Proceeding Konferensi Nasional Matematika (KNM) XXII
Universitas Bengkulu, 15-16 Juli 2024
Dipublikasikan online pada:
PATTIMURA PROCEEDING: Conference of Science and Technology
e-ISSN: 2829-3770, volume 5 issue 1, pp.101-116, Edisi November 2024

Comparison of Multiple Linear Regression and Random Forest Regression Models for House Price Prediction in Semarang City Using the CRISP-DM Method

Fransisca Mulya Sari[1*], Sugiman[2]

[1,2]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Sekaran, Gunungpati, Semarang City, 50229, Indonesia

Corresponding author's e-mail: **** mulyasarifransisca@students.unnes.ac.id***

Keywords:
House Price Prediction;
Multiple Linear Regression;
Random Forest Regression

ABSTRACT

The population density in Semarang City is increasing every year. This requires more potential land to build houses to accommodate the denser population. There are various kinds of house prices based on specifications in Semarang City. This requires the right prediction to get the desired house. This study implements and compares the performance of Multiple Linear Regression (MLR) and Random Forest Regression (RFR) models to predict house prices in Semarang City. The method used in this research is CRISP-DM (Cross-Industry Standard Process for Data Mining) as a data mining process. The data used in this research amounted to 9533 data with 8 variables obtained by web scraping. The data will go through a data preprocessing process then training the model. Next is the evaluation stage, which is carried out to measure the performance of the two models using evaluation metrics, namely R-Squared (prediction accuracy), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error). The results of this study show that the MLR model obtained a prediction accuracy 61.1% with a training and testing data division ratio of 75%: 25%. While the RFR model produces a prediction accuracy 78.4% with a training and testing data division ratio of 90%: 10%. This shows that the RFR model is the best performing model. This research successfully applied the RFR model to the streamlit web framework. The final result of this research is a website that can be used by the public to predict house prices according to criteria in Semarang City.

DOI: https://doi.org/10.30598/ppcst.2024.knmxxii.101-116

1. INTRODUCTION

Semarang City is the capital city of Central Java Province consisting of 16 districts and 117 urban village [1]. The population in Semarang City has increased from 2020 to 2022. Therefore, population density increases every year along with the increase in the population in Semarang City [2]. The increase and spread of population density in Semarang City can affect settlements, where residential settlements require land while land in urban areas is limited [3]. Therefore, there is a need for housing that can accommodate the increasing population density in Semarang City.

A house is a primary need in the form of a place to live that is needed by everyone [4]. There is a relationship between population density and housing development in Semarang City. If the population density increases, the number of houses in Semarang City will also increase [5]. This is why more and more property companies are building houses as a means of long-term investment by offering varying prices. This will make people think more when buying a house by analyzing whether the house they want to buy has a good profit value or not [6]. This will cause house prices to increase. Knowing the house price based on variables that can be a consideration for decisions can be done by using house price predictions.

Over time, predictions have been used as a tool for decision-making. The business and economic fields are fields that are quite often predicted [7]. In this context, a decision-making system is needed that can help prospective buyers make optimal home purchase decisions based on predetermined criteria and prices. Multiple Linear Regression and Random Forest Regression are two algorithms Machine Learning which is in sufficient interest to conduct a study related to prediction.

Multiple Linear Regression (MLR) is a statistical method used to find the most appropriate linear relationship between a single bound variable and a set of independent variables [8]. Random Forest Regression (RFR) is a general-purpose regression method that combines multiple random decision trees and combines their predictions by averaging [9]. Process model Data Mining CRISP-DM (Cross-Industry Standard Process for Data mining) is the standard de facto to develop the project Data Mining and the discovery of knowledge. CRISP-DM consists of six phases, namely: business understanding, data understanding, data preparation, Modelling, evaluation, and Deployment [10]. CRISP-DM has stages Deployment. This stage contains the incorporation of a pre-built model into the decision-making process, one example is creating a website. When making website, can use Web Framework one of them is Streamlit. Web framework Streamlit simply uses the Python programming language. In addition, the use of Web Framework Streamlit is unpaid and accessible to the public [11].

Based on this description, the formulation of the problem in this study is how the Multiple Linear Regression (MLR) and Random Forest Regression (RFR) models perform in predicting house prices in Semarang City, which model has the best performance, and how the procedure for applying the model with the best performance to the web framework streamlit. The purpose of this study is to determine the performance of the model, determine which model has better performance, and to apply the model with the best performance in the streamlit web framework. The essence of this study is to compare the performance of MLR and RFR models in predicting house prices in Semarang City and then create a website using the model with the best performance.

2. RESEARCH METHODS

This study uses a quantitative research method using the CRISP-DM (Cross-Industry Standard Process for Data mining) method which consists of six stages, namely business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The prediction in this study uses the Multiple Linear Regression (MLR) and Random Forest Regression (RFR) models. The stages in this study are based on the CRISP-DM stages according to the following flowchart.

