CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1. Linear regression steps

The linear regression method has a few steps in orange data mining tools to make prediction output and get the value of MSE, RMSE, MAE, and R2, here are a few steps to make output in orange data mining tools.

5.1.1. Importing CSV data

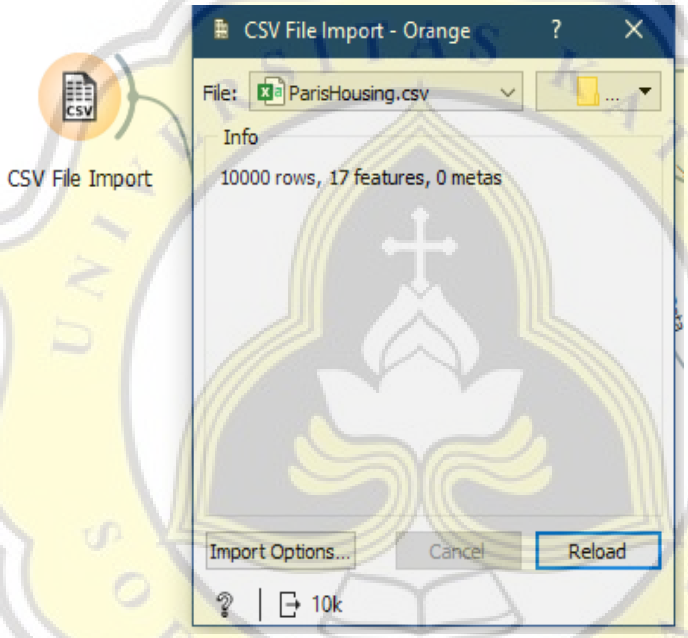

Figure 5.1 CSV Import Linear Regression

the first step is importing CSV by using the orange widget, file CSV from this study is get from keggle.com or you can also get a file CSV from another source, in this study my dataset has 10000 total data and also has 17 attributes.

