Predicting Airline Passengers Satisfaction with Classification Algorithms

Amelia Rahmanita[1], Bill Kiki[2], Meilani Yaputri[3], Mettadewi[4], Reynardthan Handoko[5]

[1] Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara
amelia.rahmanita@student.umn.ac.id
[2] Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara
bill.kiki@student.umn.ac.id
[3] Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara
meilani.yaputri@student.umn.ac.id
[4] Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara
mettadewi@student.umn.ac.id
[5]Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara
reynardthan.handoko@student.umn.ac.id

***Abstract* — The primary goal of this study is to determine if a customer or passenger would be satisfied or dissatisfied with the current service offered by airlines, based on the details of the other parameters values, feedback, and flight data. Five classification algorithms: Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest, were employed in this study. However, before that, the features that will be utilized to construct the model will be selected based on their correlation with the target variable. The AUC value, precision, and recall values of each model will then be compared to identify which modeling approach has the best performance in predicting and classifying passenger satisfaction.**

Index Terms* — *Airline; Airplane; Big Data; Classification; Data Mining; Decision Tree; K-Nearest Neighbors; KNN; Logistic Regression; Naive Bayes; Passenger; Random Forest; Satisfaction; Supervised Machine Learning.

I. INTRODUCTION

A. Background

The service sector has surpassed manufacturing as the most crucial sector of the global economy. Over the last 30 years, the commercial air transport industry has grown exponentially. There are currently over 270 international airlines that transport over 3.8 billion people each year [1]. Every day, a huge number of aircraft operate to link various geographical regions. Therefore, we may expect a huge number of passengers to travel on these aircraft every day. This is due to the fact that air travel is one of the most convenient forms of long-distance travel on both a national and worldwide scale [2]. One of the most important forms of modern mobility is air transportation. As a result, there are several airline service providers (ASPs) all over the world. Customers may select the airlines' flight services depending on their preferences because there are numerous alternatives. As a result, airline rivalry is becoming more severe, and airline service quality is attracting more attention than ever before. With rising air traffic and passenger traffic congestion, it is critical to maintain persistence and resilience. As a result, all ASPs are working hard to enhance their service quality, facility, and in-flight comfort in order to attract and please consumers. To maintain their competitive edge, airlines must continually search for methods to differentiate themselves from the competition.

In a competitive market such as the transportation sector, particularly the airline business, it is critical for enterprises to not only accurately assess what their consumers want and expect, but also to manage their own resources appropriately in satisfying those expectations. Service quality and customer satisfaction are increasingly being acknowledged as key determinants of corporate performance and strategic tools for achieving a competitive advantage. The customer's entire image and judgment of the service given is described as service quality, whereas satisfaction is an immediate reaction to consumption [3]. Because service quality is regarded as the foundation for customer satisfaction, a high degree of service quality should be given by the service provider in order to attain a high level of customer satisfaction. High-quality service has become a market necessity for air carriers, and it assists firms in gaining and retaining client loyalty.

Customer satisfaction is an emotional response that arises from a cognitive process that compares the expense of acquiring the service to the cost of receiving the service [4]. Simply defined, satisfaction is the sense of joy or disappointment that customers have when comparing a product's perceived performance to their past expectations [5]. Customers are more likely to be satisfied with an airline when the essential service quality criteria are fulfilled or exceeded. A high level of passenger satisfaction is considered a significant business advantage in order to enhance their profile among consumers and

separate themselves from the competition in such a highly competitive industry.

Customer satisfaction is crucial in motivating customers' behavioral and brand loyalty, which translates into positive reviews and valuable feedback, purchasing more products more frequently, recommending the product or service to others, and being less likely to be lost to competitors than dissatisfied customers [6]. Numerous studies show that airlines' market share, revenues, positive word of mouth, and customer retention are all dependent on consumer perceptions of service quality, and hence on customer satisfaction and loyalty [7]. Airlines are also starting to realize a significant correlation between customer satisfaction, customer retention, and profitability. Retaining current clients is considerably more profitable and cost-effective than recruiting new ones to replace those who have been lost. Unsatisfied passengers, on the other hand, may rethink using the same airline on future trips or launch bad word-of-mouth and social media campaigns that might harm the company's image, reputation, and profitability [8]. According to research, most consumers are less likely to complain to the company and are more likely to tell their friends about their poor experience.

Measuring customer satisfaction in the airline sector is becoming increasingly common and necessary, as it is critical to the industry's long-term survival and competitiveness [9]. In this situation, big data technologies and machine learning are beneficial for analyzing a massive airline's passenger database and developing highly accurate predictions or categorizations. Thus, the purpose of this research is to classify airline passengers' satisfaction using five supervised machine learning algorithms, Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest. The five algorithms will then be compared in order to determine the best and most accurate method for categorizing passenger pleasure based on the given parameter. The study opted to create a classification model for two types of satisfaction, satisfied and unsatisfied, as well as the best algorithm to go with it.

B. Problem

The airline sector is a potential contributor to the global economy since it facilitates economic development, worldwide trade, tourism, and global investment [10]. This results in the increasing use of airline services in recent years because of the time efficiency obtained when using airplane transportation. The increase in customer demand for these services then opens up opportunities and compels many airlines to improve both the quantity and quality of their services. Airline companies must improve the quality of their services to retain customers to remain loyal to using their services and compete with other airline competitors.

An airline's degree of customer satisfaction and performance may be determined using customer feedback and flight data. As a result, customer feedback is critical for any airline company [2]. Airlines can use this processed data to improve the quality and performance of their services in order to boost consumer satisfaction. This information may also be used to examine parts of their service in order to determine what sort of service consumers like. From the data processing and analysis results, it is believed that airlines would be able to attract new consumers, keep the former ones loyal, and boost customer satisfaction ratings by improving their performance.

II. LITERATURE STUDY

A. Big Data and Data Mining

Big data is a collection of data in huge volume, including both structured and unstructured data, which necessitates using a modern and sophisticated system to handle it. Big data may be anything from terabytes and petabytes in size. Big data has three primary characteristics: velocity, which is the rate at which data changes over time; volume, which is the amount of the data; and diversity, which is the kind and format of data and the method of data processing [11].

Data mining is the process of handling information from various data to discover patterns or relationships to make valid predictions. Hence, data mining can be used and useful for multiple purposes, sectors, and fields. One of them is to advance science in the field of artificial intelligence and statistics, as data mining is predicted to become a highly needed and revolutionary branch of science over the next decade, according to MIT Technology Review. Data mining can help many fields, including manufacturing, banking, insurance, marketing, aerospace, education, and health. One of these is this study, which examines airline passenger satisfaction in terms of lowering costs, improving research, growing sales, and improving the efficiency of their companies [12].