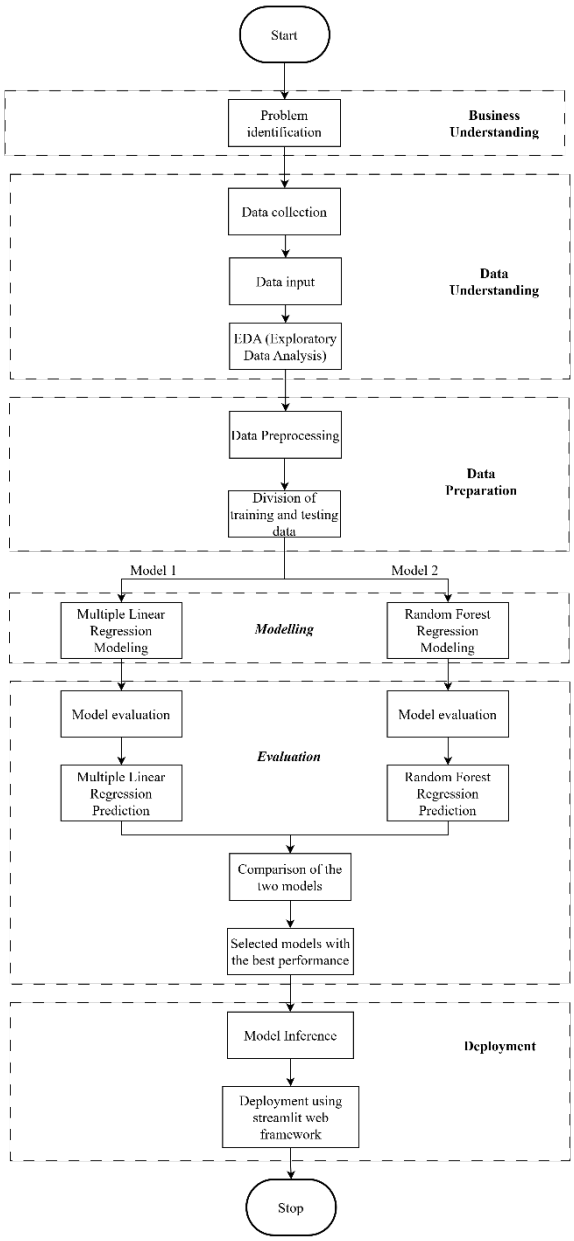

Figure 1. Data analysis flowcharts

2.1 Business Understanding

This stage is the initial stage in the CRISP-DM data mining process model, namely problem identification. The problem solved in this research is to realize a house price prediction system in Semarang City by using data from web scraping which is then developed into a house price prediction website where the data is based on input data provided by users. A suitable method is used to solve this problem using a regression algorithm that can predict the value of the input given.

2.2 Data Understanding

This stage is a stage to obtain data and understand the data that has been collected. This stage includes data collection obtained by web scraping from the website www.rumah123.com, data input, and EDA (Exploratory Data Analysis). The EDA stages used in this research include displaying datasets, data information, description of statistical data (numeric), data visualization, and correlation analysis.

2.3 Data Preparation

The data preparation phase includes all the efforts required to create the final dataset (the data to be fed into the modeling) from the raw data. This stage is to prepare the data before entering the machine learning model. This stage includes data preprocessing and data division into training data and testing data. The data preprocessing stages used in this research include cardinality, label encoding, and skewness value. The division of data into training data and testing data is done with several data division ratios with the aim of finding the data division that produces the best model prediction accuracy. The data division ratios include 90%: 10%, 80%: 20%, 75%: 25%, and 70%: 30%.

2.4 Modeling

This stage is a stage to model data that has been processed before using machine learning models, namely Multiple Linear Regression and Random Forest Regression.

Multiple linear regression is a linear regression model to analyze the relationship between one dependent variable and more than one independent variable. The model equation for multiple linear regression with continuous predictors is as follows [12].

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (1)$$

Description:
Y_i : Response for i-th subject
X_i : Predictor for i-th subject
Y_i : Error for i-th subject

Random forest is one of the supervised learning machine learning algorithms. Random forests can be used for classification and regression problems. Random Forest is a variant of bagging ensemble. Random forest is the combination of all decision trees in such a way that each tree relies on random vector values that are sampled independently and with the same distribution for all trees. The strength of a random forest is determined by using a process of selecting appropriate features for each sample (node), which can result in relatively low error measurements. Bagging uses all attributes to build an independent model whereas random forest uses only a subset of features (usually 20% of the total number of features). Random forest will be more computationally efficient by using this method. The set of independent models built by random forest is also more varied than Bagging [13].

2.5 Evaluation

This stage is to evaluate the performance of the model in predicting house prices in Semarang City. In this stage, the performance of the model will be measured using evaluation metrics, namely R-Squared to measure the accuracy of prediction as well as MSE and RMSE to measure the error rate of the model. After that, the performance of the two models will be compared based on the evaluation metrics. Then, the best performing model is selected, meaning that it has the highest R-Squared value and the lowest MSE, RMSE values.

