CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1. Implementation

After the dataset has been collected, in implementing the two algorithms chosen, namely Neural Network and Random Forest, a program called Orange is needed to analyze and find predictive results from the application of the two algorithms.

When you first create a new workflow design in the Orange program, a window will appear as shown in Figure 5.1.

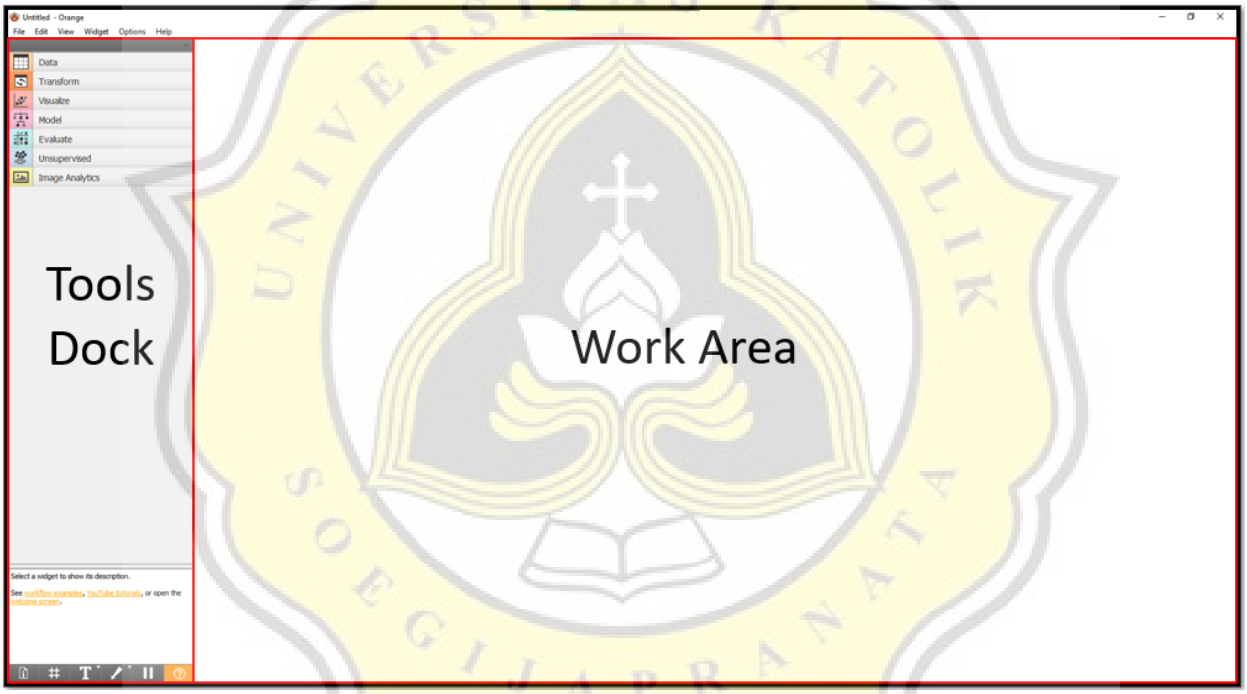


Figure 5.1 Widget (Tools Dock) and Work Area in Orange Data Mining Application

In this window there is a Widget which is a collection of tools used to process data, consisting of Data, Transform, Visualize, Model, Evaluate, Unsupervised, and Image Analysis (Optional).

The first step is to import the dataset into Orange, open the Data menu in the widget area, then select according to the dataset format you have. In this project, the dataset is comma separated values (.csv) format, so select the "CSV File Import" option. In Orange, the functions of the tools

used only need to be "drag-and-drop". This is one of the advantages of the Orange program in processing data. Then place the “CSV File Import” Menu to the work area. After that, select the menu to then search for dataset files on the computer by clicking the "browse" button like the example in Figure 5.2.