The most basic and initial analytical stage in data mining is to describe the data by summarizing its statistical features (means and standard deviations), visually evaluating it with charts and graphs, and looking for potentially significant relationships between variables. Data mining techniques such as classification are used to create models that describe important data classes and assign items in a collection to target categories. The goal of classification is to properly predict category labels and differentiate one item from another based on its characteristics. The data will be split into two data sets, training and testing, in this approach. To identify the category label, the training data will be utilized to create the

model and show the related data. The testing data, on the other hand, will be utilized to test and forecast the model in order to identify the category labels [13].

However, in general, data mining has four types of data learning procedures, which are as follows [14]:

a) Classification Learning

A set of various classified instances from the concepts is intended to be learned for classifying the unseen examples.

b) Association Learning

Different types of association between different characteristics are considered to predict a specific class value.

c) Clustering

Clustering is a method of sorting out the prediction process by forming groups of comparable values, parameters, and types.

d) Numeric Prediction

The numeric prediction is based on numerical quantities and not on a discrete class.

B. Supervised Machine Learning

Machine Learning (ML) is one of computer science's subfields that focuses on deriving patterns from data in order to enhance performance on a variety of tasks. Some of the data patterns that need to be recognized include non-linear correlations, interconnections, underlying dimensions, or subgroups of individuals. Machine learning is used to train machines to process data more efficiently and comprehend the derived information. Automated, highly flexible, and computationally intensive ways of discovering patterns in large data structures are commonly referred to as machine learning [15].

Supervised learning is a machine learning activity that involves learning a function that translates input to an output based on examples of input-output pairs [16]. It derives a process from labeled training data that consists of a collection of training instances. Supplied with external aid, supervised machine learning algorithms are ones that require external support. The input dataset is separated into two sections: train and test. The training dataset contains an output variable that must be predicted or categorized. Each algorithm learns and extracts patterns from the training set then applies them to the test set to make predictions or classify objects.

C. Logistic Regression Algorithm

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. Logistic regression is used to develop a regression model when the dependent variable is categorical. In this algorithm, the probabilities describing the possible results of a single trial are modeled using a logistic function. Logistic regression predictive model used to evaluate the relationship between the dependent variable(target) and the independent variable (predictor). The most common logistic regression models a binary outcome; something that can take two values such as true(1) or false(0), yes(1) or no(0). Each variable must have collinearity, and if there is multicollinearity, it cannot be predicted [17].

Logistic regression is useful when predicting based on the values of independent variables. Logistic regression is a non-linear regression with a dichotomous dependent variable. Logistic regression allows for the prediction of a specific outcome, such as group membership, based on variables that can be dichotomous, continuous, discrete, or a combination of all of these. The response variable is usually categorical, such as absence/presence or failure/success. When the explanatory variables are categorical or a mix of categorical and continuous, logistic regression is preferred [18].

The advantages of using logistic regression are [18]:

1. Logistic regression can be shown significant relation between dependent and independent variables.
2. Logistic regression not only classification model, also provides information related to probability.
3. Assumes that errors are sketched from a binomial distribution

The disadvantages of using logistic regression are:

1. Extremely sensitive to multicollinearity
2. The assumption on dichotomous dependent variable

D. K-Nearest Neighbors Algorithm

The K-Nearest Neighbor (KNN) algorithm is the most widely used and important classification method in data mining [19]. Furthermore, KNN is included in the supervised learning group, meaning the results of newly categorized query instances in KNN based on the majority of proximity to existing categories [20].

K-Nearest Neighbor is a simple method that classifies and saves all accessible data and calculated data based on nearest train data sets obtained by a simple majority vote of the K-Nearest Neighbor of each point [21]. It also categorizes unlabeled data points into distinct categories. This technique is simple to construct and comprehend, but it suffers from the severe disadvantage of becoming much slower as the size of the data in use rises.

The KNN algorithm becomes an alternative since it is adept at dealing with noise, is uncomplicated,

simple, and straightforward, and employs very big computerized data sets [19]. However, even as the best and most important method, K-Nearest Neighbors has numerous drawbacks, including a very high computing cost, sensitivity to irrelevant characteristics, a longer run time (lazy algorithm), and a huge memory need for storing all of the training data [22].

E. Naive Bayes Algorithm

The Naive Bayes algorithm is a supervised nonlinear and simple probabilistic algorithm based on Bayes' theorem application with the assumption of strong (naive) independence between every pair of features or variables. This algorithm is called "naive" because it makes the assumption that the occurrence (or absence) of a certain feature does not depend or unrelated to the occurrence (or absence) of other features [23]. Naive Bayes itself was first proposed by a British scientist named Thomas Bayes. Naive Bayes classifier also works great in many real-world situations such as document classification and spam filtering [24].

Naive Bayes is one of the most straightforward yet powerful classification method in Machine Learning applications due to its computational efficiency or simplicity, which allows equal contribution of all attributes to the final decision. This is what makes its technique interesting and suitable for various sectors. Naive Bayes classifier has three aspects in its main element, namely prior, posterior, and class conditional probability [24]. There are many advantages of using Naive Bayes algorithm, which include relatively easy to understand and create, takes less time to predict classes compared to other classification algorithms, and only requires small training data. However, Naive Bayes also have some disadvantages, namely "Zero Conditional Probability Problem," where it assigns zero probability or wipes out all the information in the other dataset, and dataset can be difficult, even almost impossible, to find due to the very strong assumption of independent class features [25].

F. Decision Tree Algorithm

Decision Tree is a well-known machine learning technique that may be used to solve both regression and classification problems. Decision Tree addresses the machine learning challenges by converting data into a tree representation. This approach is commonly used in data mining to create classification models. A decision tree is used to build a data model that predicts variable labels or values for use in decision-making. The model is constructed using the system's training data set (supervised learning). It presents an accurate and simple model under specified settings and examines sequential problems as well as their repercussions before making a choice [26].

A decision tree is made up of three parts: the root, the node, and the leaf. Each internal node of the tree represents an attribute, whereas each leaf node represents a class label. Each leaf is given to a class that corresponds to the most relevant target value. In other words, it is a sort of flowchart in which each point represents a test and its associated outcome. Decision trees function differently depending on the size and kind of data collection. They are resilient, can handle both category and numerical data types, and can process enormous amounts of data in a short amount of time.

There are several types of Decision Tree algorithms, such as Iterative Dichotomies 3 (ID3); Successor of ID3 (C4.5); Classification and Regression Tree (CART); Chi-Squared Automatic Interaction Detector (CHAID); Multivariate Adaptive Regression Splines (MARS); Generalized, Unbiased, Interaction, Detection, and Estimation (GUIDE); Conditional Inference Trees (CTREE), Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE); and also Quick, Unbiased, and Efficient Statistical Tree (QUEST) [27].