Mean Squared Error (MSE) is a method to evaluate the squared error in a forecast or prediction. The evaluation in question is the average squared difference between the values to be estimated. A normal MSE value is always positive (non-zero), which indicates imperfections in the forecasting or prediction results. The following shows the standard Mean Squared Error equation [14].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y - \hat{Y}_i\right)^2 \quad (2)$$

Description:
n : Number of data or samples
i : Independent variable
y : Actual value
$\hat{Y}$: Predictive value

Root Mean Square Error (RMSE) is one pf the methods used to measure the level of error in the estimation results. The error in question produces some significant differences or deviations from the estimated value. The purpose of using Root Mean Squared Error is to measure the amount of error in the analysis results obtained through various methods such as training data and test data. The following is the Root Mean Square Error equation in general [14].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y - \hat{Y}_i\right)^2}{N}} \tag{3}$$

Description:
n : Number of data or samples
i : Independent variable
y : Actual value
$\hat{Y}$: Predictive value

The Coefficient of Determination (R-Squared) is a statistical measure that shows how much the relationship between the independent variable (X) and the dependent variable (Y) is. The coefficient of determination ranges from $0 \leq R2 \leq 1$. This means that when the coefficient of determination is less than or closer to 1, the relationship between variables X and Y is getting stronger. However, if the coefficient of determination is close to or equal to 0, then the coefficient of determination becomes weaker or there is no relationship between variables X and Y. The following is the general equation for the coefficient of determination [14].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y - f(\hat{y}_i)\right)^2}{\sum_{i=1}^{n}(y - \hat{y}_i)^2} \tag{4}$$

Description:
y : Actual value dependent variable
$\hat{y}$: Predicted value dependent variable
$\hat{y}$: Mean of the value of y

2.6 Deployment

This stage is a stage for deployment the results of the machine learning model prediction with the best performance to the public by creating a website related to house price predictions in Semarang City. This stage includes model inference, setting up supporting files, upload files to GitHub repositories, log in to the streamlit page to create a website, website formation process, and results of the house price prediction website.

3. RESULTS AND DISCUSSION

3.1 Problem Identification

The problem solved is predicting house prices in Semarang City by using data from web scraping which is then developed into a house price prediction website where the data is based on input data provided by users.

3.2 Data Collection

The data of this study was collected from https://rumah123.com with the specifications of Semarang City using web scraping with the help of Octoparse software. Data collection was carried out on March 23, 2024 with 9533 data collected.

3.3 Exploratory Data Analysis (EDA)

The EDA stage is a stage to explore data with the aim of getting to know the data more deeply. The EDA stage is carried out using python via Jupyter Notebook.

3.3.1 Displaying Datasets

	House_Type	Location	Bedroom	Bathroom	Garage	Land_Area	Building_Area	Price
0	House	Tembalang, Semarang	2	1	1	60	30.0	0.3
1	House	Semarang Timur, Semarang	2	1	1	65	30.0	0.3
2	House	Gunung Pati, Semarang	2	1	1	120	32.0	0.3
3	House	Genuk, Semarang	2	1	1	60	36.0	0.3
4	House	Tembalang, Semarang	2	1	1	72	36.0	0.3

Figure 2. Web scraping results dataset

The data that has been collected through web scraping has seven independent variables, namely house type, location, bedroom, bathroom, garage, land area, building area, and one dependent variable, namely price. The price variable is in billions of rupiah, so if the price = 0.3 then it means that the price = 0.3 billion = 300 million.

3.3.2 Data Information

This stage is useful for providing information related to the data used, namely in the form of column names, the amount of non-null data, and the data type of each variable.

```
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   House_Type     9533 non-null   object
 1   Location       9533 non-null   object
 2   Bedroom        9533 non-null   int64
 3   Bathroom       9533 non-null   int64
 4   Garage         9533 non-null   int64
 5   Land_Area      9533 non-null   int64
 6   Building_Area  9533 non-null   int64
 7   Price          9533 non-null   float64
```

Figure 3. Data Information

Based on the figure, it can be concluded that the data used does not contain null values. The house type and location variables have categorical data types while the bedroom, bathroom, garage, land area, building area, and price variables have numerical data types.

3.3.3 Description of Statistical Data (Numeric)

This stage is useful for providing information related to the description of statistical data for numerical variables.

	Bedroom	Bathroom	Garage	Land_Area	Building_Area	Price
count	9533.000000	9533.000000	9533.000000	9533.000000	9533.000000	9533.000000
mean	3.326130	2.141928	1.303053	175.165425	155.920172	1.653170
std	1.350971	1.159489	0.706326	109.270013	107.200283	1.181362
min	2.000000	1.000000	1.000000	28.000000	29.000000	0.300000
25%	2.000000	1.000000	1.000000	104.000000	77.000000	0.700000
50%	3.000000	2.000000	1.000000	142.000000	124.000000	1.300000
75%	4.000000	3.000000	1.000000	210.000000	200.000000	2.300000
max	9.000000	10.000000	10.000000	912.000000	900.000000	5.000000

Figure 4. Description of Statistical Data

The description of statistical data includes calculations for each numerical variable, namely count (number of rows), mean (average), std (standard deviation), min (minimum value), 25% (first quartile value or Q1), 50% (second quartile value or Q2), 75% (third quartile value or Q3), and max (maximum value).