5.1.2. Selecting attribute

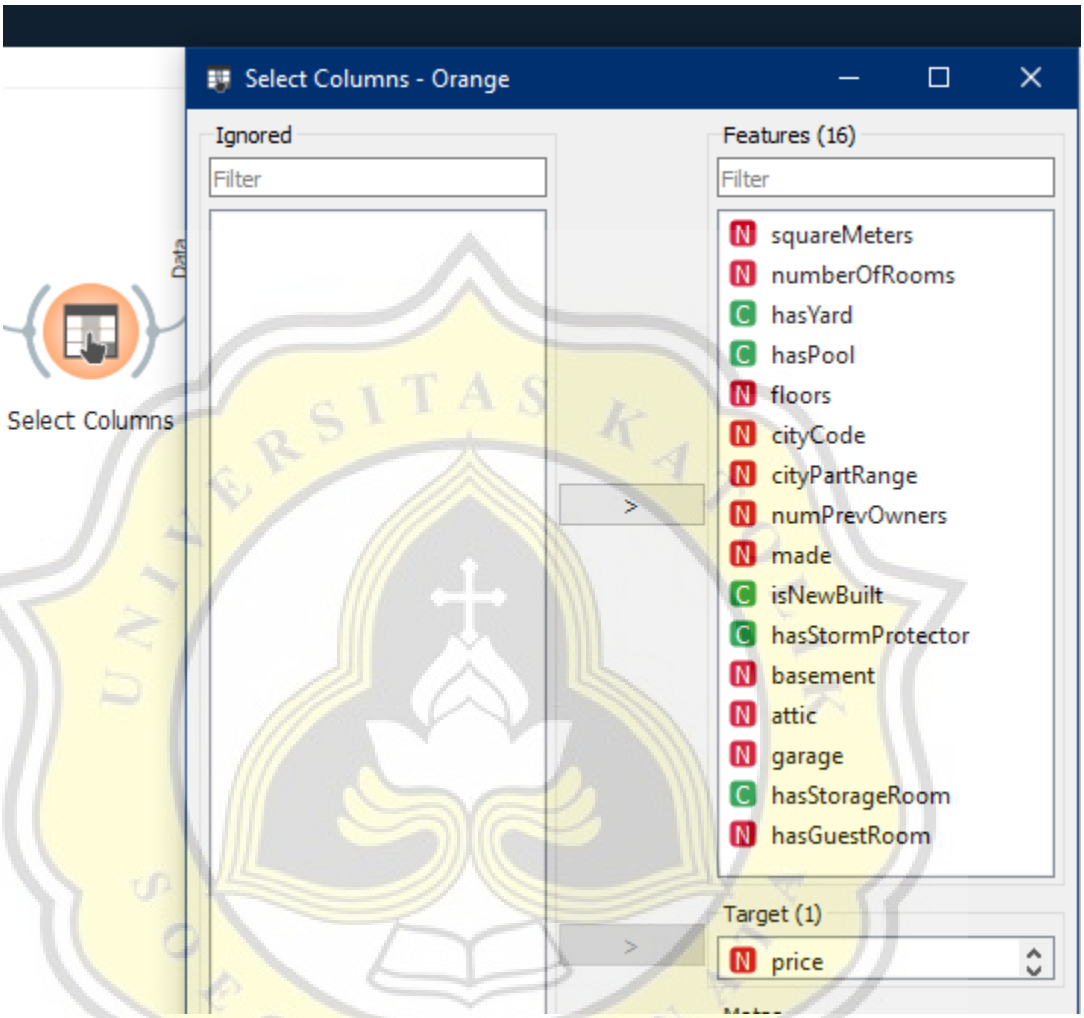

Figure 5.2 Attributes select Linear Regression

The next step is selecting the attribute target variable, by using the widget select columns in the orange data mining tool. Selecting the target variable is important in this step because this attribute has values modeled and predicted by another attribute, in this study the author used the price attribute as the target variable.

5.1.3. Data sampler

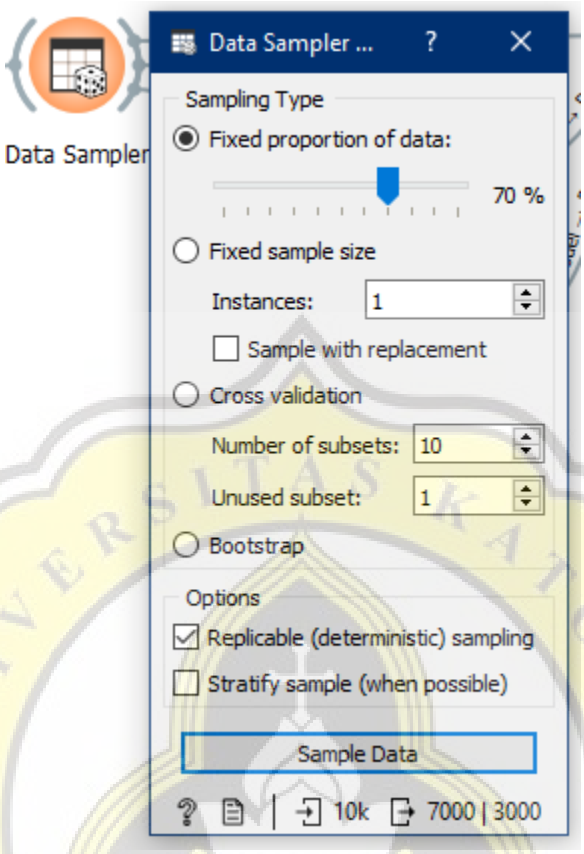

Figure 5.3 Data Sampler Linear Regression

In this step, by using the data sampler widget in the orange data mining tool, the author can divide data into two kinds of data, data testing and data training, in this widget if a fixed proportion of data have a 70% value, data training have 7000 data, and the rest of data become testing data, testing data can also called remaining data in widget.

5.1.4. Making Prediction

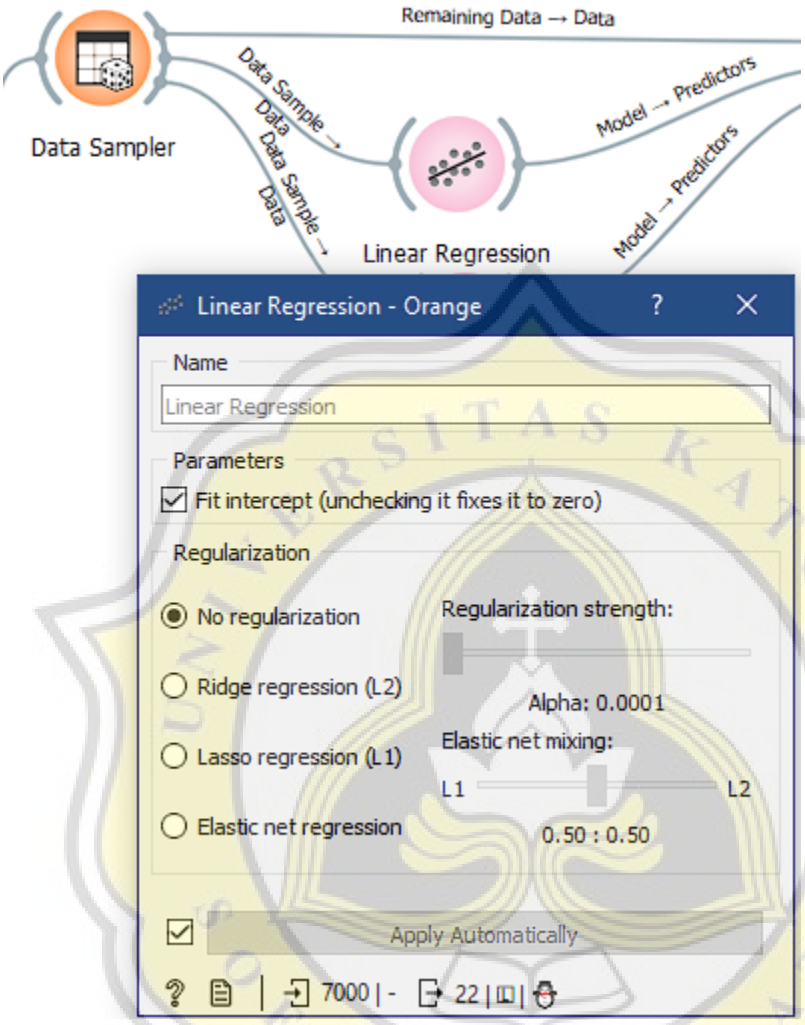

Figure 5.4 Making Prediction Linear Regression

After the data sampler, 7000 data training will be calculated and tested by using the method widget, before making a prediction, the input method widget is for output prediction, in this step author uses the linear regression widget method with fit intercept and np regulation setting on the widget.

