CHAPTER 3
RESEARCH METHODOLOGY

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 literatures as a reference topic in the case study. From the journals that the authors found, most of the journals were from the 2016 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 main data source used in this study was student dummy data totaling 372 data. There are 9 attributes in this study. The dataset used is in the form of GPA (Achievement Index) semesters 1 to 8 and student graduation status in the form of graduating on time and graduating late. From existing data, the authors will compare the accuracy of the K-NN algorithm and the Random Forest algorithm.

3. KNN (K-Nearest Neighbor) Classification Algorithm

The KNN algorithm is one of the classification methods that applies a supervised algorithm (Han, 2006) where the results of the new test sample are classified based on the majority of the categories on the K-NN. This algorithm is done by looking for groups of k objects in the training data that are closest

(similar) to objects in new data or testing data (Leidiyana, 2010) [11]. At the classification stage the KNN Algorithm uses several stages, namely split data, determining the k value, calculating the distance between the data to be evaluated and all training data, sorting the distance formed, pairing the appropriate class and calculating the accuracy value.

a. Add new dataset

At this stage, the author can determine whether the dataset of 372 data will be added or not. If you want to add a new dataset by inputting a file that can contain more or less than 372 data. So if at first the dataset is 372 data, if you want to add a new dataset of a thousand more or less by reading and loading new datasets from other files until we input n which means no or stop reading the first file until the last file we need.

b. Split Data

At this stage, the data that the author has as many as 372 data will be divided into training data and testing data. The amount of training data and testing data is determined by the probability or likelihood of a split value occurring. So for determining the value of the split is gambling or the possibility that happens. For example, we have a split value of 0.6 or 60% with 372 data. From this example the testing value has the possibility of approaching 0.6 or away from 0.6. This training value will not be much different from the split data value. From this example, the training value could be 228 and the testing value could be 144.

c. Determine the Value of K

In this step, determine the value of K which aims to find the nearest neighbor. For example, we have 100 pieces of data, then we will determine the value of k, for example 5, which means that we will find the distance between data 1 to 5. To get the optimal k value, we must find the accuracy of the k value that we specify. A value of k cannot only depend

on one nearest neighbor because it will make the classification result rigid. On the other hand, if the value of k increases, it will result in vague results. The value of k cannot exceed the amount of data being trained.

d. Calculate the distance between the evaluated data and the training data

Euclidean distance is a calculation of the distance between 2 points in euclidean space. Euclidean distance is applied to various dimensions. In this step, calculate the distance between the evaluated data and the training data by means of the Euclidean distance formula. The Euclidean distance formula is as follows:

$$d(p,q)= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p,q = Two points in the Euclidean room

qi,pi = Euclidean vector, starting from space origin (starting point)

n = space-n

e. Sort the distances formed

After getting the distance between the data, the next step is to sort the distance formed and determine the closest distance to the k value based on the smallest to largest value.

f. Pair the appropriate class

At this stage, it discusses how to predict the real student graduation status from the original data. Whether the student's graduation status is on time or late in the original data according to the predictions made. If appropriate, add 1 otherwise remain 1.

g. Predictions

From this stage the writer will start predicting using data testing, and start calculating the closest distance from the training data and data testing and sorting the predictions out of student graduation status.

h. Calculates the value of accuracy, precision, recall and f1 score

After getting the prediction of the student's graduation status, the next step is to calculate the accuracy of the predictions we have. To get the accuracy value obtained from the number of predictions in accordance with the original data divided by the number of test data then multiplied by 100%. For example the prediction of passing status on time based on original data in the form of on time or called tp (true positive), the amount of data that results from the prediction of late passing status according to the original data in the form of being late is called tn (true negative). The prediction is correct but the original data is late is called fp (false positive) and the amount of data from the late prediction but the real data is on time is called fn (false negative). So it can be concluded that the accuracy value is ((tp + tn * 100%) / (tp + tn + fn + fp)) In addition to calculating the accuracy value, the writer also calculates the precision, recall and f1 score. The formula for precision is (tp * 100%) / (tp + fp). And for the recall or sensitivity formula, namely (tp * 100%) / (tp + fn). And the formula for the f1 score is (2 * precision * recall) / (precision + recall). F1 score is the average comparison between precision and recall formed.

