low, it means that the system is producing more inaccurate predictions. The typical way to describe accuracy is as a percentage, with a greater percentage denoting better performance. To calculate the accuracy, the formula for accuracy is as stated in the Eq (3).

https://dx.doi.org/10.24191/jcrinn.v9i2.460

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{True\ Positive\ (TP) + True\ Negative\ (TN) + False\ Positive(FP) + False\ Negative(TN)} \quad (3)$$

3.4.2 Precision or Recall

Precision and recall are two evaluation metrics commonly used in classification tasks. They provide insights into the performance of a model in predicting positive instances. The precision and recall expected based on flight delay attribute is True Positive (TP). Precision measures how many of the predicted positive events turn out to be positive. It's focused on accuracy of positive forecasts. Recall or often referred to as sensitivity or true positive rate, quantifies the proportion of actual positive events that the model accurately recognized. It focuses on the model's capacity to identify good examples. To calculate precision and recall, the formulas below will be used.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Positive(FP)} \quad (4)$$

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + \ False\ Negative(TN)} \quad (5)$$

3.4.3. Specificity

Specificity is a metric that evaluates the model's accuracy in correctly identifying negative cases. It precisely measures the proportion of accurately predicted not delayed flights out of all the actual not delayed flights. Equation (6) illustrates the concept of specificity.

$$Specificity = \frac{True\ Negative\ (TN)}{True\ Negative\ (TN) + \ False\ Positivr\ (FP)} \quad (6)$$

4. RESULT AND DISCUSSION

The implementation of this study was previously mentioned and described in section 3.1. Right before the classification model training phase was implemented, the correlation among all attributes were examined by using Pearson correlation matrix to provide an overview of the patterns and relationship among them. Fig. 3 illustrate the Heatmap of correlations among attributes and aids in deciding which attribute has less correlation should be dropped, meanwhile the rest were retained.

4.1 Analysis Using Naïve Bayes Algorithm

Next, several analyses had been implemented in order to thoroughly analyzed the constructed model of Naïve Bayes for flight delayed. For the purpose of testing, the Hold-out method was employed where data for training and testing in Fig. 2 were split firstly into 80:20 ratio and secondly into 70:30. These two

https://dx.doi.org/10.24191/jcrinn.v9i2.460

sets of analyses were conducted to determine the optimal ratio of the hold-out approach that yields the most favorable outcome.

The GNB approach in Eq (2) captures the mean and variance of the attributes values by modelling the distribution of each attributes for each target class. During the training phase, class priors and likelihoods parameters were calculated based on the provided training data (both 80% and 70%). The algorithm uses Bayes' theorem to calculate the posterior probability from Eq (1) for each class based on observed features (the attributes) when making predictions. Then it was predicted that the class with the highest probability of posterior will be output.

According to Table 3, the model's performance with 70:30 data splitting is optimal when the parameter priors are set to [0.2, 0.8], achieving the highest accuracy at 93%. In comparison, both the default parameters and other tuning variations exhibit the same performance, yielding an accuracy of 92%. Furthermore, the specificity and ROC AUC scores remain consistent across all experiments, standing at 88% specificity and 97% ROC AUC score.

