CHAPTER 3
RESEARCH METHODOLOGY

3.1 Literature Journal Study

The purpose of reading journals with a similar discussion of the research to be carried out aims to be used as a reference in order to get an overview of the journals that have been read. By reading journals, you can find out and learn what is needed to make this research project. What is learned from the journal can know the basics, types of datasets and theoretical basis for calculating algorithms. The journal can be used to learn how the C4.5 algorithm works and calculation functions and to find out the types of datasets that can be solved with the algorithm. This research can be done after studying the calculation steps of the C4.5 algorithm from journals with the same discussion of the research to be carried out.

3.2 Collecting Data

The data used in this study is a dataset containing aircraft passenger satisfaction based on the facilities provided by the airline. This dataset comes from https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction since 2020. This data has a csv extension data file. The file contains test data containing nearly 26,000 passenger data. Category attributes that can be used are Inflight wifi service, departure / arrival time convienient, ease of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on-board service, leg room service, baggage handling, check-in service, inflight service and cleanliness.

3.3 Processing Data

The data obtained has an extension with the csv format. The data contains categories or attributes and subcategories or value values of the attributes. Attributes include Inflight wifi service, departure / arrival time convienient, ease of online booking, gate location, food and drink, online boarding, seat comfort, inflight entertainment, on-board service, leg room service, baggage handling, check-in service, inflight service and cleanliness. As for sub-categories or attribute value values, all attributes in this dataset have the same values, namely 0, 1, 2, 3, 4, and 5. 0 "means no value", 1 "very dissatisfied", 2 "no satisfied ", 3" sufficient ", 4" satisfied ", 5" very satisfying ". Then based on the assessment dataset, it has a decision class attribute called Satisfaction. Satisfaction has two classes of decisions, namely "Neutral or Dissatisfied" and

“Satisfied”. The data will be implemented using the C4.5 algorithm calculation. After the C4.5 algorithm program is successfully implemented, the dataset with csv extension will be called and processed using the program.

3.4 Program Implementation

This program will be made in the MYSQL programming language. The program will load the csv in the first step. After the loading is done, the program will look for the best attributes by calculating the highest gain from each attributes using the algorithm. When it is done, the program will take some of the data from a value which has the best attribute and repeat its process of looking the best attributes by calculating the best gain again. It will repeat the process until there is no more branch.

3.5 Testing

Testing is carried out when the program has been implemented. The result can be compared with the calculation of the C4.5 algorithm using the LibreOffice Spreadsheet application. The data used for testing on the LibreOffice Spreadsheet is calculated manually based on the C4.5 algorithm using the Spreadsheet application so that the data inputted is limited, containing only the top 15 data which is the data on airline passenger satisfaction from Kaggle. The purpose of the testing is to be able to perform calculations for all data contents from the dataset to be tested easily and quickly. Then after the output of the program has appeared, the decision tree image will be drawn manually based on the calculation logic of the C4.5 algorithm with the results of the table iteration of the amount of data that has been calculated from the program implementation as a result of the final analysis of the C4.5 Decision Tree algorithm.

3.6 Data Analysis

At this stage, we will discuss what will be analyzed. What is discussed in data analysis is to test how much time it takes to test the large amount of data in the implementation program. The amount of data that will be tested in the analysis is 100, 200, 500, 1000, and up to all data. So based on the amount of data tested, the purpose of this data analysis is to find out the time required for the implementation program based on the large amount of data to be tested. So that what is being analyzed is that the large amount of data being tested has different or the same results and processing time. Another thing analyzed in this stage is whether or not this algorithm determines the passenger satisfication.

3.7 Report

This research report will contain the testing of this project, abstracts and conclusions. So that this report can contain the content of the research conducted and can be used as a reference for future research. This research also contains about testing the airline passenger satisfaction dataset using the C4.5 algorithm and answering some of the questions that exist in the problem formulation.