CHAPTER 3
RESEARCH METHODOLOGY

The first step we did was search the title topic; the title topic search was done based on a collection of journal references that we had previously searched for then from there I could draw conclusions to find a title topic that had never existed before using the algorithm I used.

The second step is to find a data set that is suitable for my topic, after doing various research, finally found the appropriate and suitable research data, namely the data set in Kaggle with 11 thousand data with six variable values.

Then the third step is to try to enter the data set that has been taken and has been adjusted to the needs of entering into each algorithm with the way each algorithm works, now when you have successfully entered each algorithm, you can say it has been successful. but what we are looking for is the value of TP, FP, TN, FN to get a score of accuracy, precision, recall, F-1 score.

Then the next step after getting the previous score from each algorithm, what is done is to carry out the implementation of the score that has been searched for is input into the graph so that the process is more visible so that you know which algorithm is better and worse.

Then for clearer steps, the explanation is below in more detail

1. Identification and Literature Study

The first stage in the research is the identification stage. The identification stage is the stage where the author searches, finds, collects, examines more deeply for each existing case study. In the literature search, the writer often finds Indonesian language literature because there is very little to find literature in English. To find literature using the internet with the URL address https://scholar.google.com/ researchers collect at least 10 pieces of literature as reference topics in the case study. From the journals that the authors found, most of the journals were from the 2015 to 2020 publication years. The steps in this Literature collection were intended as reference materials and sources of information in solving a problem at hand.

2. Data and Variables

The data and variables used are chili producer data, namely 129,881 data. Dataset taken from Kaggle's source. The data used here are 5 attributes. The data used are Satisfaction, Age, Flight distance, Gate location, Inflight WIFI service, Inflight entertainment, Online support, Ease of Online booking, On-board service, Cleanliness, Departure Delay in Minutes, and Arrival Delay in Minutes. The existing data will be used to compare the accuracy between the SVM algorithm and the Random Forest algorithm.

3. Perform data processing

After getting data from the kaggle website, the next step is to enter the data processing step. This data processing stage is the initial stage in analyzing the results in processing a data. Data processing consists of split datasets, data normalization techniques and testing the model formed using split split validation. Here are some steps when processing data.

a. Split Dataset

In this stage, from the amount of existing data, the data will be divided into 2 parts, namely training data and testing data. For the distribution of the amount of training data and testing data with a ratio of 70% for training data and 30% for testing data. By doing a split dataset aims to facilitate the processing of very large data.

b. Data normalization

The data normalization stage is carried out by re-scaling the dataset and aims to compare data sets from various factors or using different units. This technique rescales the distribution by using the ratio of the distance of each value from the minimum value in each data set to the range of values in each data set. The normalization technique begins by performing min-max statistics for scaling where at this stage the data will be approximated with data that is scaled in a fixed range of 0 to 1. The normalization technique is different from standardization because normalization will end with a smaller standard deviation. After performing min-max for scaling, the next step is to perform feature scaling, which is the method used to normalize the range of independent variables or data features. The purpose of this

method is to scale the feature vector components such that the complete vector has a length of one.

c. Cross Validation Split

For testing the model formed from data processing, a cross validation split is carried out. Cross validation split is a method used to evaluate the performance of a model or algorithm and obtain maximum accuracy results where the data is separated into two subsets, namely learning process data and validation or evaluation data. To find out the performance of an algorithm model, it is done by experimenting with k or folds that will be formed. Cross validation split is also a validation technique of developing a split validation model where the validation measures training errors by testing with test data or test data. By testing using a cross-validation split, the results obtained can be maximized but can reach more efficient tests.

4. Making a data classification with SVM (Support Vector Machine)

Support Vector Machine (SVM) is a classification method that works by finding the hyperplane with the largest margin. In other words, SVM is a technique that uses 2 points (2 vectors), which in turn these 2 points will form a dividing line (or border if 3 dimensions or more). The boundary line/side that is formed from these two vectors is called a hyperplane. The hyperplane is a data dividing line between classes. Margin is the distance between the hyperplane and the closest data in each class. The data closest to the hyperplane in each class is called the support vector (J. Yunliang, et al., 2010). SVM is a relatively new technique (1995) for making predictions, both in the case of classification and regression, which is very popular in recent times. SVM is in the same class as Artificial Neural Network (ANN) in terms of functions and problem conditions that can be solved. Both are included in the supervised learning class, wherein its implementation there needs to be a training stage and followed by a testing stage. In SVM it takes a kernel to transform data into a higher dimensional space called kernel space which is useful for linearly separating data.

5. Making a data classification with Random Forest Classification Algorithm.

Random forest is an algorithm used to classify large amounts of data. Random forest classification is done by merging trees by conducting training on the sample data owned. The classification process in a random forest begins with breaking the existing sample data into a random decision tree. After the tree is formed, voting will be carried out on each class from the sample data. Then, combine the votes from each class and then take the most votes. By using random forest in the data classification, it will produce the best vote. Random forest is a combination of each good tree which is then combined into one model. Random Forest depends on a random vector value with the same distribution in all trees where each decision tree has a maximum depth. Random forest is an algorithm used to classify large amounts of data. Random forest is a combination of each tree (tree) from a good Decision Tree model and then combined into one model. The use of more trees will affect the accuracy that will be obtained for the better. Determination of the classification by random forest is taken based on the voting results of the formed tree.

6. Result

After going through the classification process between the two algorithms and getting an accuracy value, precision, recall and f1 score. The values of accuracy, precision, recall and f1 score are obtained from the results of the confusion matrix of the two algorithms. The confusion matrix is a comparison of the classification results carried out by the model with the actual classification results in the form of TP (True Positive), TN (True Negative), FN (False Negative) and FP (False Positive). So, the authors can conclude that each algorithm has its own advantages and disadvantages in terms of accuracy.

7. Report Writing

In this study, the authors made a report that discusses the process of making research procedures from the beginning to the end of the system testing and analysis stage. This research also discusses the comparison between the SVM algorithm and the Random Forest in the classification. In making this report, the authors added suggestions for further research.