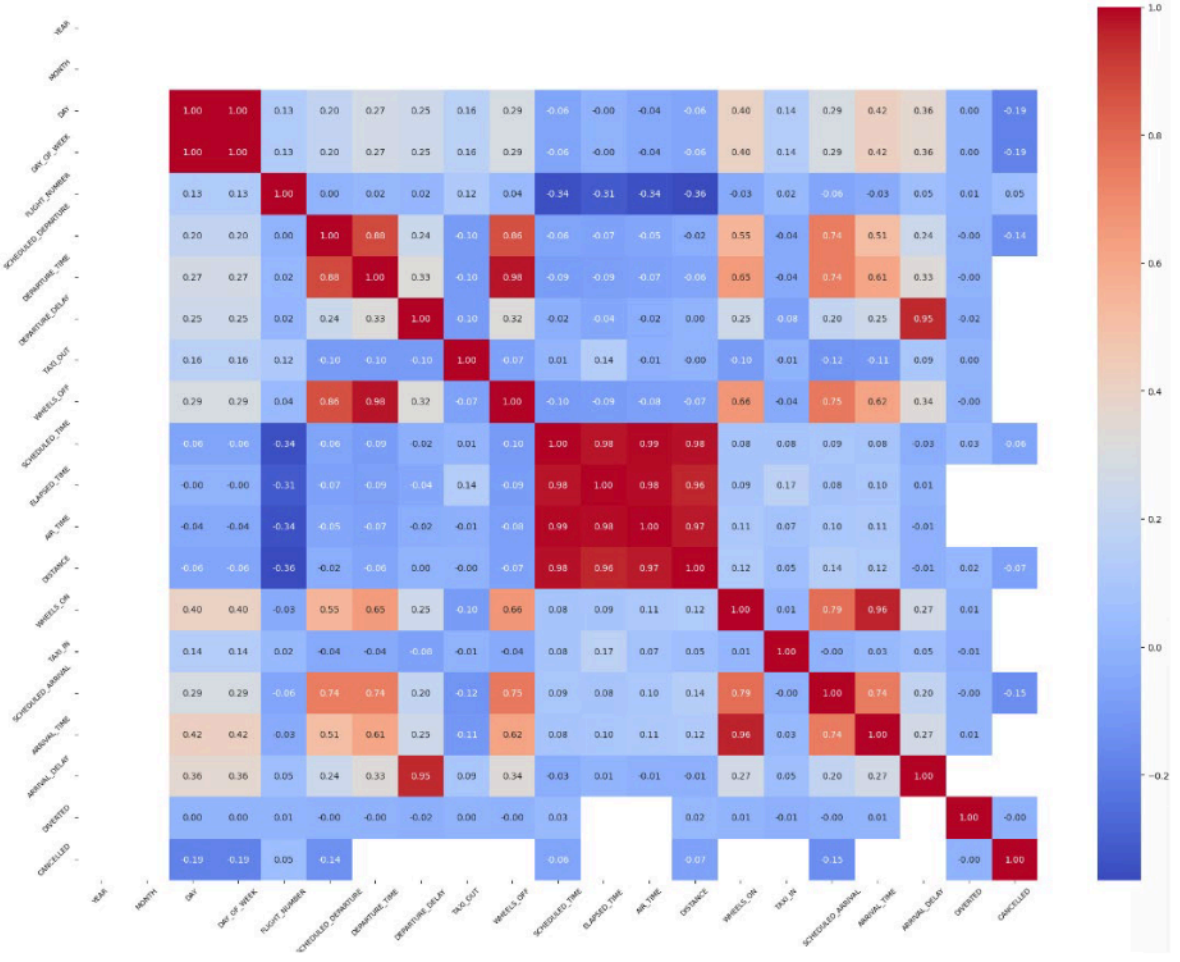

Fig. 3. Heatmap of correlation among attributes in flight delayed dataset.

https://dx.doi.org/10.24191/jcrinn.v9i2.460

Table 3. The model's performance for 70:30 data splitting with parameters tuning

Parameters	Accuracy	Specificity	ROC AUC score
Default	92%	88%	97%
Prior = [0.2, 0.8]	93%	88%	97%
Prior = [0.3, 0.7]	92%	88%	97%
Var_smoothing=1e-8	92%	88%	97%

Based on the data presented in Table 4, the model achieves optimal performance with a 80:20 data splitting when the parameter priors are set to [0.2, 0.8], attaining the highest accuracy of 93%. In contrast, both the default parameters and other tuning variations show similar performance, resulting in an accuracy of 92%. Additionally, the specificity and ROC AUC scores remain constant across all experiments, registering at 87% specificity and a 96% ROC AUC score. After analyzing both Table 3 and Table 4, we have determined that adjusting the hyperparameters in the Naïve Bayes algorithm does not have a substantial influence on increasing the performance of the model.

