p-ISSN: 2252-3006
e-ISSN: 2685-2411

Comparison Of Decision Tree, Linear Regression, and Random Forest Regressor Models for Predicting House Prices

Desak Nyoman Mulya Cahyani [a1], Ni Putu Sutramiani [a2], Kadek Indah Permata [a3], Dicky Umardhani [a4], Putu Yoga Pramana Putra [a5], Ni Wayan Emmy Rosiana Dewi [a6], I Wayan Agus Surya Darma [a7]
[a]Information Technology, Udayana University, Indonesia
e-mail: [2]sutramiani@unud.ac.id, [6]emmyrosiana@unud.ac.id, [7]suryarave@gmail.com

Abstrak

Rumah merupakan kebutuhan utama yang memberikan kenyamanan dan keamanan bagi penghuninya. Dalam konteks investasi, rumah juga menjadi pilihan yang menjanjikan karena dapat mengalami perubahan harga. Oleh karena itu, diperlukan sistem yang dapat memprediksi harga rumah bagi investor dan pembeli. Penelitian ini bertujuan untuk membandingkan kinerja tiga model prediksi harga rumah, yaitu decision tree, linear regression, dan random forest regressor. Dalam penelitian ini, digunakan dataset house price prediction dengan melakukan tahap eksplorasi data, pra-pemrosesan, pemodelan, dan perbandingan model. Hasil penelitian menunjukkan bahwa random forest regressor memberikan kinerja prediksi yang paling baik dengan metrik evaluasi yang lebih rendah, seperti MAE, MSE, RMSE, dan R2 Score, sehingga dapat dijadikan pilihan terbaik dalam memprediksi harga rumah dan memberikan informasi berharga bagi investor dan pembeli.

Kata kunci: *Prediksi Harga Rumah, Data Mining, Decision Tree, Linear Regression, Random Forest Regressor*

Abstract

A home is a basic requirement that offers comfort and security to its occupants. Because they are subject to price fluctuations, houses are also a potential option in an investing setting. As a result, buyers and investors require a system that can forecast house values. This study compares the effectiveness of decision trees, linear regression, and random forest regressors as models for predicting home prices. The dataset for predicting home prices was used in this study to conduct data exploration, pre-processing, modeling, and model comparison stages. The study's findings demonstrate that the random forest regressor offers the best prediction performance with lower assessment metrics, including MAE, MSE, RMSE, and R2 Score, making it the best option for predicting house prices and other financial outcomes.

Keywords: *House Price Prediction, Data Mining, Decision Tree, Linear Regression, Random Forest Regressor*

1. Introduction

A house is a basic requirement since it provides a place to live, a place to rest, and a place to escape the heat and rain. The house you own must match the comfort criteria so that the people who live in it feel at ease and safe.

One of the most pressing human needs is the desire to buy and possess a home. Furthermore, if the house is in a strategic location, the selling price will be affected. Property entrepreneurs will build or buy houses as a form of investment. Houses, like gold, can be utilized as a future investment because housing prices might fluctuate at any time. When investing, don't forget to forecast price movements to avoid losing money [1]. As a result, a system that can forecast property values for investors and home buyers is required.

In data mining, decision trees, linear regression, and random forest regressor models are frequently used to create predictions. However, no research has been conducted to compare the three models, particularly for predicting property values. As a result, the purpose of this study was to examine the performance and accuracy of decision tree, linear regression, and random forest regressor models for predicting house values.

According to Batubara D et al research, Prediction Analysis of Delays in Electricity Payments Using a Comparison of Decision Tree and Support Vector Machine Classification Methods, the date of salary receipt is the factor causing consumer delays in paying electricity bills. The C4.5 algorithm produced 70% accurate results, while SVM produced 67%. Based on this, it is possible to infer that the c4.5 algorithm is the best solution for solving the problems presented with the data listed [2].

Apriliah et al conducted a study that used the Random Forest Classification Algorithm to predict the possibility of diabetes in its early stages. The current research compares numerous approaches, including the SVM algorithm and Naive Bayes. The results of the comparison demonstrate that the Random Forest approach produces the greatest accuracy value, 97.88%. Based on this, it may be inferred that random forest classification can more correctly forecast the possibility of diabetes than other classification methods [3].

According to research by Prasojo et al, with his research, Prediction Analysis of the Feasibility of Giving Loans Using the Random Forest approach, the Random Forest approach has a performance of 83%. Based on this, it is possible to conclude that the Credit Feasibility Analysis Using General Random Forest (German Credit Data Case Study) was carried out effectively [4].

Heru Herwanto et al research on the use of linear regression algorithms to predict rice crop yields. The linear regression algorithm accounts for 94.51% of the variation in crop yield values depending on the independent variables measured, which are land area, seed variety, number of seeds, urea fertilizer, and Phonska NPK fertilizer. Meanwhile, the remaining 5.49% is influenced by various factors. The average RMSE accuracy value is 0.432. This demonstrates that a forecast model's fluctuation in values is near to accurate [5].

Haryanto et al research on comparing machine learning algorithms when predicting house prices, elements that can influence house prices include land area, building area, number of bedrooms, bathrooms, and garages. In this research, two algorithms, multiple linear regression, and random forest regressors, will be compared. The results showed that the random forest regressor method performed the best, with an accurate score of 81.6% [6].

In this research, we compare data mining models to predict house prices. The data mining models used are decision trees, linear regression, and random forest regressors.

2. Research Method

The quantitative-descriptive method was adopted in this research. The quantitative method is a systematic procedure of investigating a condition using data and facts in the form of numbers. The data used was collected from Kaggle.com as secondary data.