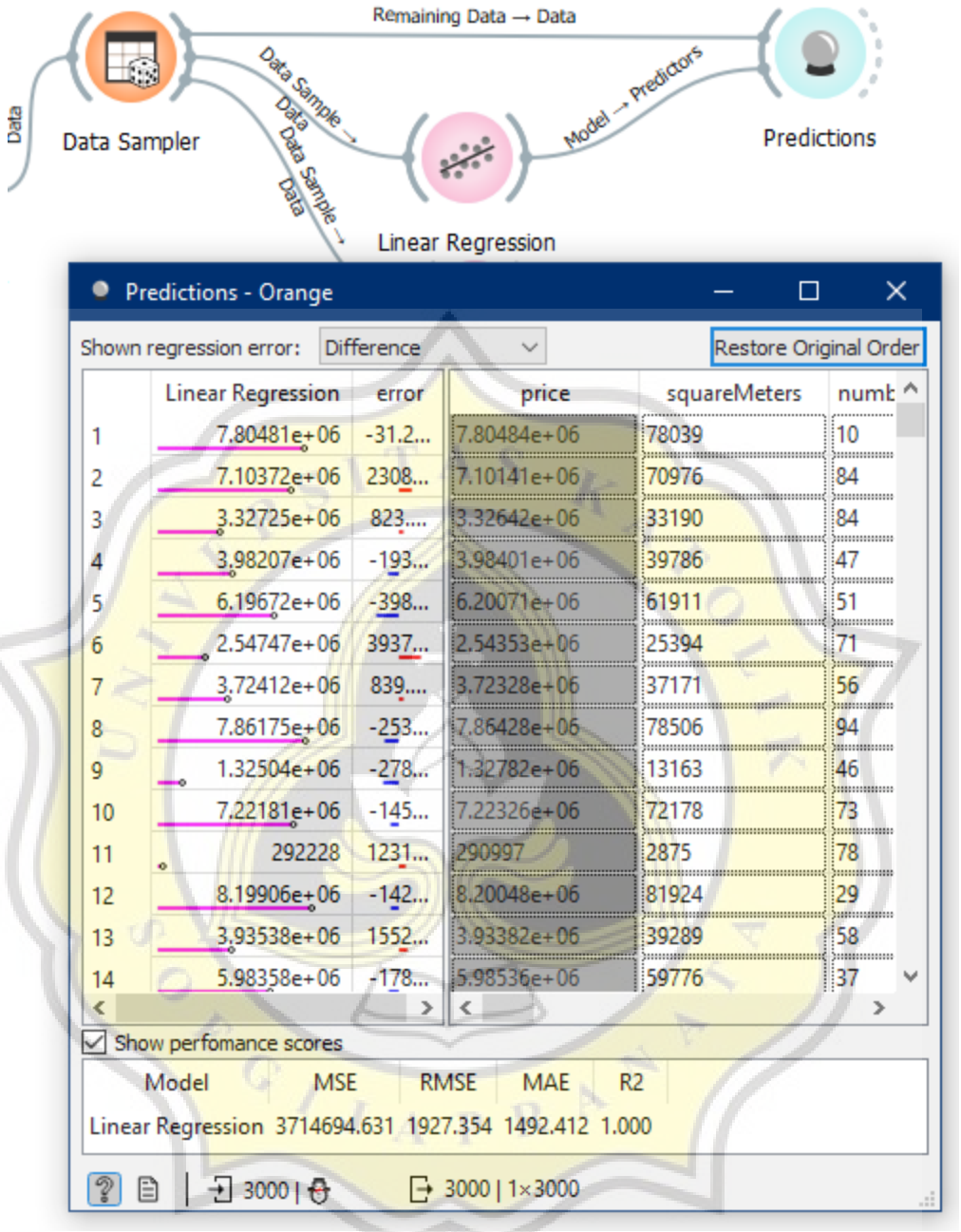

Figure 5.5 Linear regression prediction output

By using the prediction widget, the author can get output for linear regression method, and the author also get, MSE, RMSE, MAE, and R2 prediction output values as the accuracy of linear regression method matches with this dataset and regression, there's also a formula to get MSE, RMSE, MAE, R2

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

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

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

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
∑	Summation

5.2. Random forest steps

The Random forest method also has a few steps in orange data mining tools to make prediction output and get the value of MSE, RMSE, MAE, and R2, here are a few steps to make output in orange data mining tools.