Table 4. The model's performance for 80:20 data splitting with parameters tuning

Parameters	Accuracy	Specificity	ROC AUC score
Default	92%	87%	96%
Prior =[0.2, 0.8]	93%	87%	96%
Prior =[0.3, 0.7]	92%	87%	96%
Var_smoothing=1e-8	92%	87%	96%

4.2 Comparing Naïve Bayes Performance with Others Algorithm

In this part, the result of each algorithm will be compared. The algorithm used for the comparison is K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) and Gaussian Naïve Bayes. The parameter for each algorithm is set to default except Gaussian Naïve Bayes where the priors is set to = [0.2, 0.8]. Table 5 shows the result of each algorithm and it can be concluded that Gaussian Naïve Bayes have the better evaluation performance compared to KNN and SVM. The performance of Gaussian Naïve Bayes is much better than KNN and SVM with 93% accuracy and 97% ROC AUC score.

Table 5. Comparing Naïve Bayes model's performance with two other algorithms

Algorithm	Accuracy	ROC AUC score
Gaussian Naïve Bayes (priors = [0.2, 0.8]	93%	97%
K-Nearest Neighbors (KNN)	85%	63%
Support Vector Machine	84%	92%

https://dx.doi.org/10.24191/jcrinn.v9i2.460

Fig 6, 7 and 8 show the ROC AUC graph for each algorithm. From those three graphs, it can be concluded that Gaussian Naïve Bayes have the highest ROC score which is 0.96 followed by SVM with 0.92 and KNN with 0.63.

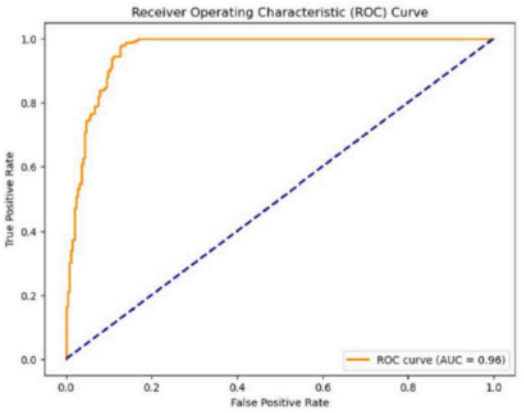

Fig. 6. ROC curve for Gaussian Naïve Bayes

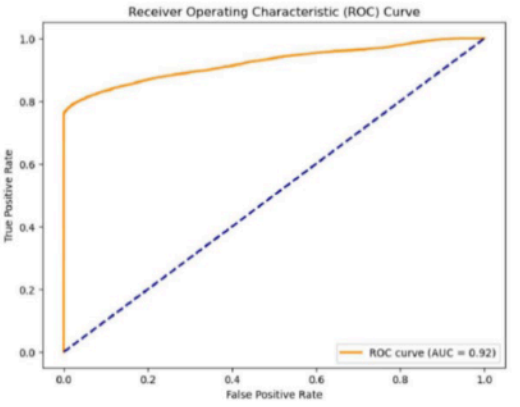

Fig. 7. ROC curve for Support Vector Machine (SVM)

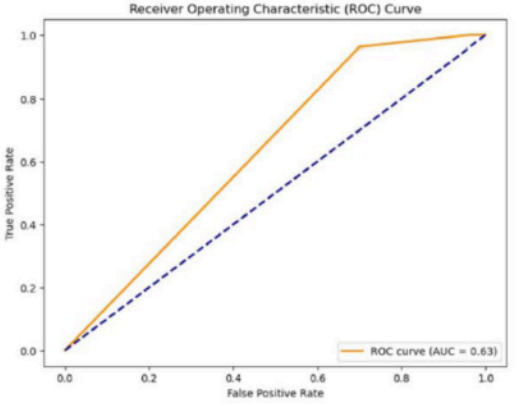

Fig. 8. ROC curve for K-Nearest Neighbor (KNN)