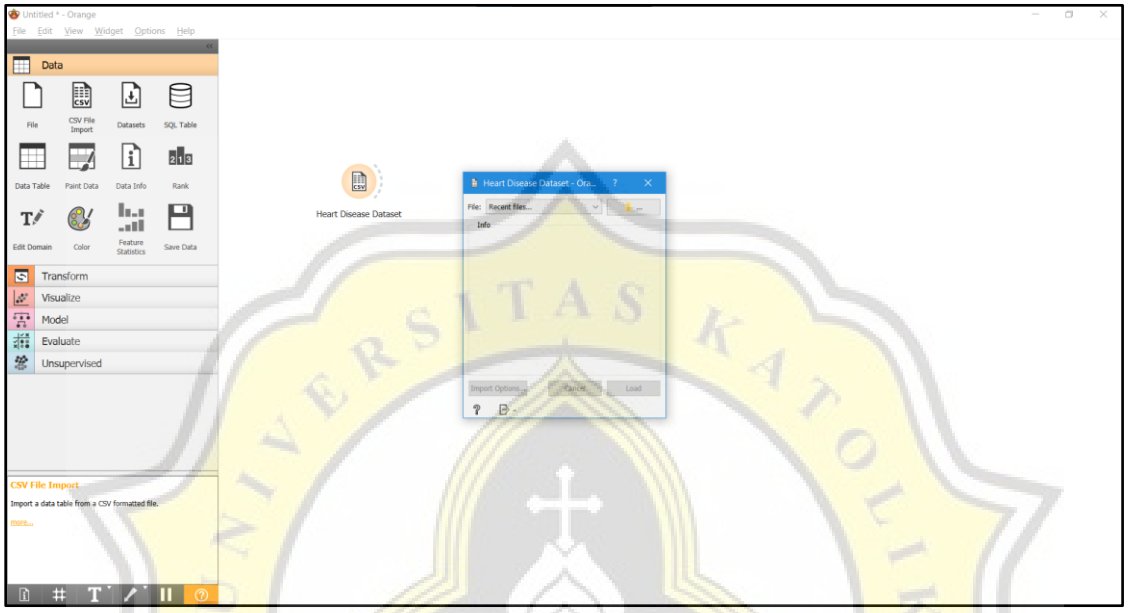
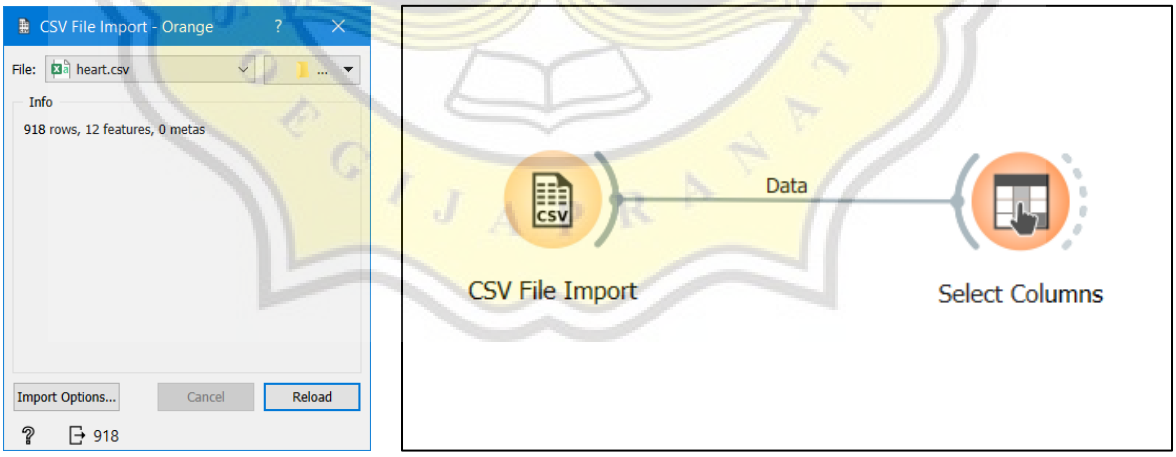


Figure 5.2 Import CSV File on Orange Data Mining

If the file is successfully imported, it will look like the one in Figure 5.3.



(a) (b)

Figure 5.3 (a) Display of CSV File Import Window on Orange (b) Linking Between Menus on Orange

After the CSV file is successfully imported, the dataset will enter the Transform stage, where the data will be processed first before entering the next stage. This transformation process is one of the data preprocessing techniques. At this stage, the Select Column and Data Sampler menus are used in the Category Transform on the widget menu. The first step to take is to connect the dataset with the select column first as in the example in Figure 5.4 (a) to select the column to be used as the target feature where the column will be searched for by a trained algorithm. In this project, the heart disease column was selected as the target feature as shown in Figure 5.4 (b).