The Decision Tree method has the benefit of requiring less work for data preparation during pre-processing. Decision Tree additionally does not need data normalization or scaling. A missing value has no effect on the decision tree's decision-making process. Finally, the Decision Tree model is extremely intuitive in terms of creating visuals that are simple to explain and comprehend in order to aid in decision-making or drawing conclusions. The drawback of utilizing the Decision Tree method is that its computations might be more complex than those of other algorithms at times. Small changes in the data might generate significant changes in the decision tree structure, resulting in instability. Furthermore, these methods frequently necessitate a longer period to create the model.

G. Random Forest Algorithm

The random forest algorithm is a supervised method that generates and combines several decision trees at random. Random forest fits a number of decision trees on diverse subsamples of datasets and utilizes the average to increase the prediction accuracy of the model while avoiding overfitting. The size of the sub-sample is always the same as the size of the original input sample, but the samples are generated using replacement.

Random Forests are frequently used to address classification, regression, and other types of issues. Two factors contribute to the randomness of this method, namely [28]:

- Each tree grows on a unique bootstrap sample drawn at random from the training data.

- During decision tree creation, a sample of m variables from the original data set is taken and the best one is utilized in each split node.

Random Forest may also be used to determine the significance of variables. This is accomplished through the use of OOB data. Each variable m is permuted randomly, and the permuted OOB situations are recursively passed down the tree. The significance value of variable m is calculated by subtracting the number of properly categorized cases using permuted data from the number of correctly classified instances using non-permuted data. Although these values vary for each tree, the average of each value throughout the whole forest yields a raw significance score for each variable [29].

Random Forest includes an internal method for estimating its generalization error. This process is referred to as the out-of-bag (OOB) error estimate. In tree construction, bootstrap sampling was performed on just 2/3 of the original data. While the remaining 1/3 is categorized according to the tree generated and utilized to evaluate the algorithm's effectiveness. The OOB error estimate is the average prediction errors for each training instance y using a tree that does not include y in the bootstrap sample. When the RF is created, all training cases traverse each tree, and the proximity matrix for each instance is derived using the pair of examples that arrive at the same terminal node. Numerous studies have demonstrated that Random Forest performs well in regression and classification in a variety of disciplines, including financial forecasting, remote sensing, and genetic and biological research. Additionally, RF outperforms other techniques such as partial least squares regression, support vector machine, and neural network [28].

III. METHODOLOGY

A. *Object of Research*

The study focuses on preparing and analyzing consumer satisfaction data based on airline services. The parameters used to measure customer satisfaction include satisfaction, gender, customer type, age, type of travel, class, flight distance, seat comfort, departure/arrival time convenient, food and drink, gate location, flight WiFi service, inflight entertainment, online support, ease of online booking, on-board service, leg room service, baggage handling, check-in service, cleanliness, online boarding, departure delay in minutes, and arrival delay in minutes.

The dataset we use contains approximately 130,000 observations and 24 variables, of which all observations represent airline passengers. If specified, the 24 variables consist of 5 categorical variables, 14 ordinal categorical variables, and 5 numerical variables, with satisfaction as the target variable.

B. *Methods of Collecting Data*

The research data used is secondary data gathered from other parties for research purposes. The data was obtained via an actual survey technique, according to the data provider's assertion. Because of the restricted time available and the limited number of sources available to collect data, secondary data was chosen and used for this study.

The secondary data used in this study was downloaded via the Kaggle website. Specifically, Airline Customer Satisfaction published by Prachi Gupta on the Kaggle website (https://www.kaggle.com/datasets/prachi15gupta98/airline-passenger-satisfaction) will be used. Kaggle is a website or platform dedicated to Data Science and Machine Learning in general. This website offers a wide range of datasets on numerous topics that may be utilized as raw data in research. In addition, Kaggle frequently hosts Data Science contests and acts as an online community service provider for Data Science activists to exchange expertise.

C. *Methods of Research*

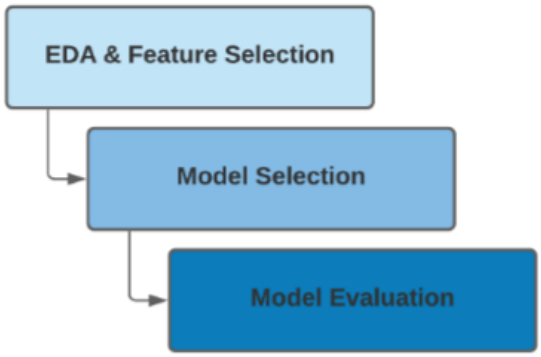


Fig 1. The Research Framework

To analyze the data, the programming language Python in the program Jupyter Notebook is utilized. First the selected data in the format .csv will be loaded into the Jupyter Notebook.

The missing data is then handled by eliminating missing values such as NaN (Not a Number) and unused variables. Following that, we alter the data by changing the data type of our target variable from categorical to numeric data type, satisfied = 1, neutral or dissatisfied = 0. Data transformation on class feature (Ticket type) with, Eco: Economy, Eco Plus: Economy, and Business: Business. All variables whose data types have been modified are subsequently saved into a new data frame. We also remove some variables; Departure Delay In Minutes, Arrival Delay In Minutes features, and ID column.

Furthermore, various data variables will be displayed in graphs based on their data type utilizing

descriptive statistical approaches to aid in data analysis. Barplots are used to illustrate the data distribution of categorical variables, while boxplots are used to illustrate the data distribution of numerical variables.

Following exploratory data analysis, the data frame is separated into two parts, training and testing, with an 80:20 composition ratio, which 80% for training set and 5-fold cross validation while 20% as a test set. This split's results will subsequently be used in the airline's customer satisfaction classification system. The classification algorithms utilized in this study are Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest.

The prediction procedure is then carried out by predicting the completed model from each algorithm created with testing data based on training data. The prediction's outcome is generated and shown in a confusion matrix and ROC curve. Also, the most contributing feature variables will be identified from running the feature importance function. Finally, the mean performance of cross-validation graphs of each algorithm, which contains the AUC, precision, and recall values, will be compared in order to find the best and most accurate classification method.