3.3.4 Visualization of Pairplots

Figure 5. Pairplot

Visualization of pairplot is useful for knowing the relationship between two variables. If the result resembles a straight line, then the relationship is linear. If not, then the relationship is non-linear. Based on the figure, it is concluded that the relationship between the two variables is diffuse so that it is a non-linear relationship.

3.3.5 Distribution of house sales in Semarang City

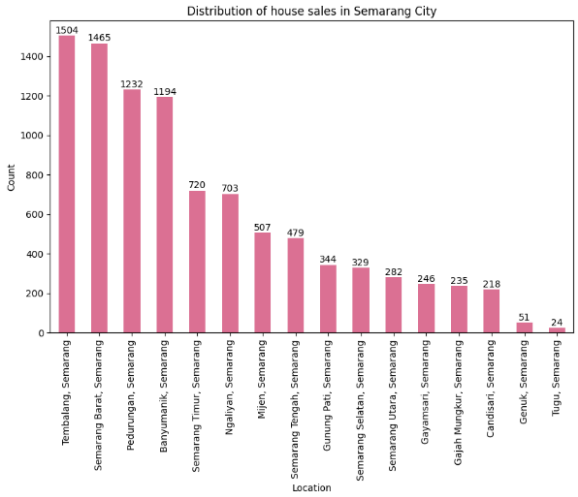

Figure 6. Distribution of house sales in Semarang City

Based on this data, Tembalang is the district with the most house sales in Semarang City. Meanwhile, Tugu is the district with the least house sales.

3.3.6 Correlation Analysis

It is used to determine whether there is a relationship between two variables, as well as to calculate the strength and direction of the relationship. The value of the relationship between these two variables ranges from $-1 \leq$ correlation ≤ 1.

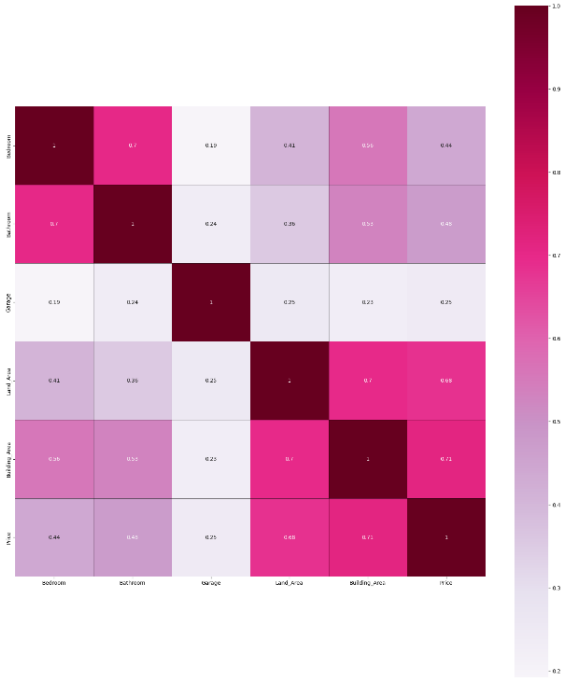

Figure 7. Correlation analysis

The results obtained the correlation level between the building area and price is the highest correlation with a value of 0.71. Meanwhile, the correlation rate between garage and price has the lowest correlation with a value of 0.25.

3.4 Data Preprocessing

It is the process of identifying, correcting, and eliminating errors in data to ensure data quality before machine learning modeling.

3.4.1 Cardinality

Cardinality (unique value) affects the encoding process. This is because categorical variables will be converted to numerical. So, it is important to be able to determine the right encoding technique based on the number of unique values.

Table 1. Cardinality

Variable	Cardinality
House type	3
Location	16
Bedroom	8
Bathroom	10
Garage	10
Land area	500
Building area	363
Price	487

The unique value of the house type variable is 3 while the unique value of the location variable is 16. Therefore, an encoding technique is needed that maintains data in one column so that there are not many new columns. This will also affect the deployment process. On that basis, encoding labels are needed.

3.4.2 Label Encoding

It is a technique that converts any data that is categorically typed into a numerical type that keeps the data in a single column.

Table 2. Label encoding

Categorical Variables	Label	Information
House type	0	House
	1	Featured house
	2	Premier house
Location	0	Banyumanik
	1	Candisari
	2	Gajah Mungkur
	3	Gayamsari
	4	Genuk
	5	Gunung Pati
	6	Mijen
	7	Ngaliyan
	8	Pedurungan
	9	Semarang Barat
	10	Semarang Selatan
	11	Semarang Tengah
	12	Semarang Timur
	13	Semarang Utara
	14	Tembalang
	15	Tugu

3.4.3 Skewness value

Skewness is a statistical measure used to assess the symmetry or asymmetry of a data distribution.