5.2.1. Importing CSV

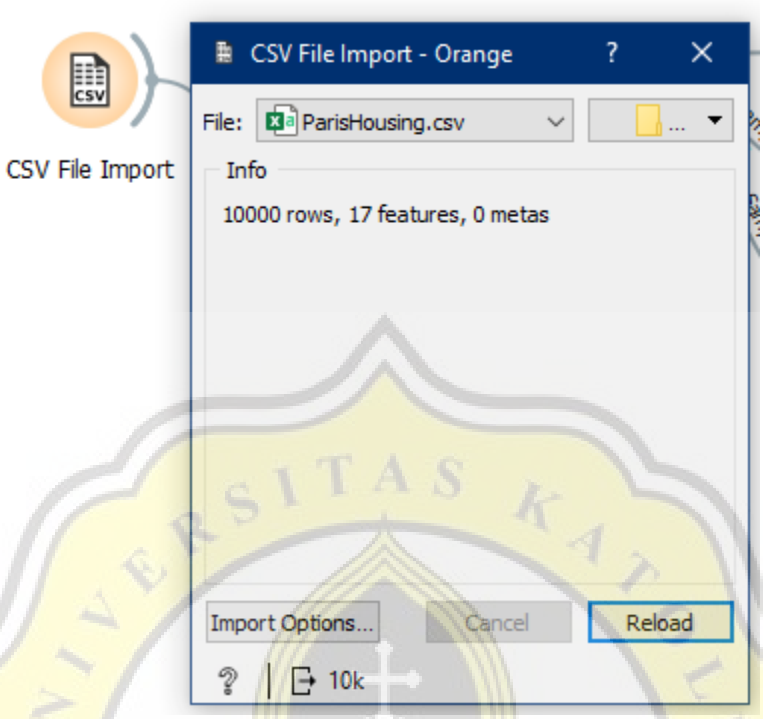

Figure 5.6 CSV Import Random Forest

Same as the first step in the linear regression step, the first step is importing CSV by using the orange widget, file CSV from this study is get from keggle.com or you can also get a file CSV from another source, in this study my dataset has 10000 total data and also has 17 attributes.