IV. Result and Analysis

A. *Exploratory Data Analysis*

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 69066 entries, 0 to 69065
Data columns (total 22 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Satisfaction                       69066 non-null  int64
 1   Gender                             69066 non-null  object
 2   Customer Type                      69066 non-null  object
 3   Age                                69066 non-null  int64
 4   Type Of Travel                     69066 non-null  object
 5   Class                              69066 non-null  object
 6   Flight Distance                    69066 non-null  int64
 7   Inflight Wifi Service              69066 non-null  int64
 8   Departure/Arrival Time Convenience 69066 non-null  int64
 9   Ease Of Online Booking             69066 non-null  int64
 10  Gate Location                      69066 non-null  int64
 11  Food And Drink                     69066 non-null  int64
 12  Online Boarding                    69066 non-null  int64
 13  Seat Comfort                       69066 non-null  int64
 14  Inflight Entertainment             69066 non-null  int64
 15  On-board Service                   69066 non-null  int64
 16  Leg Room                           69066 non-null  int64
 17  Baggage Handling                   69066 non-null  int64
 18  Checkin Service                    69066 non-null  int64
 19  Inflight Service                   69066 non-null  int64
 20  Cleanliness                        69066 non-null  int64
 21  Total Delay                        69066 non-null  float64
```

Fig 2. Structure of The Data

The figure above displays the structure of the current data being analyzed which contains approximately 70,000 observations and 22 variables after omitting the NAs and unnecessary variables.

1) Visualization of Categorical Variables Using Barplot

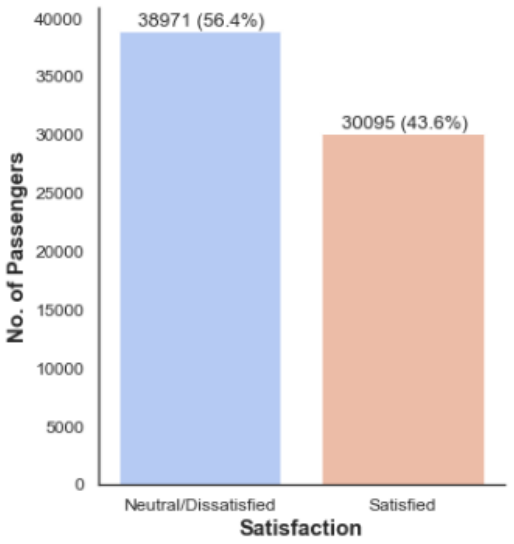


Fig 3. Visualization of Satisfaction

Based on the Satisfaction barplot above, the number of neutral/dissatisfied customers are more than satisfied customers. Therefore, it can be concluded that the research subjects are dominated by customers neutral/dissatisfied with the airline services.

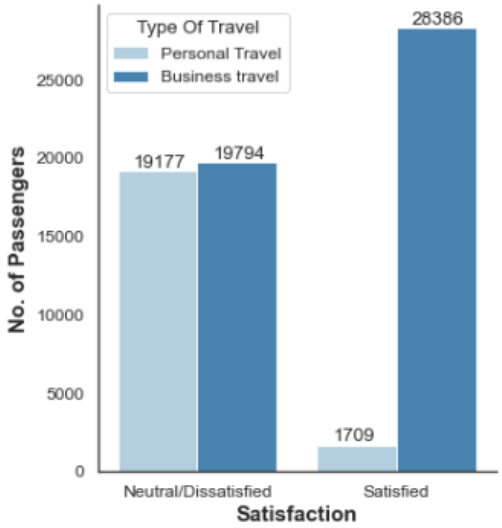


Fig 4. Tabulation and Visualization of Satisfaction by Travel Type

Based on the Satisfaction by Customer Type barplot above, the customers are mostly satisfied and there are more business travel types than personal travel in general. Therefore, it can be concluded that our research subjects are dominated by business travel type customers satisfied with the airline services.

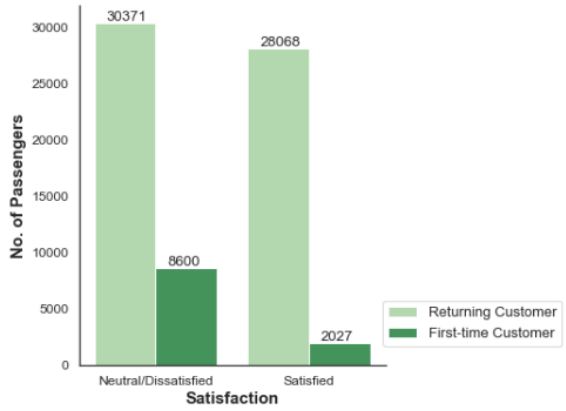


Fig 5. Tabulation and Visualization of Satisfaction by Customer Type

Based on the Satisfaction by Customer Type barplot above, the customers are mostly dissatisfied and there are more returning customers than first-time customers in general. Therefore, it can be concluded that our research subjects are dominated by returning customers dissatisfied with the airline services.

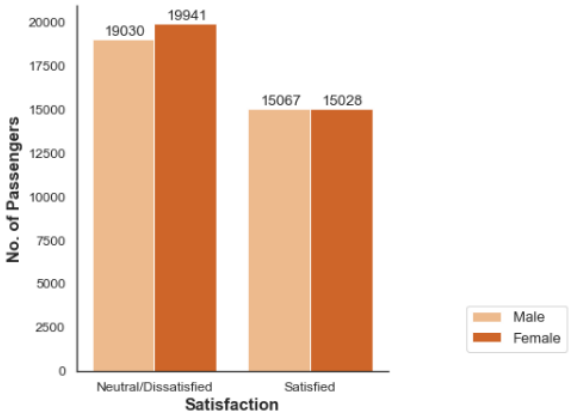


Fig 6. Tabulation and Visualization of Satisfaction by Gender