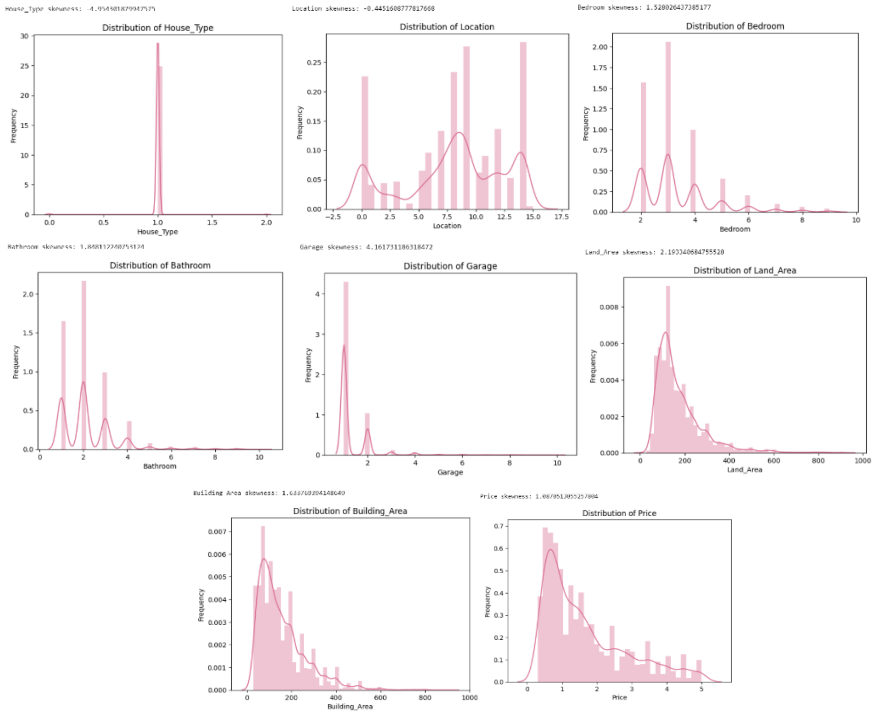

Figure 8. Skewness value

The house type, bedroom, bathroom, garage, land area, building area, and price variables have a positive skewness value or skewed to the right. Meanwhile, the location variable has a negative skewness value or skewed to the left.

3.5 Data Split

It is a stage of dividing data into training data and testing data. Training data is used to build a model. Meanwhile, testing data is used to test the performance of the model that has been trained. The comparison ratios used include 90%:10%, 80%:20%, 75%:25%, and 70%:30% with a total dataset of 9533.

3.6 Modeling

3.6.1 Multiple Linear Regression (MLR)

MLR models are more suitable for predicting linear data. Model definition using MLR using the parameters fit_intercept = True, copy_X = True, n_jobs = True, positive = False. MLR model is then trained with the .fit method using data training. Then, the MLR model will be tested for prediction using data testing with the .predict method.

3.6.2 Random Forest Regression (RFR)

The RFR model is more suitable for predicting non-linear data because it can better capture data variability. Model definition using RFR using parameters n_estimators = 200, criterion = 'friedman_mse', max_depth = None, min_samples_split = 2, min_samples_leaf = 1, max_features = 1.0, bootstrap = True, random_state = 0, and n_jobs = -1. RFR model is then trained with the .fit method using data training. Then, the MLR model will be tested for prediction using data testing with the .predict method.

3.7 Model Evaluation

3.7.1 Multiple Linear Regression (MLR)

The MLR model has performance results that are assessed based on the following evaluation metric values.

Table 3. MLR model evaluation

Ratio	Model evaluation		
	R-Squared	MSE	RMSE
90%:10%	60,5%	0,554	0,744
80%:20%	61,0%	0,533	0,730
75%:25%	**61,1%**	**0,547**	**0,739**
70%:30%	60,5%	0,547	0,739

The comparison ratio of training data of 75% and testing data of 25% resulted in the highest R Squared value (prediction accuracy) among other comparison ratios, which was 61.1% with an MSE value of 0.547 and an RMSE of 0.739. The MLR model has intercept and slope values that form the linear regression equation as follows.

$$y = 0.003402177953959251 + (0.1074744881200282 \times X1) + (-0.0017567874928904196 \times X2) + (-0.032665518913645025 \times X3) + (0.1562735834557301 \times X4) + (0.06362591765915206 \times X5) + (0.0037247407240702185 \times X6) + (0.004520382474806696 \times X7).$$

Where house type (X_1), location (X_2), bedroom (X_3), bathroom (X_4), garage (X_5), land area (X_6), and building area (X_7).

