CHAPTER 3
RESEARCH METHODOLOGY

3.1. Data Collection

Table 3.1. shows the sample dataset obtained at *https://www.kaggle.com/datasets/yasserh/walmart-dataset* and used in this study. The data used to carry out the test amounted to 6.435. There are 8 attributes in the dataset, namely Store, Date, Weekly Sales, Holiday Flag, Temperature, Fuel Prices, CPI, and Unemployment. Fuel price is the cost of fuel in that area. This value is used as the selling price, which is used as the basis for determining Weekly Sales, in which sales will be influenced by the presence or absence of a Holiday Flag. The purpose of conducting data trials is to see how Tree and the Random Forest algorithm perform calculations.

Table 3.1. Data Sample

Store	Date	Weekly Sales	Holiday Flag	Temperature	Fuel Price	CPI	Unemployment
1	5/2/2010	1643691	0	42.31	2.572	211.0964	8.106
1	12/2/2010	1641957	1	38.51	2.548	211.2422	8.106
1	19-02-2010	1611968	0	39.93	2.514	211.2891	8.106
1	26-02-2010	1409728	0	46.63	2.561	211.3196	8.106
1	5/3/2010	1554807	0	46.5	2.625	211.3501	8.106
1	12/3/2010	1439542	0	57.79	2.667	211.3806	8.106

3.2. Program

The first step is to import the CSV file into Orange. This dataset has 8 attributes namely, Store, Date, Weekly Sales, Holiday Flag, Temperature, Fuel Prices, CPI, and Unemployment. Of the 8 attributes, one is selected to be used as the target. The predetermined target is used as a reference in determining the prediction results. After that prepare a Data Sampler from the imported CSV data and arrange the data split into two parts, namely Data Training and Data Testing. The tests were carried out three times, for example, 60% Data Training and 40% Data Testing, 70% Data Training and 30% Data Testing, and 80% Data Training and 20% Data Testing.

The data that has been set, which is called the Training Data, is then processed by the Random Forest algorithm and the Tree algorithm. As for the rest of the data or what is called Data Testing, From this process, namely Data Training and Data Testing, produces predictions, where in the predictions it is seen and compared which results are better, in this case, the results that have a low value, because a low value means that it has fewer errors.

3.3. Application and Model

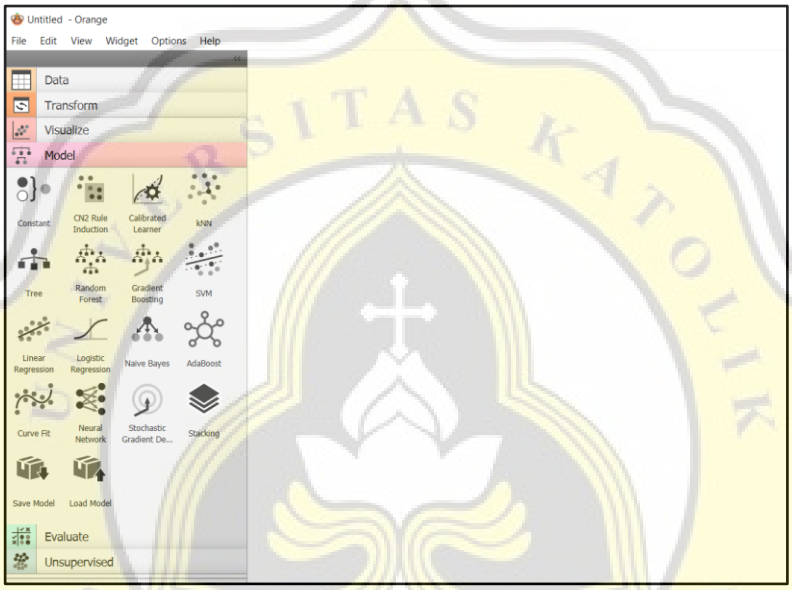

Figure 3.1 Orange Application

Data processing is done in the Orange application. The model used in testing is the Regression model. Regression is an analytical technique to identify a relationship between two or more variables. Regression aims to find a function that models data by minimizing the error or the difference between the predicted value and the actual value, wherein the Regression model the lowest value is a good value which means that the value has fewer errors. Tests were carried out using two algorithms, namely Random Forest and Tree. Random Forest is one of the algorithms used in the regression model.

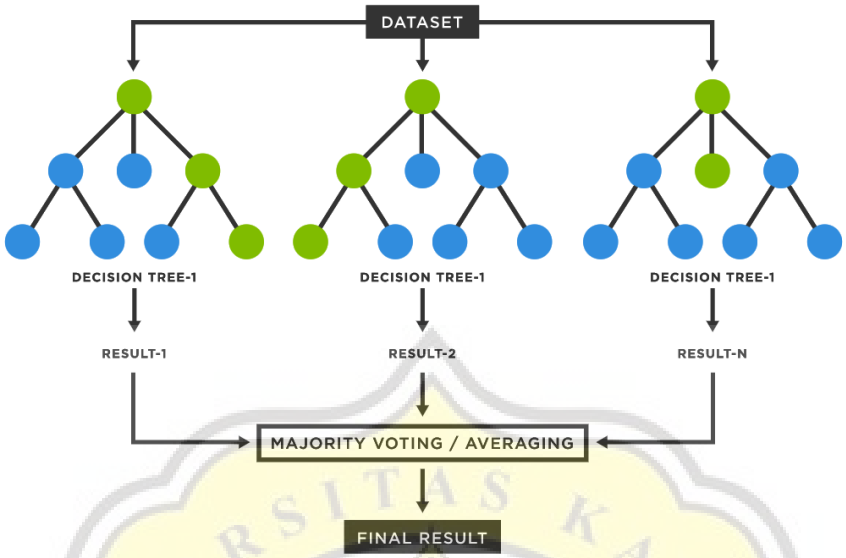

Figure 3.2 Random Forest

Random Forest works by building several decision trees and combining them to get more stable and accurate predictions. The Random Forest algorithm selects samples randomly from the dataset used, then creates a Decision Tree for each sample that has been made. The prediction results are obtained from each Decision Tree that has been made. The Decision Tree itself is a flow chart that is shaped like a tree that has a root node that is used to collect data. This algorithm was chosen because several algorithms that have been tested are no better than Random Forest and Tree. In the final test results, a comparison of each algorithm is carried out, and it is seen which algorithm has the lowest score.