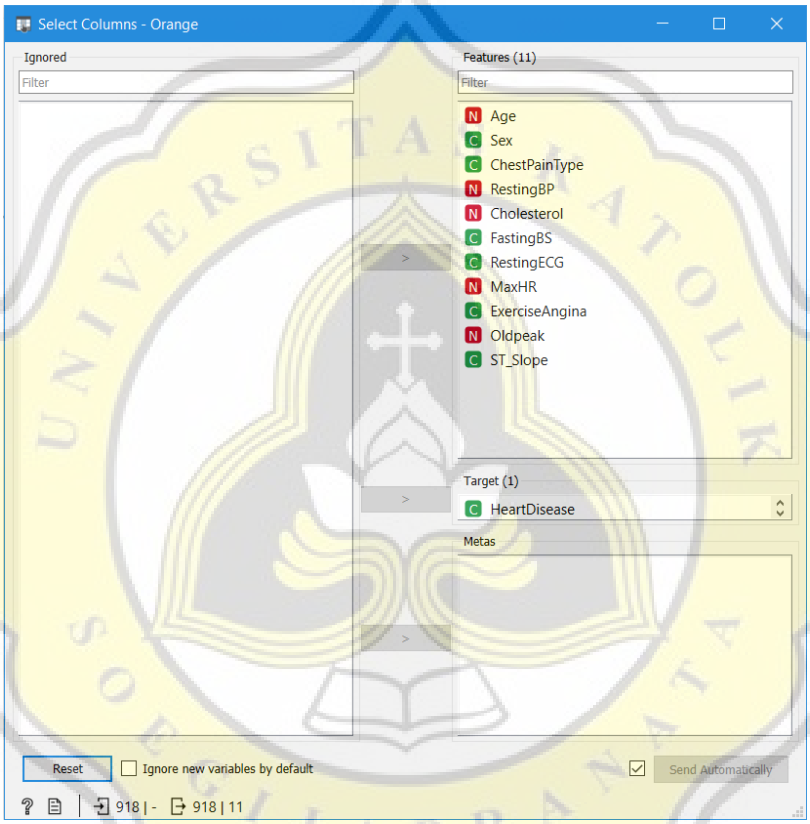


Figure 5.4 Select Column Window

After setting the data on the select column menu, then the data will be connected and sorted randomly, where the dataset will be separated to be used as sample data as the data used to train the algorithm and the remaining data as the remaining data to be used for the testing process. In this process, you can set how much data will be used in the sampling process. There are 4 types of sampling in the Orange program, which we will use in this project is the "Fixed Proportion of Data" option. This option will make the proportion of data that will be used for the sampling

process with a percent scale as shown in Figure 5.5. There will be 3 scenarios for the data used to train the algorithm, namely 20%, 50% and 80%.

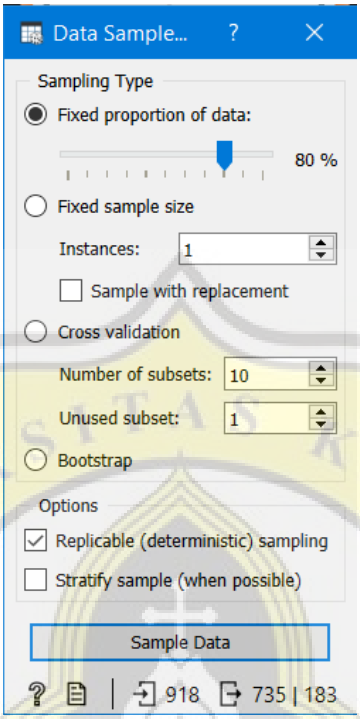
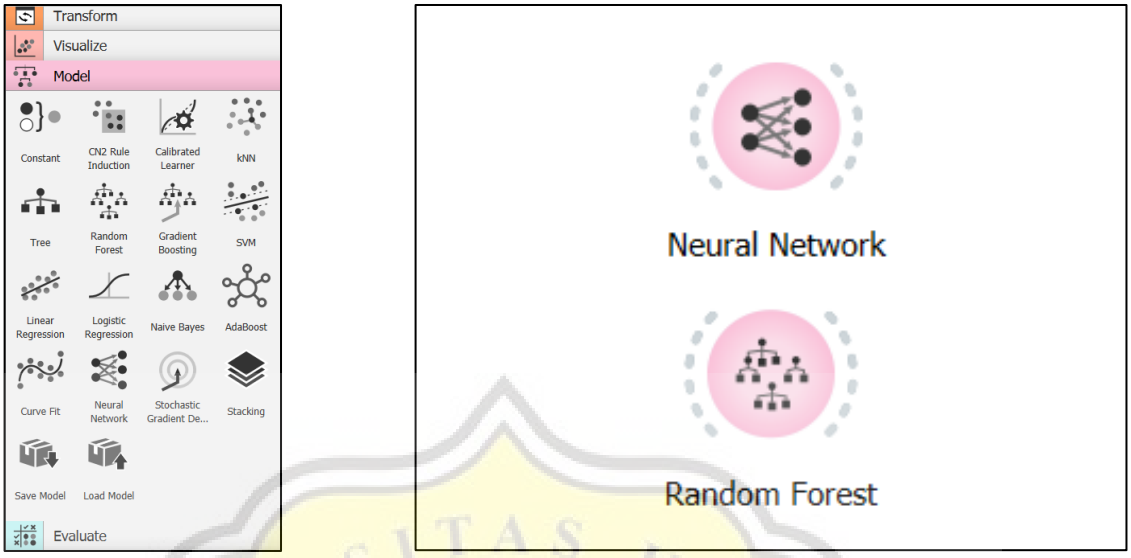