The following is a comparison Table of the actual price and the prediction price as well as the visualization of the MLR model.

	Actual	Predicted	Error
1842	0.645	0.866547	0.221547
1824	0.640	0.821477	0.181477
5622	1.500	1.972295	0.472295
2861	0.800	1.312936	0.512936
6651	1.990	1.948165	0.041835
2702	0.775	0.910790	0.135790
841	0.490	0.529375	0.039375
1536	0.595	0.885428	0.290428
1156	0.535	0.777509	0.242509
1478	0.585	1.008155	0.423155

Figure 9. Comparison of actual price and predicted price MLR model

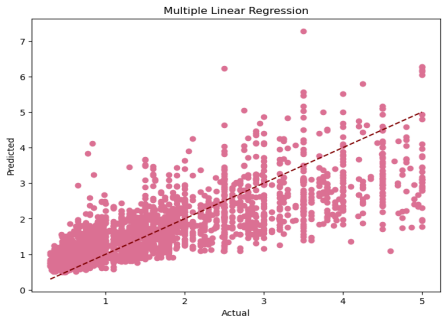

Figure 10. MLR Visualization

Predictions using the MLR model have a poor regression line and have a high difference between the actual price and predicted price. This is because the MLR model is not a suitable model for the data used because the data used has a distributed or non-linear data distribution.

3.7.2 Random Forest Regression (RFR)

The RFR model has performance results that are assessed based on the following evaluation metric values.

Table 4. Evaluation of the RFR model

Ratio	Model evaluation		
	R-Squared	MSE	RMSE
90%:10%	**78,4%**	**0,302**	**0,550**
80%:20%	75,2%	0,339	0,582
75%:25%	74,4%	0,360	0,600
70%:30%	74,2%	0,357	0,597

Comparison ratio Data Training by 90% and Data Testing by 10% resulting in the highest R-Squared value (prediction accuracy) among other divisions, which is 78.4% with an MSE value of 0.302 and RMSE of 0.550. The following is a Table comparing the original price with the predicted price and the visualization of the RFR model.

	Actual	Predicted	Error
1842	0.645	0.845018	0.200018
1824	0.640	0.651621	0.011621
5622	1.500	2.458300	0.958300
2861	0.800	1.550872	0.750872
6651	1.990	2.107658	0.117658
2702	0.775	0.755967	0.019033
841	0.490	0.432133	0.057867
1536	0.595	0.654346	0.059346
1156	0.535	0.485982	0.049018
1478	0.585	0.573742	0.011258

Figure 11. Comparison of actual price and predicted price RFR model

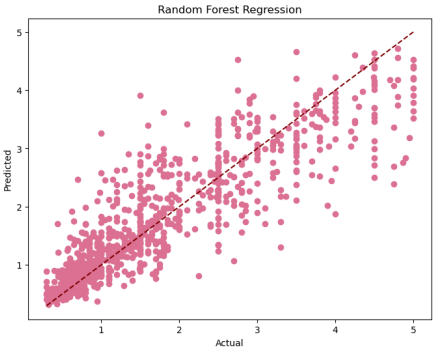

Figure 12. RFR Visualization

The prediction of the RFR model is quite good because it has a predicted price spread that is close to the identity line and has a lower difference between the original price and the prediction. This is because the RFR model is a suitable model for the data used because the data used has a distributed or non-linear data distribution.

3.7.3 Comparison of the performance of the two models

The following is a comparison Table of MLR and RFR model performance.

Table 5. Model performance comparison

Model	Ratio	Model evaluation		
		R-Squared	MSE	RMSE
MLR	75%:25%	61,1%	0,547	0,739
RFR	**90%:10%**	**078,4%**	**0,302**	**0,550**

Based on the table, the MLR model has a prediction accuracy result of 61.1% with an MSE of 0.547 and an RMSE of 0.739. The MSE and RMSE results in this model are quite large, indicating that there is a relatively high error value. Meanwhile, the RFR model has a prediction accuracy result of 78.4% with an MSE of 0.302 and an RMSE of 0.550. The MSE and RMSE results in this model are smaller compared to the MLR model, indicating that the error value is relatively lower.

Based on the explanation that has been explained, it can be concluded that the RFR model has higher prediction accuracy results with lower MSE and RMSE values compared to the MLR model. This is because

the RFR model has more parameters and is a model that can capture non-linear data. So that the RFR model is very suitable for the data used. Therefore, the RFR model will be chosen to enter the next stage.

3.7.4 Variable Importance

One of the outputs produced in the RFR model is the variable importance or the importance value of each independent variable in predicting the value of the dependent variable.

