CHAPTER 1
INTRODUCTION

1.1 Background

In classifying data, a complete and comprehensive data collection is required. Classification is the process of analyzing data and producing models that can predict the future. Clustering is a data analysis method that functions to group data with the same characteristics into the same area and data with different characteristics to other regions. Classification aims to group data and estimate the possibilities that occur based on the same class and different classes of an object. Classification is very important in predicting the possibilities that arise so that the data used can be developed and utilized in processing very large amounts of data. The large use of data can produce mixed and varied results. Classification can be based on a value, group or other things. Algorithms that can be used in data classification techniques are naive bayes, k-nearest neighbor (KNN), support vector machine (SVM), decision tree, random forest and so on. Each algorithm has advantages and disadvantages as well as a different level of accuracy.

The problem is how to share the data we have which is the benchmark in data classification. The large amount of data used will affect the results of the data classification. Another problem is the method we use in classifying the data itself. The method used can have a positive or negative impact on the classification itself. Classification also requires objects to be predicted such as predicting a value from a certain course. Another problem that arises is how to compare the algorithms used with one another.

Based on the above statement, the authors apply the classification of student data based on their graduation status and grades from semester 1 to semester 8 to find better accuracy values than the algorithm used. The algorithms used in this study are the Knn (K-Nearest Neighbor) algorithm and the Random Forest algorithm. The method used in the Knn algorithm is the brute force search method which aims to find the best k value from each k value that appears. In addition, the KNN algorithm also divides data based on the amount of training data and testing data used. The KNN algorithm looks for the distance between the input data and the data tested. For testing using the amount of data that is displit. The KNN algorithm also looks for accuracy, precision, recall and f1 score.

The Random Forest algorithm is different from Knn because it forms a model for the emerging decision trees. The method used in the Random Forest Algorithm is the K Folds Cross Validation Split where the data is divided based on the folds determined for model evaluation. This algorithm also divides the data based on the training data and testing data used. For testing it uses the amount of data that is disciplined and then looks for the value of accuracy, precision, recall and f1 score. When you get the accuracy value of the two algorithms, the authors compare the results of the accuracy of the two algorithms.

1.2 Problem Formulation

The following are some of the questions that we want to prove in this project, namely:

1. How to classify data from the KNN Algorithm and the Random Forest Algorithm ?
2. How is the comparison result between the KNN algorithm and the Random Forest algorithm to predict student graduation status?

1.3 Scope

In this project, the writer uses Python version 3.0, for the data used in this research is dummy data in the form of student data, amounting to 372 data and can be added with other datasets. The author uses the process of dividing the data into training data and testing data in data classification. After dividing it into 2 parts, the next step is to calculate the accuracy, precision, recall and f1 score of the two algorithms, namely the KNN algorithm and the Random Forest algorithm. This project will also focus on the comparison of the KNN algorithm and the Random Forest algorithm in classifying student data in terms of its accuracy.

1.4 Objective

The main objective to be achieved in this project is to classify student data from the KNN algorithm and the random forest algorithm. The second objective of this project is to calculate the accuracy of the two algorithms. So that from the accuracy value obtained, it can be concluded that the comparison of the two algorithms.