Figure 5.5 Display Selecting the Sampling Type on the Data Sampler

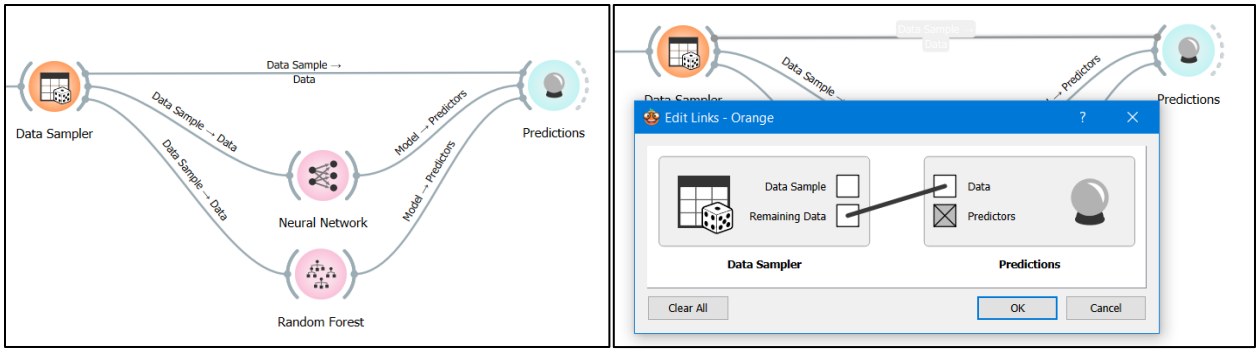
After setting the sampler data, the transform process is After setting the sampler data, the transform process in this project has been completed and will enter the next process, namely the Model. In this process, the data that has been selected from the sampler data process will be entered into the selected algorithm, this time the algorithm used to process heart disease patient data is a Neural Network and Random Forest. Then we will drag-and-drop the Neural Network and Random Forest menus from the widget menu to the work area like the example in Figure 5.6.



(a) (b)

Figure 5.6 (a) The Models Menu Widget on Orange (b) Neural Network and Random Forest Algorithms in the Work Area

Enter the last process, namely Evaluation. At this stage the algorithm that has been trained will be tested with the same dataset with different data values to produce value results or predictions on the data that is already available. The data used here is different from that used in the process of training the algorithm. At this stage, the menu used is "Predictions" to predict the results of training the algorithm on new data. In order to produce new prediction results. The Predictions menu will be linked to the two selected algorithms and connected to the Data Sampler process as shown in Figure 5.7 (a). In the Data Sampler connector with Predictions, the data used for testing will be changed, the data used is the remaining data that is not used in the training process. Double-click on the connector between the Data Sampler and the Predictions, then connect the Remaining Data box on the Data Sampler with the Data on the Predictions like example 5.7 (b)



(a) (b)

Figure 5.7 (a) Inter-Process Connection on Orange (b) Setting Remaining Data as Data Testing Algorithm

After all processes are connected, we will see the results of testing from training the selected algorithm. Double-click on the predictions process to see the test results of the algorithm predicting the results of the heart attack dataset, when the test results are out, the result score for the accuracy of the algorithm in predicting the results is also there. The display of the test results will show the AUC, CA, F1, Precision, and Recall scores as shown in Figure 5.8.