Based on the Satisfaction by Gender barplot above, the frequency of satisfied women is less than men, and the frequency of dissatisfied women is more than men. Therefore, it can be concluded that our research subjects are dominated by women being mostly dissatisfied with the airline services, while men mostly satisfied.

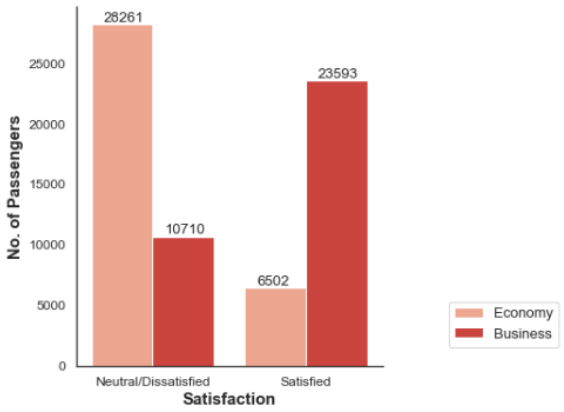


Fig 7. Tabulation and Visualization of Satisfaction by Travel Class

Based on the Satisfaction by Travel Class barplot above, the business class type are mostly satisfied and the economy class type are mostly dissatisfied. Therefore, it can be concluded that our research subjects are dominated by business-class customers satisfied with the airline services.

2) Visualization of Numeric and Categorical Variables Using Boxplot

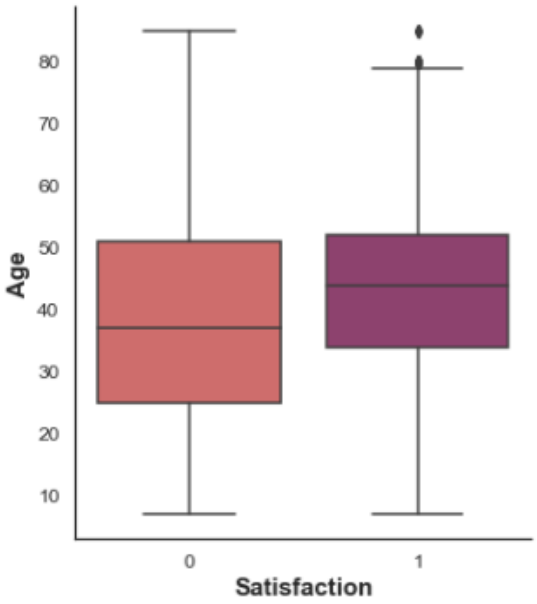


Fig 8. Visualization of Satisfaction by Age

Based on the Satisfaction by Age boxplot above, the average age of satisfied customers is higher than the average of dissatisfied customers. This means the satisfied customers are older on average than the dissatisfied customers.

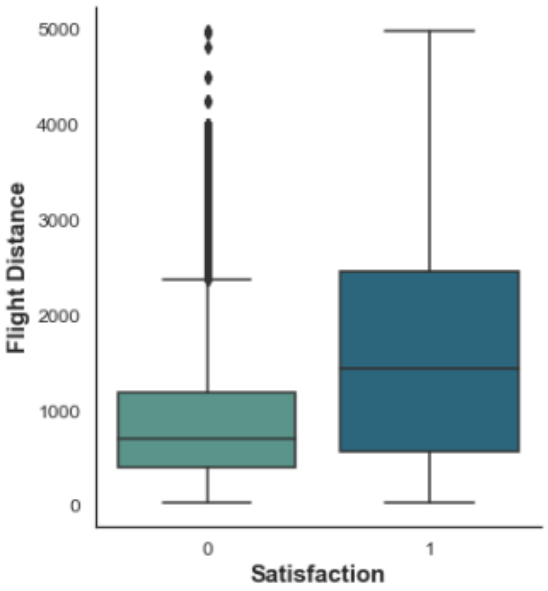


Fig 9. Visualization of Satisfaction by Flight Distance

Based on the Satisfaction by Flight Distance boxplot above, the average distance of satisfied customers is higher than the dissatisfied customers. This means the flight distance of the satisfied customer is further than the dissatisfied customers.

B. Feature Selection

1) KDE Plots

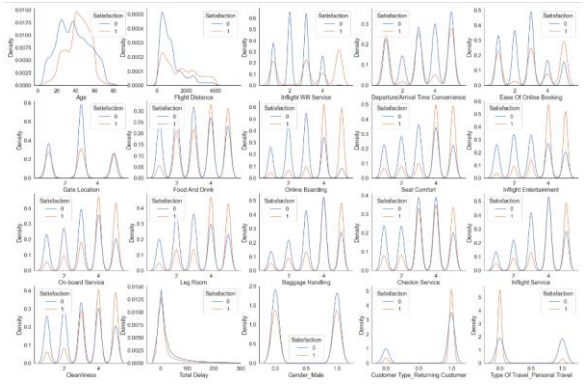


Fig 10. KDE Plots of Features

KDE is used to make visualizations of continuous and non-parametric data variables that show the probability density. The 'Gate Location' feature appears to be lacking '2' and '4' points, indicating an inconsistency as passengers are unlikely to achieve this score. The distribution of satisfaction in the 'Gender' feature is nearly the same for both, suggesting that it is negatively aligned with goals and has therefore been omitted.

2) Correlation Matrix and Heatmap

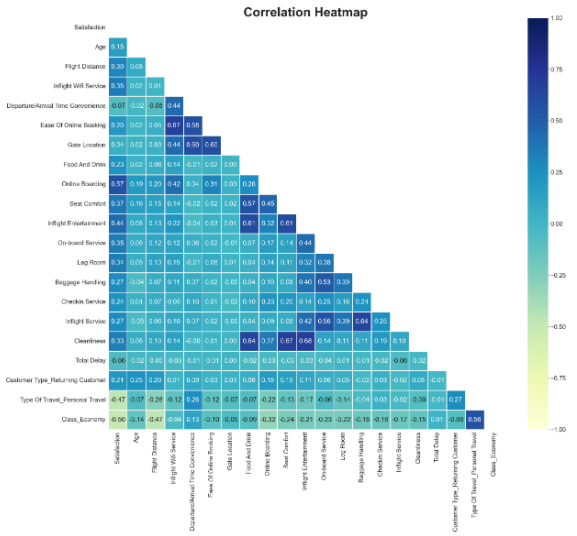


Fig 11. Correlation Heat Map

The heatmap above shows the relationship or correlation between the variables in the data. Darker colors and values close to positive 1 indicate a strong positive or unidirectional relationship between these variables. Meanwhile, a lighter color and a value close to negative 1, indicates a strong negative or opposite relationship between these variables. Features 'Age,' 'Departure / Arrival Time Convenience,' 'Gate Location,' and 'Total Delay' have poor correlation with target variable of 0.15 and below.

3) LASSO Path

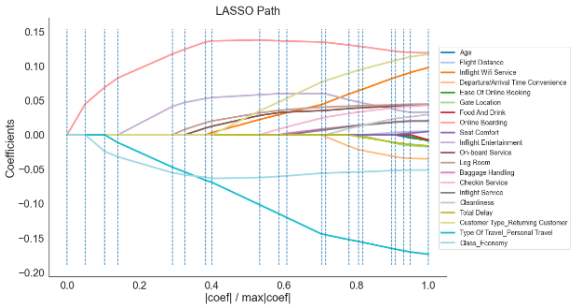


Fig 12. Plot of LASSO Regression

Lasso is used for performing variable selection by using a technique that regularizes and shrinks the coefficient estimates towards zero. When the hyper alpha parameter increases, the least significant feature has a linear coefficient that decreases to zero the fastest. Based on the plot, the features/variables are: 'Food and Drink', 'Ease of Online Booking', 'Age', 'Flight Distance', 'Total Delay', and 'Gate Location'.

C. Modeling

1) Logistic Regression Model