5. DISCUSSION AND CONCLUSION

This study specifically aimed to develop a reliable flight delay prediction model using the Naïve Bayes algorithm on the benchmarking dataset. The implementation has been carefully designed starting at pre-processing step and classification approach for the flight delayed dataset and Gaussian Naïve Bayes algorithm. To evaluate the performance of the flight delay prediction model using the Naïve Bayes algorithm the training and testing process was evaluated using three metrics: accuracy, specificity, and ROC AUC score and tested with hold out method of split test and incorporate with parameters tuning specifically to tune the best performance from Gaussian Naïve Bayes algorithm. Additionally, the algorithm of Gaussian

https://dx.doi.org/10.24191/jcrinn.v9i2.460

Naïve Bayes algorithm was compared with other algorithms, such as K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) to determine their relative performance.

In the split test, the algorithm demonstrated its highest performance when the data was split at a ratio of 70:30. As for parameter tuning, the key parameter adjusted for this algorithm is "priors," which was set to [0.2, 0.8]. Furthermore, a thorough comparison of performance with other algorithms was conducted, specifically with K-Nearest Neighbors (KNN) and Support Vector Machine (SVM). The results of this comparison indicate that Naïve Bayes outperforms the other models, with SVM following closely behind and KNN ranking third in terms of predictive accuracy. These findings validate the effectiveness of the chosen Naïve Bayes algorithm in addressing the challenges of predicting flight delays outlined in the project. Moreover, the inherent nature of the Naïve Bayes formula allows for the inclusion of conditional probability information regarding uncertain events, which might be useful in formulating a model for predicting flight delays.

There are several limitations faced when conducting the flight delay prediction using the Naïve Bayes algorithm project, including limited handling of numeric data. The Naïve Bayes algorithm performs well with categorical data, but it faces challenges when handling numeric data. The algorithm assumes a normal distribution of numeric features, which might not align with the distribution of real-world flight data. Moreover, since most of the flight data comes in numeric form, this limitation can impact the algorithm's ability to accurately model and predict flight delays. The second limitation is data imbalance. Imbalances in the dataset, such as an excessive number of on-time flights compared to delayed flights, can impact the algorithm's effectiveness in accurately predicting delays. The Naïve Bayes algorithm may exhibit bias towards the majority class, leading to potential challenges in accurately capturing and predicting instances of flight delays. The

6. ACKNOWLEDGEMENTS/FUNDING

The authors would like to acknowledge the support of Universiti Teknologi Mara (UiTM), Cawangan Pulau Pinang and Universiti Teknologi Mara (UiTM), Cawangan Terengganu specifically to College of Computing, Informatics, and Mathematics, UiTM Cawangan Terengganu for the support given in the collaborative project and research.

7. CONFLICT OF INTEREST STATEMENT

There is no conflict of interests in the research.

8. AUTHORS' CONTRIBUTIONS

Ahmad Adib Baihaqi Shukri: Carried out the research and wrote the first draft. **Syarifah Adilah Mohamed Yusoff**: Supervised the project and edit the article. **Saiful Nizam Warris** and **Mohd Saifulnizam Abu Bakar:** validation of data and result. **Rozita Kadar**: Conceptualisation, writing- review and editing.