5.2.2. Selecting attribute

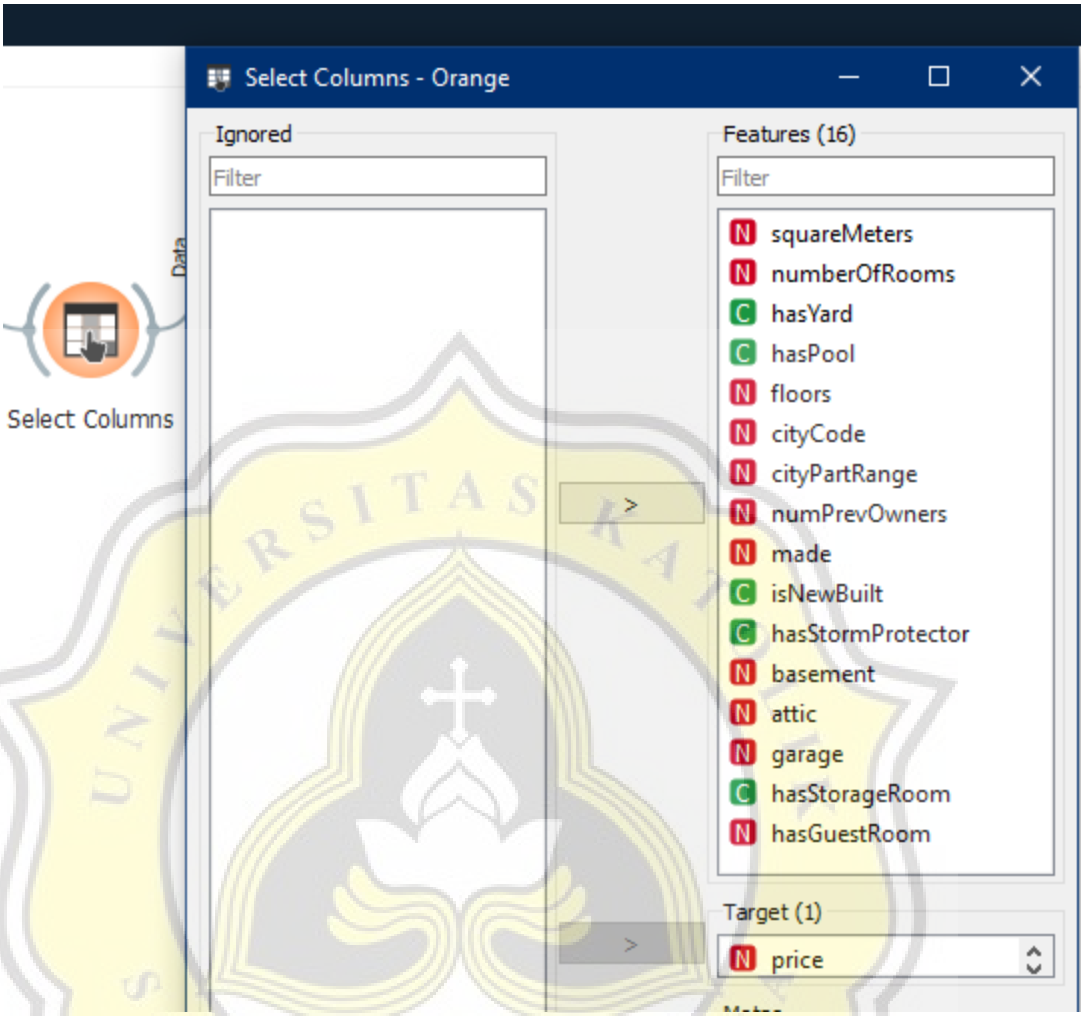

Figure 5.7 Selecting Attributes Random Forest

Same as linear regression steps, The next step is selecting the attribute target variable, by using the widget select columns in the orange data mining tool, selecting the target variable is important in this step because this attribute has values modeled and predicted by another attribute, in this study author used the price attribute as the target variable.

5.2.3. Data sampler

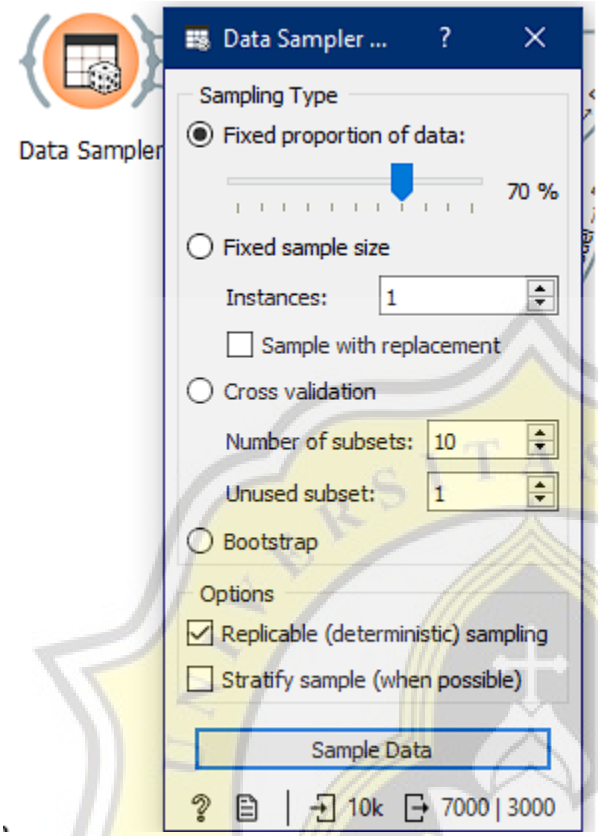

Figure 5.8 CSV Import Random Forest

Also same as linear regression steps, in this step, by using the data sampler widget in the orange data mining tool, the author can divide data into two kinds of data, data testing and data training, in this widget if a fixed proportion of data have 70% value, data training have 7000 data, and the rest of data become testing data, testing data can also called remaining data in widget.