```
AUC Score of Logistic Regression Model is :  95.75%
Precision of Logistic Regression Model is :  88.26%
Recall of Logistic Regression Model is :  86.89%
```

Fig 13. Logistic Regression Model Performance Scores

Based on the final result of model performance, the model with the Logistic Regression model has an overall AUC score of 95.75%, precision value of 88.26%, and recall value of 86.89%.

2) K-Nearest Neighbors Model

```
AUC Score of K-Nearest Neighbors Model is :  97.89%
Precision of K-Nearest Neighbors Model is :  95.39%
Recall of K-Nearest Neighbors Model is :  90.21%
```

Fig 14. K-Nearest Neighbors Model Performance Scores

Based on the final result of model performance, the model with the K-Nearest Neighbors model has an overall AUC score of 97.89%, precision value of 95.39%, and recall value of 90.21%.

3) Naive Bayes Model

```
AUC Score of Naive Bayes Model is :  94.72%
Precision of Naive Bayes Model is :  89.83%
Recall of Naive Bayes Model is :  84.05%
```

Fig 15. Naive Bayes Model Performance Scores

Based on the final result of model performance, the model with the Naive Bayes model has an overall AUC score of 94.72%, precision value of 89.83%, and recall value of 84.05%.

4) Decision Tree Model

```
AUC Score of Decision Tree Model is :  97.95%
Precision of Decision Tree Model is :  96.30%
Recall of Decision Tree Model is :  92.71%
```

Fig 16. Decision Tree Model Performance Scores

Based on the final result of model performance, the model with the Decision Tree Model model has an overall AUC score of 97.95%, precision value of 96.30%, and recall value of 92.71%.

5) Random Forest Model

```
AUC Score of Random Forest Model is :  99.29%
Precision of Random Forest Model is :  97.28%
Recall of Random Forest Model is :  93.58%
```

Fig 17. Random Forest Model Performance Scores

Based on the final result of model performance, the model with the Random Forest model has an overall AUC score of 99.29%, precision value of 97.28%, and recall value of 93.58%.

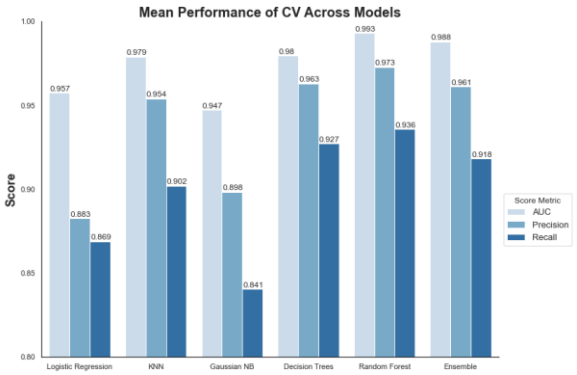
6) Ensemble Model

```
AUC Score of Ensemble Model is :  98.79%
Precision of Ensemble Model is :  96.11%
Recall of Ensemble Model is :  91.82%
```

Fig 18. Ensemble Model Performance Scores

Based on the final result of model performance, the model with the Ensemble model has an overall AUC score of 98.79%, precision value of 96.11%, and recall value of 91.82%.

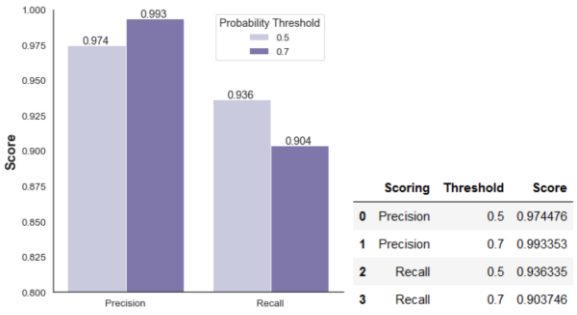
D. Model Comparison



	Model	Scoring	Score
0	Logistic Regression	AUC	0.957475
1	Logistic Regression	Precision	0.882574
2	Logistic Regression	Recall	0.868921
3	KNN	AUC	0.978937
4	KNN	Precision	0.953942
5	KNN	Recall	0.902068
6	Gaussian NB	AUC	0.947164
7	Gaussian NB	Precision	0.898347
8	Gaussian NB	Recall	0.840534
9	Decision Trees	AUC	0.979510
10	Decision Trees	Precision	0.963001
11	Decision Trees	Recall	0.927053
12	Random Forest	AUC	0.992917
13	Random Forest	Precision	0.972766
14	Random Forest	Recall	0.935849
15	Ensemble	AUC	0.987866
16	Ensemble	Precision	0.961116
17	Ensemble	Recall	0.918154

Fig 19. Model Comparison Based on Mean Performance

By running the GridSearchCV algorithm on Scikit-Learn, the optimal model and hyper parameters are: k-Nearest (k = 7), Logistics Regression (C = 0.04), Decision Tree (Max Depth= 12), and Random Forest (Max Depth= 17). Based on the Model Comparison of Mean Performance graph, Random Forest is the best algorithm with AUC 0.99, Precision 0.97, and Recall 0.94.



	Scoring	Threshold	Score
0	Precision	0.5	0.974476
1	Precision	0.7	0.993353
2	Recall	0.5	0.936335
3	Recall	0.7	0.903746

Fig 20. Random Forest Model Comparison with Different Threshold

High precision will be more relevant for this business problem. The estimation of the positive class model, 'Satisfied,' must be very accurate in order to accurately determine the critical factors that contribute to customer satisfaction.

The probability default threshold of Random Forest is 0.5. After changing to 0.7, the precision increased from 0.97 to 0.99. Since this is consistent and in accordance with the model's objectives, Random Forest (Max Depth = 17) with a threshold of 0.7 becomes our final model.

E. Model Evaluation

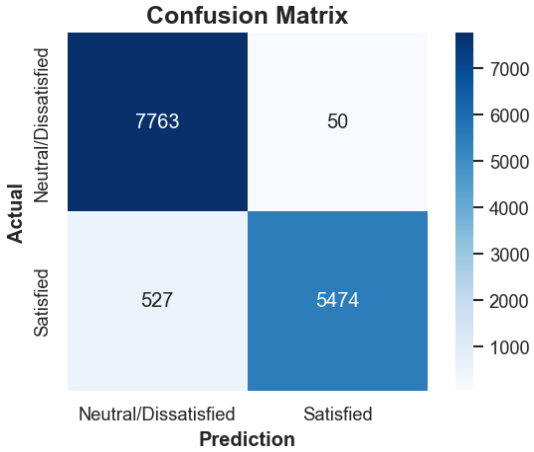


Fig 21. Confusion Matrix of Random Forest Model

In Confusion Matrix of Random Forest model above, there are 5474 passengers who are predicted as satisfied and are actually satisfied (the true positive); 50 passengers who are predicted as satisfied and are actually neutral/dissatisfied (the false positive); 7763 passengers who are predicted as dissatisfied and are actually neutral/dissatisfied (the true negative); and 527 passengers who are predicted as neutral/dissatisfied and are actually satisfied (the false negative).

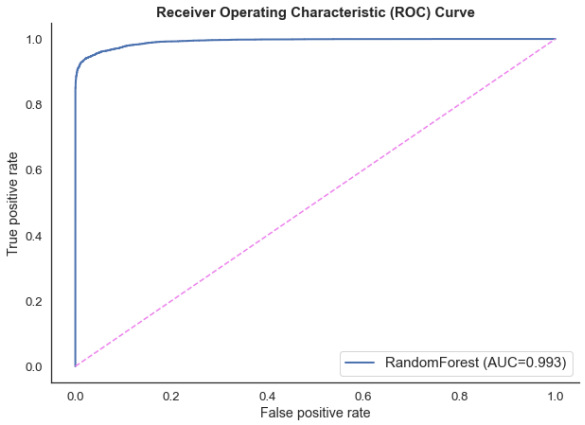


Fig 22. ROC Curve of Random Forest Model

Based on the results of the ROC curve plot above, the Random Forest model has an AUC score of 0.993. This value ranges between 0.9 and 0.1, which means that this model is classified as a model with excellent performance at predictive task.

