CHAPTER 4
ANALYSIS AND DESIGN

4.1. Analysis

The data used in 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. 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. The data is processed using an application called Orange. The initial step is to enter the CSV file into Orange, then from this data, the target data is determined to be used as a prediction.

The next step is the Data Sampler to determine the proportion of data 60%, 70%, and 80%, the data is also called Data Training. The data is then processed using the Tree algorithm and the Random Forest algorithm from each of the predetermined data proportions. The remaining data from the Data Sampler is also called Data Testing. The results shown are then compared to which is better, Random Forest or Tree.

4.2. Random Forest and Tree

Random Forest is a machine learning algorithm used to classify large amounts of data. Random forest and Tree algorithms can be used for classification and regression problems. Random forest is also a combination of individual trees from a good decision tree model combined into a single model. Decision Tree is used to predict classes for a given data set.

The Decision Tree algorithm starts at the root node of the tree. Compare the root attribute value with the record attribute. Based on the comparison, the algorithm follows branches and proceeds to the next node. Decision Tree is a tree-like flowchart with a root node used to collect data, nodes inside the root node containing questions about the data, and leaf nodes used to solve problems and make decisions.

Decision Tree classifies data samples that do not have known classes into existing classes. Using more trees increases the accuracy you get. The classification decisions made by Random Forest are based on the voting results of the formed trees.

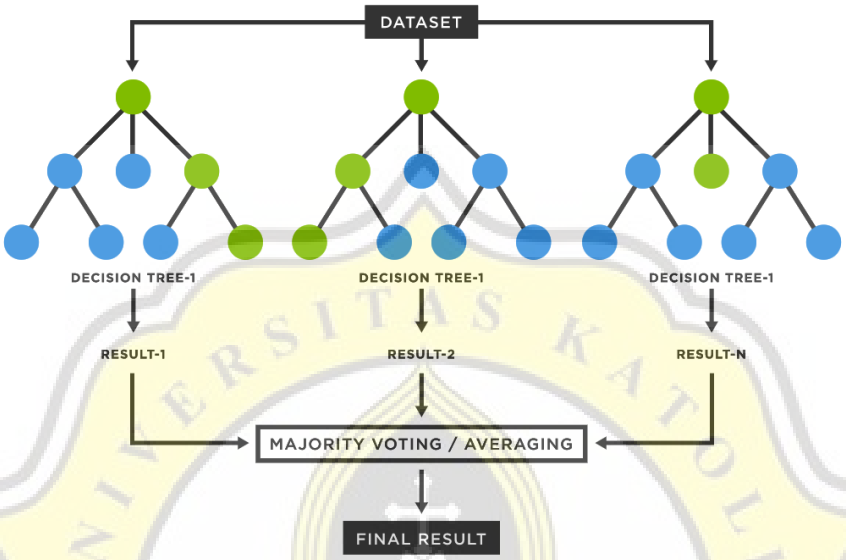

Figure 4.1 Random Forest

The workings of the Random Forest algorithm can be described in the following steps:

1. The algorithm selects random samples from the provided dataset.
2. Create a decision tree for each selected sample. Then the prediction results will be obtained from each decision tree that has been made.
3. A voting process is performed for each prediction result. For classification problems, the mode (most frequent value) will be used, while for regression problems, the mean (average value) will be used.
4. The algorithm will choose the most voted prediction result as the final prediction.

4.3. Design

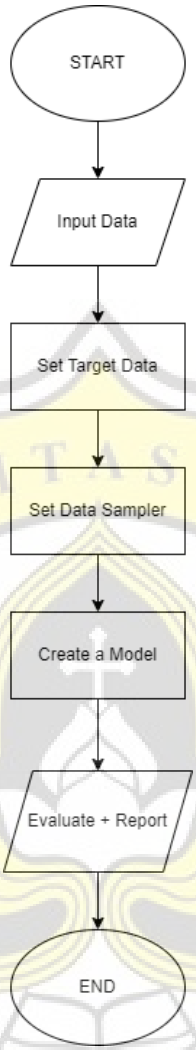

Figure 4.1 Flowchart Orange

The first thing to do is enter data into Orange. The data that has been entered has data attributes, one of the data attributes is determined by one of the target data used as a prediction. If you have determined the next data target, that is, set the Data Sampler by determining the proportion of data used in the test. If you have determined the proportion of data used, the next step is to create a regression model. In this study, the models created are Random Forest and Tree.

4.4. Function

This study uses the Regression method which has three functions, namely MSE (Mean Squared Error), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error). All three functions are used in performing the final calculation.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(target - prediction)^2 \quad (1)$$

In function (1), n is the sample size, Σ is the sum, $target$ is the original value, and $prediction$ is the predicted result of the data. The entire function (1) describes the *MSE* calculation.

Mean Squared Error measures how close the regression line is to the fixed points of the record. It is a risk feature similar to the cost of loss prediction squared error. The mean squared error is calculated by taking the average, mainly mean, squared error of the records pertaining to the feature.

$$RMSE = \sqrt{MSE} \quad (2)$$

In function (2), $\sqrt{MSE}$ is the division of the MSE result in function (1). The entire function (2) describes the $RMSE$ calculation.

Root Mean Square Error (RMSE) is the significance of the prediction mistake rate, wherein the smaller (toward 0) the RMSE value, the greater correct the prediction effects will be. Root Mean Squared Error (RMSE) is one manner to assess linear regression fashions with the aid of measuring the accuracy of a model`s forecast effects.

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|target - prediction| \quad (3)$$

In function (3), n is the sample size, Σ is the sum, $target$ is the original value, and $prediction$ is the predicted result of the data. The entire function (3) describes the *MAE* calculation.

MAE measures the common signs of the mistake in a fixed of predictions, without thinking about its direction. It is the common check pattern of absolute variations among predictions and real observations in which all character variations have identical weights.