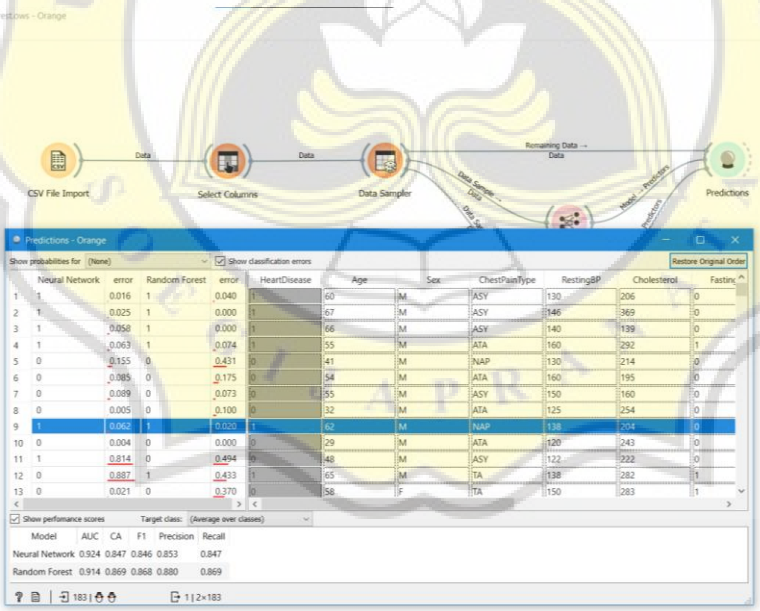


Figure 5.8 Prediction Test Results and Algorithm Accuracy Scores on Orange Data Mining

5.2. Results

The results of training the two algorithms using the selected datasets produce satisfactory predictive results. From the prediction results of the two algorithms, both algorithms produce high accuracy results. By doing various scenarios by changing the ratio of the amount of training data in the dataset. The ratio of the amount of training data used is 20%, 50% and 80%. By using this scenario, the results of the prediction accuracy score for heart attacks are produced as shown in the graph in Figure 5.9.

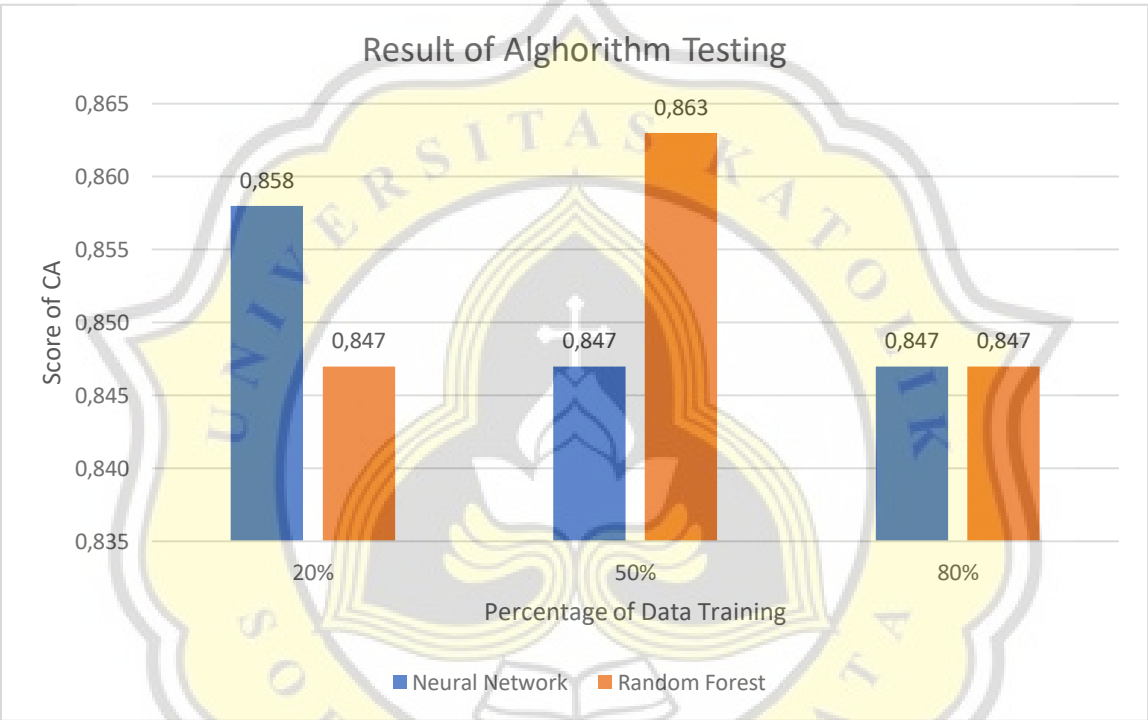


Figure 5.9 Graph of Neural Network and Random Forest Algorithm Accuracy Scores

In Figure 5.9, it can be seen that the results of the accuracy of the two algorithms are the lowest at 0.847 and the highest at 0.863. This score is included in the results of high accuracy. For a small amount of training data, we can see that the accuracy level of the Neural Network algorithm is higher than the Random Forest algorithm with a score of 0.858. At a training data ratio of 50%, the Random Forest algorithm has an accuracy score of 0.863, higher than the Neural Network at the same training data ratio. Even being the highest compared to other scenarios and in different algorithms. Meanwhile, at a training data ratio of 80%, both algorithms produce the same accuracy score at 0.847.