```
Precision Score of The Best Model (Random Forest) is :  99.09%
Recall Score of The Best Model (Random Forest) is :  91.22%
F1-Score of The Best Model (Random Forest) is :  94.99%
AUC Score of The Best Model (Random Forest) is :  99.30%
```

Fig 23. Final Performance of Random Forest Model

Thus, it can also be concluded from both confusion matrix and ROC curve that the final performance result of Random Forest model are precision score of 99.09%, recall score of 91.22%, f1-score of 94.99% and AUC score of 99.30%.

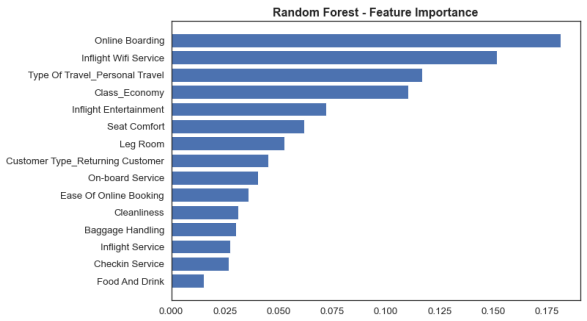


Fig 24. Feature Importance Plot of Random Forest Model

Since using the Random Forest algorithm, we can use functions to calculate the most important variables using feature importance function. Based on the plot above, the variables Online Boarding, Inflight Wifi Service, and Type Of Travel_Personal Travel are the 3 most important variables.

F. Model Simulation

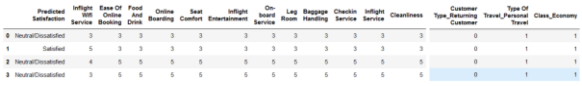
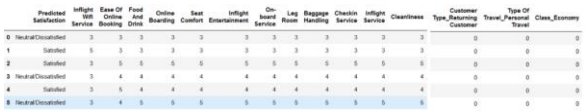


Fig 25. Simulation of First-Time Customer Satisfaction Personal Travel in Economy Class

When we started allocating all categories to the average ranking (rating: 3) for economy customers on a personal trip, the model was uncertain as to whether the consumer would be satisfied. The model predicts that the consumer will be satisfied if we boost the level of In-Flight Wifi Connectivity to excellent (rating: 5), with other groups performing averagely. Even though the In-Flight WiFi Facility is degraded when the rest of the tier is set to extremely fine, the model predicts that the customer will not be satisfied.



	Predicted Satisfaction	Inflight Wifi Service	Ease Of Online Booking	Food And Drink	Online Boarding	Seat Comfort	Inflight Entertainment	On-board Service	Leg Room	Baggage Handling	Checkin Service	Inflight Service	Cleanliness	Customer Type_Returning Customer	Type Of Travel_Personal Travel	Class_Economy
0	Neutral/Dissatisfied	3	3	3	3	3	3	3	3	3	3	3	3	0	0	0
1	Satisfied	5	3	3	3	3	3	3	3	3	3	3	3	0	0	0
2	Satisfied	3	5	5	5	5	5	5	5	5	5	5	5	0	0	0
3	Neutral/Dissatisfied	3	4	4	4	4	4	4	4	4	4	4	4	0	0	0
4	Satisfied	3	5	4	4	4	4	4	4	4	4	4	4	0	0	0
5	Neutral/Dissatisfied	3	4	5	5	5	5	5	5	5	5	5	5	0	0	0

Fig 26. Simulation of First-Time Customer Satisfaction Business Travel in Business Class

The model implies business travelers will be more easily satisfied. This model predicts that business clients will be satisfied despite the lower In-Flight WiFi Coverage level compared to the rest of the segment, which is rated as very good. However, as we proceeded to downgrade in other categories, this model predicts that consumers would be pleased the ranking for Ease of Online Shopping is set to at least very good (rating:5)

V. CONCLUSIONS

For airlines, we developed a very accurate classification model to assist them in identifying important constraints and enhancing passenger satisfaction. Classification of a target variable using few different algorithms will generate varying levels of accuracy, precision, recall, sensitivity, and specificity. In this study, several classification algorithms are used to build predictive models such as: Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest. An algorithm is deemed to be good at classifying if its accuracy rate exceeds 50%. However, in order to avoid misclassification, the established model must fulfill the requirements of having a high degree of AUC score, precision, and recall values when choosing the optimal method for the classification function.

Based on the findings of several mentioned classification algorithms, it can be inferred that the Random Forest algorithm with 0.7 threshold is the best predictive/classifier model in this research. This result is based on its greater AUC score, precision, and recall values than the rest. Overall, the Random Forest model with 0.7 threshold condition has a precision of 99.3% and recall of 90.4%. As the result, for datasets and target variables with the same or similar specifications as this study, the author proposes the Random Forest algorithm as the best and most suitable option for building future prediction/classification model.

Furthermore, based on numerous simulations and feature importance identification, we advise that airlines prioritize optimizing the In-Flight Wi-Fi Service experience. In order to encourage more economy class passengers to utilize in-flight Wi-Fi, airlines may, for instance, provide new tools to make accessing Wi-Fi easier or reduce the cost of accessing Wi-Fi in order to increase the number of passengers that utilize the service. In addition, airlines must focus on the Ease of Online Booking, as business passengers value flexibility and comfort during flights.

ACKNOWLEDGMENT

The author would like to thank Mr. Rudi Sutomo, S.Kom., M.Si., M.Kom. as the supervisor for the Data Modelling course, Information Systems Study Program, Multimedia Nusantara University, who has shared all the knowledge and insights that are very useful for the continuity of this research. The author would also like to thank Mr. Rudi Sutomo, who has been willing to provide consultation, input, direction, and guidance since this research was planned.

REFERENCES

[1] H. Y. Khudhair, A. Jusoh, A. Mardani and K. M. Nor, "Quality Seekers as Moderating Effects between Service Quality and Customer Satisfaction in Airline Industry," *International Review of Management and Marketing*, vol. 9, no. 4, pp. 74-79, 2019.