9. REFERENCES

Biswas, M., Dhoom, T., & Barua, S. (2018). Weather forecast prediction: An integrated approach for analyzing and measuring weather data. *International Journal of Computer Applications*, *182*(34), 20-24.

https://dx.doi.org/10.24191/jcrinn.v9i2.460

Buhalis, D. (2004). eAirlines: Strategic and tactical use of ICTs in the airline industry. *Information & Management*, *41*(7), 805-825. https://doi.org/10.1016/j.im.2003.08.015

Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine learning approach for flight departure delay prediction and analysis. *Transportation Research Record*, *2674*(8), 145-159. https://doi.org/10.1177/0361198120930014

Freestone, R. (2009). Planning, sustainability and airport-led urban development. *International Planning Studies*, *14*(2), 161-176. https://doi.org/10.1080/13563470903021217

Gnaneswar, B., & Jebarani, M. E. (2017). A review on prediction and diagnosis of heart failure. In *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-3). IEEE Xplore. https://doi.org/10.1109/ICIIECS.2017.8276033

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2019). 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

Gultepe, I., Sharman, R., Williams, P. D., Zhou, B., Ellrod, G., Minnis, P., Trier, S., Griffin, S., Yum, S. S., & Gharabaghi, B. (2019). A review of high impact weather for aviation meteorology. *Pure and Applied Geophysics*, 176, 1869-1921. https://doi.org/10.1007/s00024-019-02168-6

Hayes, A. (2023). *Bayes' Theorem: What It Is, the Formula, and Examples.* https://www.investopedia.com/terms/b/bayes-theorem.asp

Khaksar, H., & Sheikholeslami, A. (2019). Airline delay prediction by machine learning algorithms. *Scientia Iranica*, *26*(5), 2689-2702. https://doi.org/10.24200/SCI.2017.20020

Kulesa, G. (2003*)*. Weather and aviation: How does weather affect the safety and operations of airports and aviation, and how does FAA work to manage weather-related effects? *In The Potential Impacts of Climate Change on Transportation US Department of Transportation Center for Climate Change and Environmental Forecasting*. US Environmental Protection Agency; US Department of Energy; and US Global Change Research Program. http://climate.dot.gov/documents/workshop1002/kulesa.pdf

Nigam, R., & Govinda, K. (2017). Cloud based flight delay prediction using logistic regression. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 662-667). https://doi.org/10.1109/ISS1.2017.8389254

Ray, S. (2023). Naive bayes classifier explained: Applications and practice problems of naive bayes classifier. *Analytics Vidhya*, *11*(9).

Ramalingam, K. (2007). Challenges in Indian civil aviation and opportunities to designers and manufacturers-2007. *Journal on Design and Manufacturing Technologies*, *1*(1), 5-10.

Tao, J., Man, H., & Yanling, L. (2021). Flight delay prediction based on LightGBM. In *2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)* (pp. 1248-1251). IEEE Xplore. https://doi.org/10.1109/ICCASIT53235.2021.9633431

Tang, Y. (2021). Airline flight delay prediction using machine learning models. In *the Proceedings of 2021 5th International Conference on E-Business and Internet* (pp. 151-154). ACM Digital Library. https://doi.org/10.1145/3497701.3497725

Venkatesh, Ranjitha, K.V., Venkatesh Prasad, B.S. (2020). Optimization scheme for text classification using machine learning naïve bayes classifier. In A. Kumar, M. Paprzycki, & V. Gunjan, (eds) *ICDSMLA 2019. Lecture Notes in Electrical Engineering*, vol 601. Springer.

https://dx.doi.org/10.24191/jcrinn.v9i2.460

https://doi.org/10.1007/978-981-15-1420-3_61

Wang, J., & Pan, W. (2022). Flight delay prediction based on ARIMA. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)* (pp. 186-190). IEEE Xplore. https://doi.org/10.1109/ICCEAI55464.2022.00047

Wang, Z., Liu, H., & Chu, F. (2022). Flight arrival delay time prediction based on machine learning. *In 2022 3rd Asia Conference on Computers and Communications (ACCC)* (pp. 35-39). IEEE Xplore. https://doi.org/10.1109/ACCC58361.2022.00013

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277-2293. https://doi.org/10.1007/s00500-020-05297-6

https://dx.doi.org/10.24191/jcrinn.v9i2.460