5.2.4. Making prediction

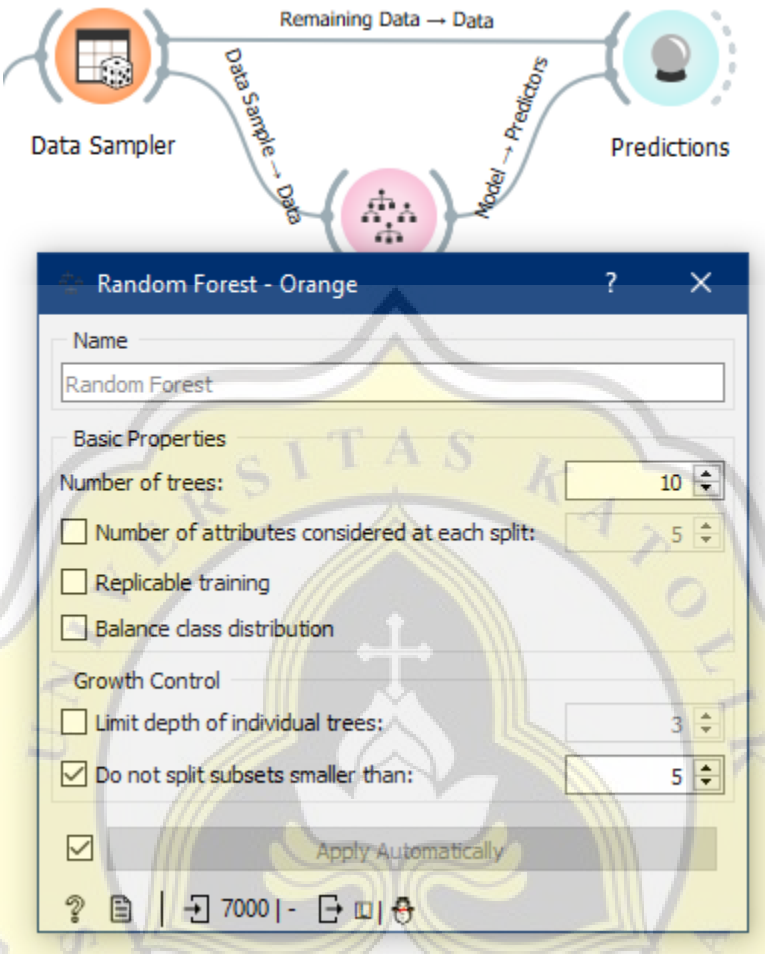

Figure 5.9 Making Prediction Random Forest

After the data sampler, 7000 data training will be calculated and tested by using the method widget, before making a prediction, the input method widget is for output prediction, in this step author using the random forest widget method with fit intercept and np regulation setting on the widget.

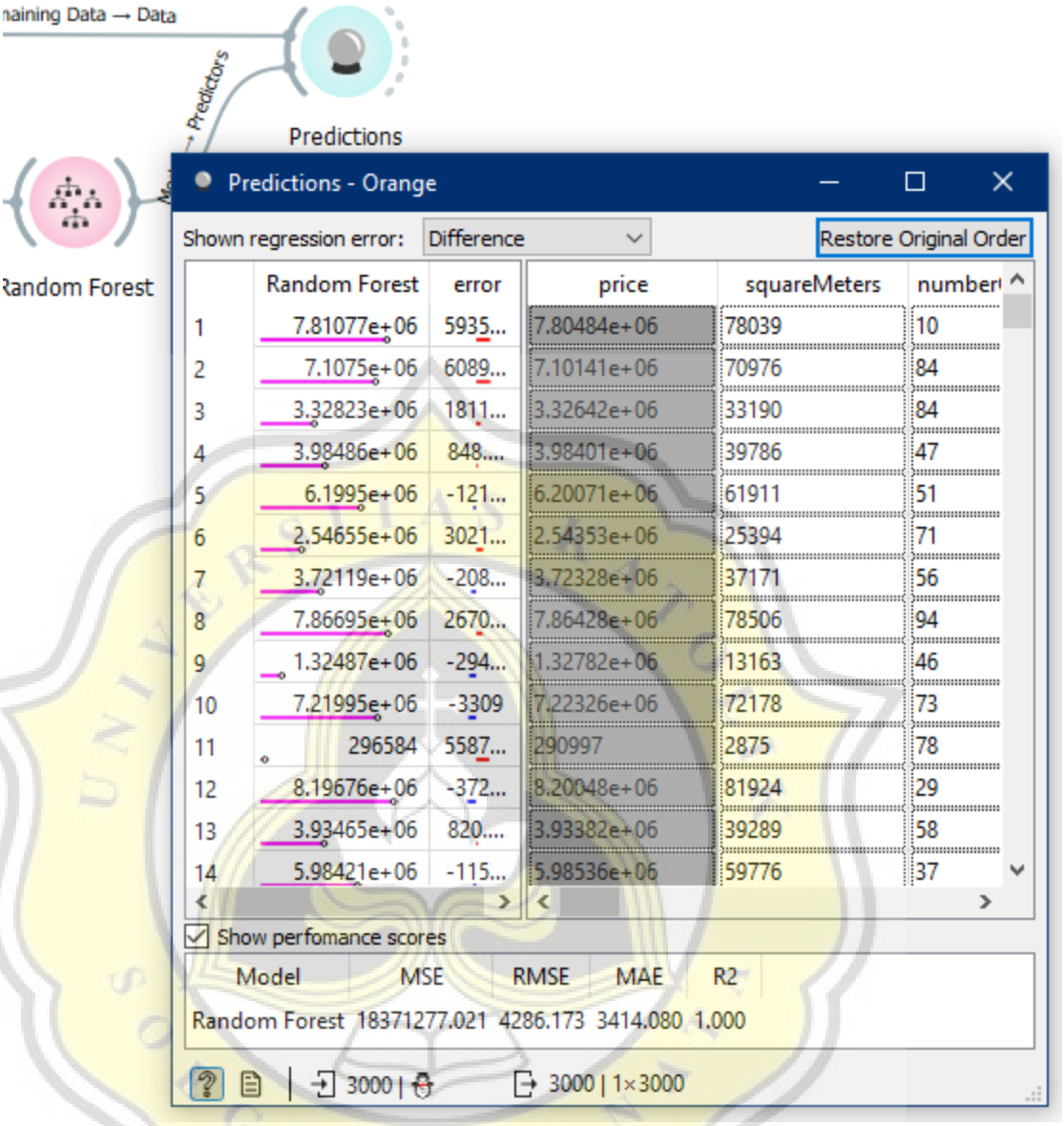

Figure 5.10 Prediction Output Random Forest

By using the prediction widget, the author can get output for the random forest method, and the author also get, MSE, RMSE, MAE, and R2 prediction outpuy values as the accuracy of the random forest method matches with this dataset and regression, there's also a formula to get MSE, RMSE, MAE, R2

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

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

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

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
∑	Summation

5.3. Results

After getting the output from the two methods, I compare both of them with different configurations and a total data sampler for testing data, that contains 60%, 70%, 80%, and 90% for both algorithms to predict which of them has less error, and better accuracy.