Table 6. Model Performance comparison

Variable	Importance value
Land area	0.582662
Building area	0.239076
Location	0.080150
Bathroom	0.041182
Bedroom	0.037755
Garage	0.018353
House type	0.000822

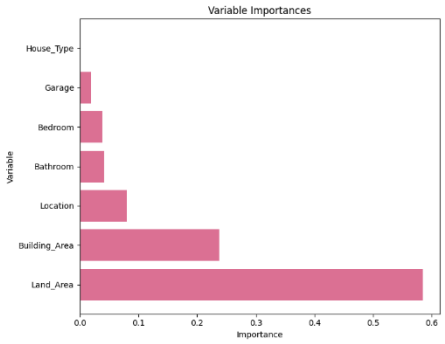

Figure 13. Variable importance

Based on the Table and Figure above, it is concluded that the importance value between land area and price has the highest result. This means that the variable of land area greatly contributes to the prediction of house prices. Meanwhile, the house type variable has the lowest importance value. This means that house type does not contribute much to the prediction of house prices

3.8 Deployment

3.8.2 Model inference

Model inference is the process of using a machine learning model that has been trained to make predictions or produce outputs based on new data that the model has never seen before. This stage is very important before the model is made into a website.

3.8.2 Setting up supporting files

The following are the files needed as supporting files to deploy to the streamlit web framework.

Table 7. Deployment support files

Files with the format	Fill
.csv	Data used from web scraping
.sav	Trained random forest regression model
.ipynb (notebook)	The entire modeling process from start to finish
.txt	Versions of the libraries used in this program
.png	Supporting photos for website display
.py	The model used, the things that will be displayed on the website page, and directly connected to the streamlit web framework

3.8.3 Upload files to GitHub repositories

Once all the files that need to be prepared are gathered, the next step is to create a new repository on the GitHub platform. Make sure to have a GitHub account before creating a new repository. The maximum limit for uploading files in GitHub repositories is 25 MB, while .sav files are 109 MB in size. Therefore, .sav file must be compressed first by creating .zip. After that, upload all the necessary files into the GitHub repository that was created earlier.

3.8.4 Log in to the streamlit page to create a website

Once everything is done, then open the https://share.streamlit.io/ to head to the streamlit framework website. On this page, click the "create app" section in the upper right corner then a display will appear as shown in the following image.

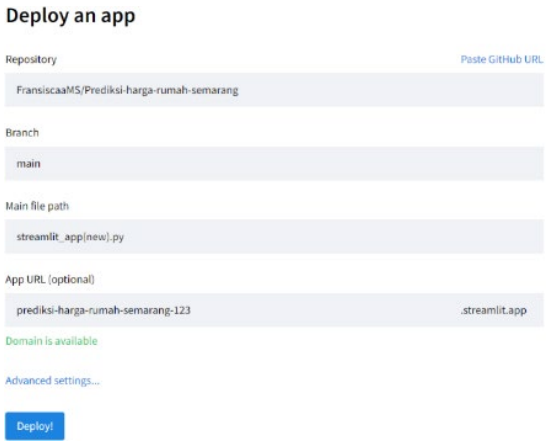

Figure 14. Streamlit page view

Description

Repository : The name of the GitHub repository that was created earlier.
Branch : Branch will be immediately adjusted by the system because it is already connected to the selected GitHub repository.
Main file path : The name of the file with the pre-created .py format.
App URL : The URL to be created, can custom the URL itself.

3.8.5 Website formation process

After everything is filled in, then click "Deploy!", then streamlit will process the program deployment. This process does not take too long, the progress of the deployment can also be monitored from the "Manage App" in the lower right corner to monitor whether there are errors or not. If there is an error on the streamlit website, then immediately fix the error by opening the file in .py format in the GitHub repository.

Then edit the file using the "edit file" feature so that files with the .py format can be repaired directly through GitHub. When done, save changes. Because GitHub is directly connected to streamlit, the streamlit website page will automatically redo the deployment process according to the update made to the file with the .py format earlier. Do this until there are no errors on the streamlit website.

3.8.6 Results of the house price prediction website

If there are no errors, then the website is complete. The streamlit website in this study has the following URLs https://house-price-prediction-semarang-city.streamlit.app/.When the deployment process is complete, the website will look like this.

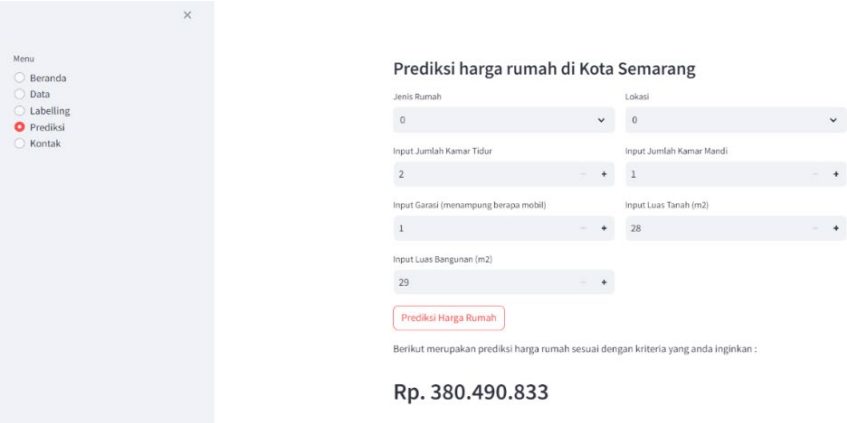

Figure 15. Website appearance

This website is created as a sidebar with several pages, namely the home page, data, labeling, predictions, and contacts. The home page contains greetings for website visitors and information related to the usability of the website. The data page contains a view of the data used. The labeling page contains related to the labeling of categorical data into numerical with the aim that website users know that the labeling will be used for prediction pages. Meanwhile, the prediction page contains variable inputs in accordance with the criteria of website users to predict house prices in Semarang City. The contact page is a page that contains the researcher's contact information so that if there is criticism and suggestions, can directly contact the available contacts.