[2] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," *Journal of Big Data*, vol. 6, no. 62, pp. 1-16, 2019.

[3] B. H. Hayadi, J.-M. Kim, K. Hulliyah and H. T. Sukmana, "Predicting Airline Passengeer Satisfaction with Classification Algorithms," *International Journal of Informatics and Information System*, vol. 4, no. 1, pp. 82-94, 2021.

[4] E. Park, Y. Jang, J. Kim, N. J. Jeong, K. Bae and A. P. d. Pobil, "Determinants of customer satisfaction with airline services: An analysis of customer feedback big data," *Journal of Retailing and Consumer Services*, vol. 51, pp. 186-190, 2019.

[5] S. Tsafarakis, T. Kokotas and A. Pantouvakis, "A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement," *Journal of Air*

Transport Management, vol. 68, pp. 61-75, 2018.

[6] R. Hussain, "The Mediating role of customer satisfaction: evidence from the airline industry," *Asia Pasific Journal of Marketing and Logistic*, vol. 28, no. 2, pp. 234-255, 2016.

[7] M. K. Koklic, M. Kukar-Kinney and S. Vegelj, "An investigation of customer satisfaction with low-cost and full-service airline companies," *Journal of Business Research*, vol. 80, pp. 188-196, 2017.

[8] F. R. Lucini, L. M. Tonetto, F. S. Fogliatto and M. J. Anzanello, "Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews," *Journal of Air Transport Management*, vol. 83, 2020.

[9] S. Suresh, T. G. Balachandran and S. Sendilvelan, "Empirical Investigation of Airline Service Quality and Passenger Satisfaction in India," *International Journal of Performability Engineering*, vol. 13, no. 2, pp. 109-118, 2017.

[10] S. Tahanisaz and S. Shokouhyar, "Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry," *Journal of Air Transport Management*, vol. 83, no. 2, 2020.

[11] N. Madaan, U. Kumar and S. K. Jha, "Big Data Analytics: A Literature Review Paper," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT)*, vol. 8, no. 10, 2020.

[12] M. A. Jassim and S. N. Abdulwahid, "Data Mining preparation: Process, Techniques and Major Issues in Data Analysis," *IOP Conf. Series: Materials Science and Engineering*, vol. 1090, no. 1, 2021.

[13] K. Sumathi, S. Kannan and K. Nagarajan, "Data Mining: Analysis of student database using Classification Techniques," *International Journal of Computer Applications*, vol. 141, no. 8, pp. 22-27, 2016.

[14] A. R. Sreenivasa, A. V. Ramana and S. Ramakrishna, "Implementing the Data Mining Approaches to Classify the Images with Visual Words," *International Journal of Recent Technology and Engineering*, vol. 7, no. 6, pp. 901-909, 2019.

[15] T. Jiang, J. L. Gradus and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behavior Therapy*, vol. 51, no. 5, pp. 675-687, 2020.

[16] B. Mahesh, "Machine Learning Algorithms -A Review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381-386, 2020.

[17] U. Salamah and D. Ramayanti, "Implementation of Logistic Regression Algorithm for Complaint Text Classification in Indonesian Ministry of Marine and Fisheries," *International Journal of Computer Techniques*, vol. 5, no. 5, pp. 74-78, 2018.

[18] E. u. Hassan, Z. Zainuddin and S. Nordin, "A Review of Financial Distress Prediction Models: Logistic Regression and Multivariate Discriminant Analysis," *Indian-Pacific Journal of Accounting and Finance (IPJAF)*, vol. 1, no. 3, pp. 13-23, 2017.

[19] O. M. I. Gazalba and N. G. I. Reza, "Comparative Analysis of K-Nearest Neighbor and Modified K-Nearest Neighbor Algorithm for Data Classification," *International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, vol. 2, pp. 294-298, 2017.

[20] A. M. S. I. Dewi and I. B. G. Dwidasmara, "Implementation Of The K-Nearest Neighbor (KNN) Algorithm For Classification Of Obesity Levels," *Jurnal Elektronik Ilmu Komputer Udayana*, vol. 9, no. 2, pp. 277-284, 2020.

[21] K. U. Syaliman, E. N. and O. S. Sitompul, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *Journal of Physics Conference Series*, vol. 978, no. 1, 2018.

[22] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," *International Journal of Applied Science and Technology*, vol. 7, no. 2, pp. 5-18, 2017.

[23] A. Ghazzawi and B. Alharbi, "Analysis of Customer Complaints Data using Data Mining Techniques," *Procedia Computer Science*, vol. 163, pp. 62-69, 2019.

[24] A. Wibawa, D. M. P. Murti, R. P. Adiperkasa and A. C. Kurniawan, "Naïve Bayes Classifier for Journal Quartile Classification," *International Journal of Recent Contributions from Engineering Science & IT*, vol. 7, no. 2, pp. 91-99, 2019.

[25] P. Kaviani and S. Dhotre, "Short Survey on Naive Bayes Algorithm," *International Journal of Advance Engineering and Research Development*, vol. 4, no. 11, pp. 607-611, 2017.

[26] K. Mittal, D. Khanduja and P. C. Tewari, "An Insight into “Decision Tree Analysis”," *World Wide Journal of Multidisciplinary Research and Development*, vol. 3, no. 12, pp. 111-115, 2017.

[27] B. T. Jijo and A. M. Abdulazeez, "Classification Based on Decision Tree

Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 1, pp. 20-28, 2021.

[28] J. "IMPLEMENTASI ALGORITMA RANDOM FOREST UNTUK KLASIFIKASI KATEGORI BERITA," Tangerang, 2021.

[29] A. D. Kulkarni and B. Lowe, "Random Forest Algorithm for Land Cover Classification," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 4, no. 3, pp. 58-63, 2016.