4. Random Forest Classification Algorithm

The Random Forest algorithm is a method of classifying data that is determined by the voting results of the formed tree. Voting results are determined by the largest number of votes. In the process of classification the random forest algorithm will be successful if all trees have been formed. The Random Forest algorithm classification process consists of dividing data into k folds and based on attribute values, calculating the Gini index of the data displayed, choosing the best split point from the dataset, creating terminal nodes, creating child splits for nodes or creating new terminals, making a decision tree, makes predictions with a decision tree, creates a random subsample of the dataset, makes predictions with bagged tree lists and calculates accuracy and scores. The following are the steps for the classification of the Random Forest algorithm:

a. Add new dataset

At this stage the same as the basic step of adding a new dataset, the author can determine whether a dataset of 372 data will be added or not. If you want to add a new dataset by inputting a file that can contain more or less than 372 data. So if at first the dataset is 372 data, if you want to add a new dataset of a thousand more or less by reading and loading new datasets from other files until we input n which means no or stop reading the first file until the last file we need.

b. Split the dataset into k folds and based on attribute values

At this stage the dataset will be divided into as many as k, then the k partition data as testing data and training data. At this stage the aim is to validate the accuracy of the model built on a certain dataset. Then divide the training data based on attribute values.

c. Calculates the Gini index

Gini index is a method for determining the level of inequality that occurs. If gini is 0 it indicates perfect equality. If Gini is 1 then inequality

occurs. At this stage it aims to measure the probability of certain classified variables when randomly selected. Here's the Gini index formula:

$$\text{Gini} = 1 - \sum_{i=1}^{n} P_i(y_i + y_i - 1)$$

P*i* = variable mean / score

n = number of observations

F*i* = income value

Gini= Gini Coefficient

d. Choose the best split point

The next step is to choose the best split point from the Gini dataset calculation or the majority voting of each of the best values that the split point has. Majority voting here is meant to take the best value from each tree that is formed. At this stage using feature selection that aims to minimize errors that appear.

e. Create a terminal node

The next step is to create a terminal node for the tree branch. Terminal node is a node that is not split. At this stage it contains the results of the nodes that are not displaced.

f. Create a child split for the nodes

The next step is to create a leaf node or child split or terminal node based on the depth of branching. At this stage the looping of nodes is repeated until the node condition cannot be displaced. At this stage, it will check that the depth is greater than the maximum depth, then it will create a new terminal node.

g. Make a decision tree

Next is making a decision tree as the right decision making. At this stage, create a root that comes from the best split value that appears. Root here represents the entire sample or population.

h. Make predictions with a decision tree

After creating a decision tree, the next step is to make a decision tree prediction based on the rows of nodes in the index. If the row variable in the index node is less than the node value and if isinstance is the left node in the dictionary it returns predict in the left node in the row variable and if not, returns the left node variable. And another if the instance is in the left node in the dictionary and returns the prediction variable on the right node in the row and otherwise returns to the right node.

i. Generates a random subsample from the replaced dataset

At this stage, the author creates a random subsample of the dataset by randomly selecting rows from the dataset and adding a new list. This stage can be repeated for a fixed number of rows or until the size of the new dataset matches the size ratio of the original dataset.

j. Make predictions with a list of bagged trees

At this stage the authors make predictions based on a list of bagged trees based on the existing rows and trees. Predictions are taken from predictive decision trees that are formed from looping trees inside the trees

k. Random Forest Algorithm

In this step we run the function from sample, create_tree and predict from bagging predict. At this stage using iteration to get sample results, trees and predictions from bagging_predict.

k. Calculates the value of accuracy, precision, recall and f1 score

After determining the prediction with sample data, the next step is to calculate the score of the trees that appear, accuracy, precision, gain, f1 score. The score is obtained from the number of correct predictions with the original data and the number of n_folds values and multiplied by 100%. The n_folds value here is the number of folds in the amount of data. To calculate the value of accuracy, precision, recall and f1 score using the same formula as calculating the value of value, precision, recall and f1 score in the KNN algorithm.

5. Result

After going through the classification process between the two algorithms and getting an accuracy value. So the authors can conclude that each algorithm has its own advantages and disadvantages in terms of accuracy.

6.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 KNN algorithm and the Random Forest in the classification of student data. In making this report, the authors added suggestions for further research.