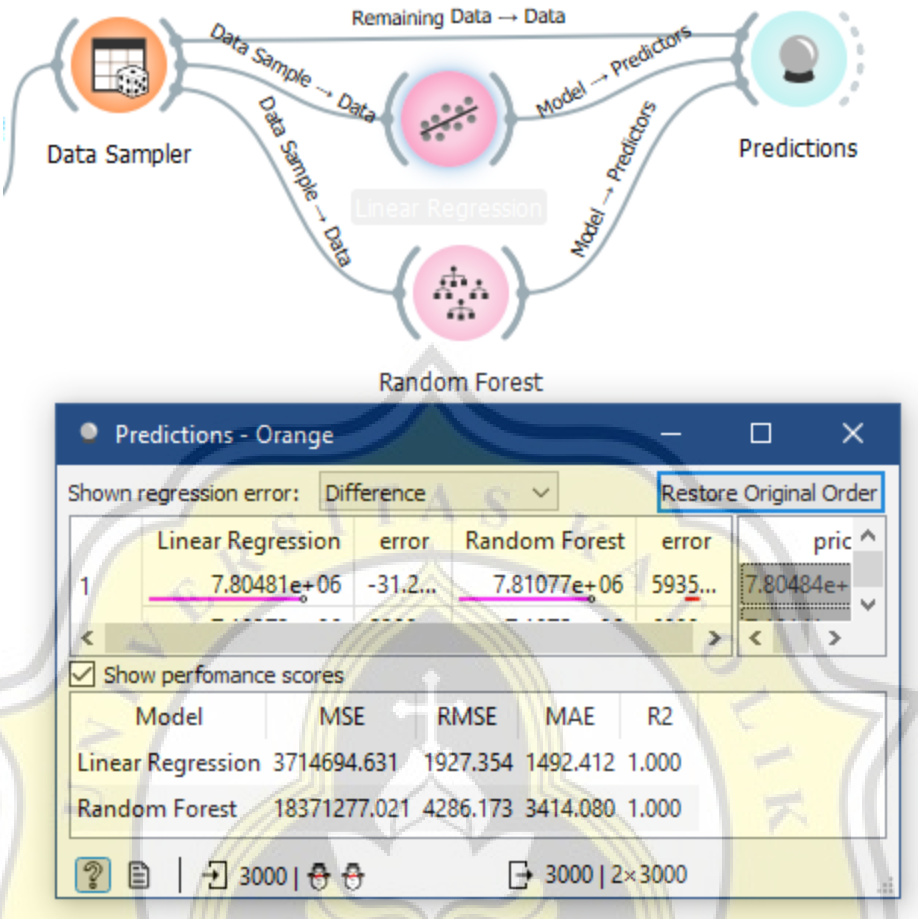

Figure 5.11 Linear Regression and Random Forest Orange Output

Table 5.1. Comparison MSE, RMSE, and MAE

DATA SAMPLER				
60%				
Model	***MSE***	***RMSE***	***MAE***	***R2***
Linear Regression	3646332.812	1909.537	1483.621	1
Random Forest	19580516.28	4424.988	3531.362	1
70%				
Model	***MSE***	***RMSE***	***MAE***	***R2***
Linear Regression	3714694.631	1927.354	1492.412	1
Random Forest	19217399.76	4382.765	3492.295	1
80%				
Model	***MSE***	***RMSE***	***MAE***	***R2***

Linear Regression	3626176.129	1904.252	1468.744	1
Random Forest	17599983.41	4195.233	3342.905	1
90%				
Model	***MSE***	***RMSE***	***MAE***	***R2***
Linear Regression	3837046.601	1958.838	1500.071	1
Random Forest	16887107.84	4109.393	3264.84	1

This table is a comparison between the total in the data sampler widget, in this study, I compared the 60%, 70%, 80%, and 90% total data sampler, and then also have output MSE, RMSE, MAE, and R2 as an output, less value output have meant that its better method for this dataset.