4. CONCLUSIONS

The performance of the MLR and RFR models in predicting house prices in Semarang City can be seen from the value of the evaluation metrics. The MLR model produced the highest R-Squared (prediction accuracy) value of 61.1%, MSE of 0.547, and RMSE of 0.739 with a ratio of 75%:25% of training and testing data. The RFR model produced the highest R-Squared (prediction accuracy) value of 78.4%, MSE of 0.302, and RMSE of 0.550 with a comparison ratio of training and testing data of 90%:10%.

Therefore, it can be concluded that the model that has the best results is the RFR model to predict house prices in Semarang City.

The procedure for applying the RFR model to the streamlit web framework includes several steps, namely the model inference, setting up supporting files, upload files to GitHub repositories, log in to the streamlit page to create a website, website formation process, and results of the house price prediction website. This website is created as a sidebar with several pages, namely the home page, data, labeling, predictions, and contacts that are useful as general information. The results of the website can be useful for the community because it can be used by the public.

REFERENCES

[1] Pemerintah Kota Semarang, “Gambaran Umum Kota Semarang.” Accessed: Feb. 19, 2024. [Online]. Available: https://semarangkota.go.id/mainmenu/detail/profil

[2] BPS, “Luas Wilayah, Jumlah Penduduk, dan Kepadatan Penduduk (Jiwa/km2), 2020-2022.” Accessed: Feb. 19, 2024. [Online]. Available: https://semarangkota.bps.go.id/indicator/12/48/1/kepadatan-penduduk.html

[3] H. Haidir and I. Rudiarto, “Lahan Potensial Permukiman Di Kota Semarang,” Tataloka, vol. 21, no. 4, p. 575, 2019, doi: 10.14710/tataloka.21.4.575-588.

[4] Z. Faradilla Daldiri, M. Rafly, and I. Veritawati, “Clustering Daftar Harga Rumah di Jakarta Dengan Algoritma K-Means,” J. Informatics Adv. Comput., vol. 3, no. 2, 2022, [Online]. Available: https://www.kaggle.com/datasets/wisnuanggara/daf

[5] A. S. Arsandi, I. Ismiyati, and F. Hermawan, “Hubungan Pertumbuhan Penduduk dan Infrastruktur di Kota Semarang,” J. Karya Tek. Sipil, vol. 6, no. 4, pp. 15–29, 2017, [Online]. Available: https://www.neliti.com/publications/188593/hubungan-pertumbuhan-penduduk-dan-infrastruktur-di-kota-semarang

[6] E. F. Rahayuningtyas, F. N. Rahayu, and Y. Azhar, "Prediksi Harga Rumah Menggunakan General Regression Neural Network," J. Inform., vol. 8, no. 1, pp. 59–66, 2021, doi: 10.31294/ji.v8i1.9036.
[7] G. N. Ayuni and D. Fitrianah, "Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ," J. Telemat., vol. 14, no. 2, pp. 79–86, 2019, [Online]. Available: https://journal.ithb.ac.id/telematika/article/view/321
[8] N. Krzywinski, M., & Altman, "Multiple linear regression," Nat. Methods, pp. 12 12, 1103–4 ., 2015, doi: https://doi.org/10.4135/9781506326139.n453.
[9] G. Biau and E. Scornet, "A random forest guided tour," Test, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.
[10] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," Procedia Comput. Sci., vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.
[11] A. Putranto, N. L. Azizah, and I. R. I. Astutik, "Sistem Prediksi Penyakit Jantung Berbasis Web Menggunakan Metode Svm Dan Framework," J. Penerapan Sist. Inf. (Komputer Manajemen), vol. 4, no. 2, pp. 442–452, 2023, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
[12] J. Harlan, ANALISIS REGRESI LINEAR, 1st ed., vol. 2. Depok: Gunadarma, 2018.
[13] Suyanto, Machine learning tingkat dasar dan lanjut, 1st ed. Bandung: Informatika, 2018.
[14] A. Eka and A. Juarna, "Prediksi Pro duksi Daging Sapi Nasional dengan Meto de Regresi Linier dan Regresi Polinomial," J. Ilm. Komputasi, vol. 20, no. 2, pp. 209–215, 2021, doi: 10.32409/jikstik.20.2.2722.