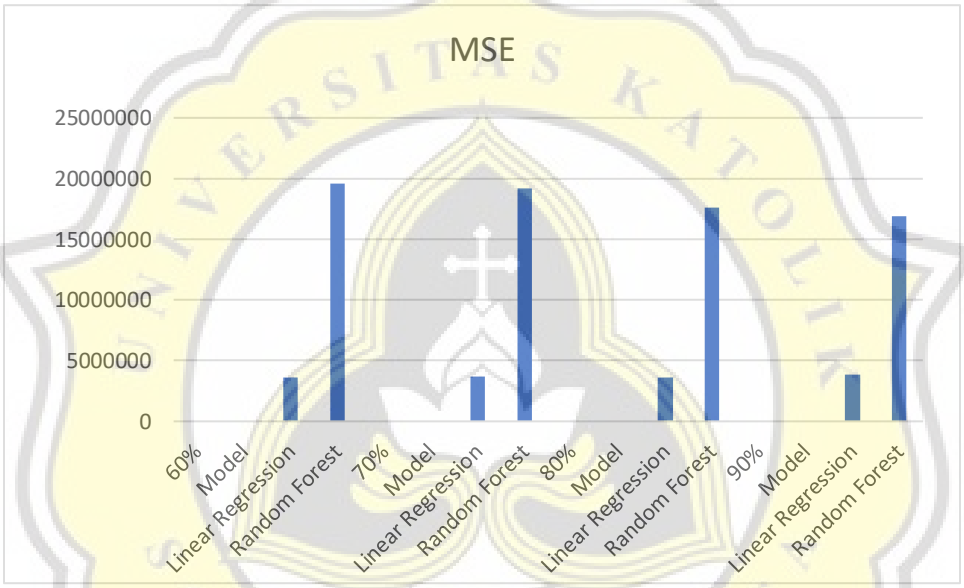

Figure 5.12 Graph comparison

MSE graph compassion by 60%, 70%, 80%, and 90% data sampler, in this graph linear regression, has less output than random forest output, in every data sampler, this means linear regression method has better output than random forest method in every data sampler tested data in every sampling.

Table 5.2. Comparison MAE and RMSE

Sample Data	Model	RMSE	MAE
60%	Linear regression	1909.537	1483.621
	Random forest	4424.988	3531.362

70%	Linear regression	1927.354	1492.412
	Random forest	4382.765	3492.295
80%	Linear regression	1904.252	1468.744
	Random forest	4195.233	3342.905
90%	Linear regression	1958.838	1500.071
	Random forest	4109.393	3264.84

This is a table for comparing RMSE and MAE values prediction output, in this table RMSE and MAE for linear regression and random forest output values prediction is shown, also linear regression have less output than random forest in every data sampler in 60%, 70%, 80%, and 90%.

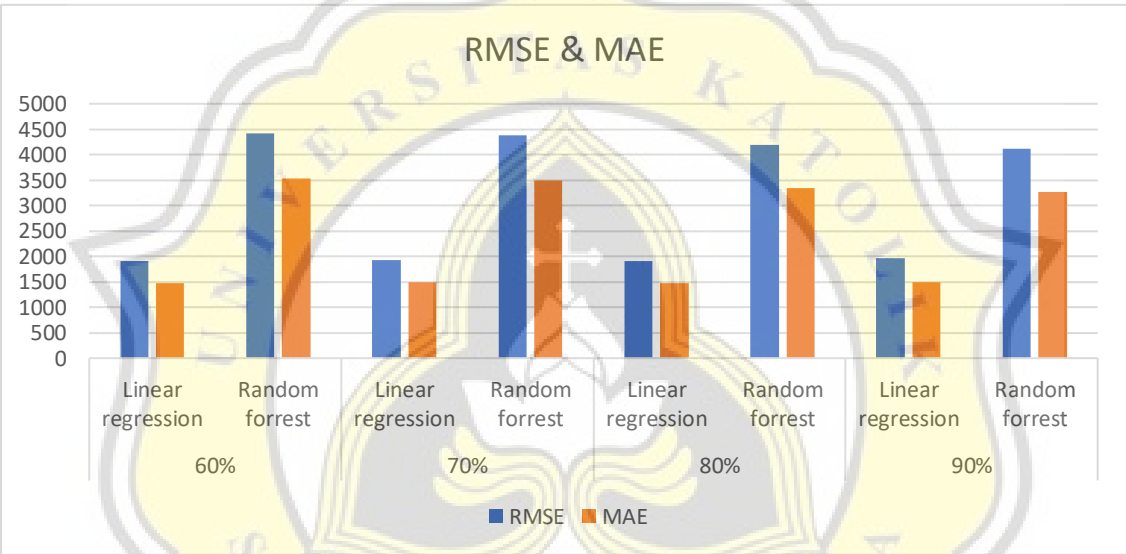

Figure 5.13 RMSE & MAE Graph

Based on output prediction for both method with calculating MSE, RMSE, MAE, AND R2, Linear regression have less error than random forest, in this case, linear regression is more suitable than random forest in this dataset, because linear regression only calculated predictor (xi) and response variable (y), the formula calculating all of the prediction output in this down below, the less the number in output its mean method is suitable for the dataset.

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

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

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

Also, the random forest in this prediction has to make a lot of random output and does not have any fixed output, for example, if split data have the same amount of 70% random forest always makes different output with less than 2% difference in the number in output prediction.

Also, the random forest function in orange builds a set of decision trees. Each tree is developed from a bootstrap sample from the training data. When developing individual trees, an arbitrary subset of attributes is drawn (hence the term "Random"), from which the best attribute for the split is selected. The final model is based on the majority vote from individually developed trees in the forest, because of the random selection tree in the random forest algorithm, this calculates random forest method has a lot of errors in this dataset because price attributes in this study have a lot of value and sometimes have a little of value.