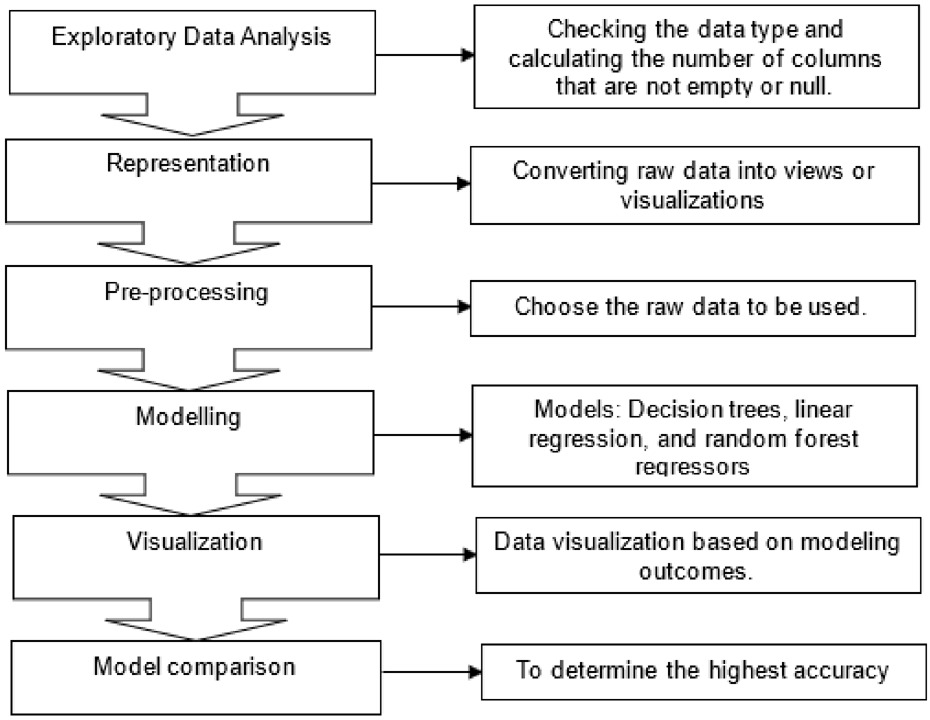

Figure 1. Flow of Research

The first step is exploratory data analysis, that is the process of investigating the contents of a dataset. Data exploration entails determining the kind or types of data in the dataset and estimating the number of values in each column that are not empty or null. The second one is representation. This stage converts the raw data into a display or visualization. The current research uses a heatmap to highlight the correlation between numerical variables and the relationship between each data feature and the goal data. Preprocessing is the stage in which data is prepared for use. This is the stage in which the raw data is sorted to match the needed data.

While modeling is the use of an algorithmic model to process data to draw conclusions from it. Decision trees, linear regression, and random forest regression are the algorithms used. Depending on the model utilized, this stage will have varied results. The values supplied are the MAE, MSE, RMSE, and R2 scores.

a. MAE is used to calculate the average absolute difference between anticipated and actual values.
b. MSE calculates the average squared error between predicted and actual data.
c. The square root of MSE is RMSE.
d. R2 is used to calculate how well the model explains fluctuations in the target variable. R2 has a value between 0 and 1. The closer R2 is to one, the better the model utilized.

Data visualization is carried out depending on the modeling results. Data can be visualized using boxplots, histograms, and other methods. After each model has produced data, a comparison process is performed to find the model with the highest accuracy.

3. Literature Study

3.1 Data Mining

Data mining is a process or stage in which fascinating patterns in interesting data are sought using certain approaches and models [7]. There are various models that can be used to predict house prices, including the decision tree model, linear regression, and random forest regressor. When modeling the links between attributes used in predicting property prices, each model takes a different approach and has different characteristics.

In general, the data mining function is used to figure out the kinds of patterns that exist inside the database. Two categories can be used to classify data mining functions [8]:

p-ISSN: 2252-3006
e-ISSN: 2685-2411

1. Descriptive.
   Descriptive features that explain the overall characteristics of the data in the database. More descriptive language is used to explain or characterize the data in the database.
2. Predictive.
   Predictive algorithms to forecast the information stored in the database.

3.2 Decision tree

Decision trees are algorithms that are often employed in decision making by converting criteria into interconnected nodes that form a tree structure [9]. A decision tree is a decision prediction model that has a hierarchical structure or tree that divides data based on separation criteria based on existing attributes. Each tree will have a branch representing an attribute that must be met to go to the next branch. This approach has advantages when dealing with non-linear interactions and provides a straightforward understanding of the elements that influence property prices.

3.3 Linear regression

Linear regression is a statistical strategy for predicting output results by establishing mathematical correlations between variables [10]. This strategy seeks a linear relationship between the input and goal variables. Linear regression is useful when the relationship between qualities and property prices is linear. This model gives a straight-line equation that describes the relationship between property prices and qualities.

3.4 Random Forest Regressor

When applying the regression method, a random forest regressor is an ensemble model composed of several decision trees [11][12][13][14][15]. A random forest is a model that combines several decision trees. Random forest regressors are commonly employed when the target variable is continuous or numeric and the purpose is to make continuous number predictions, such as house price predictions.

4. Result and Discussion

4.1. Datasets

The dataset used is a house price forecast dataset with 1460 rows and 81 columns. Kaggle was used to collect datasets. The dataset is separated into two parts: training data and test data, with 80% training data and 20% test data. The information includes several attributes that can affect the price of a house.

4.2. Data Preprocessing

The stage of preprocessing is carried out by selecting the house price dataset. The preprocessing phase functions to eliminate invalid data, inconsistent data, and duplicate data. The data before preprocessing is shown below.

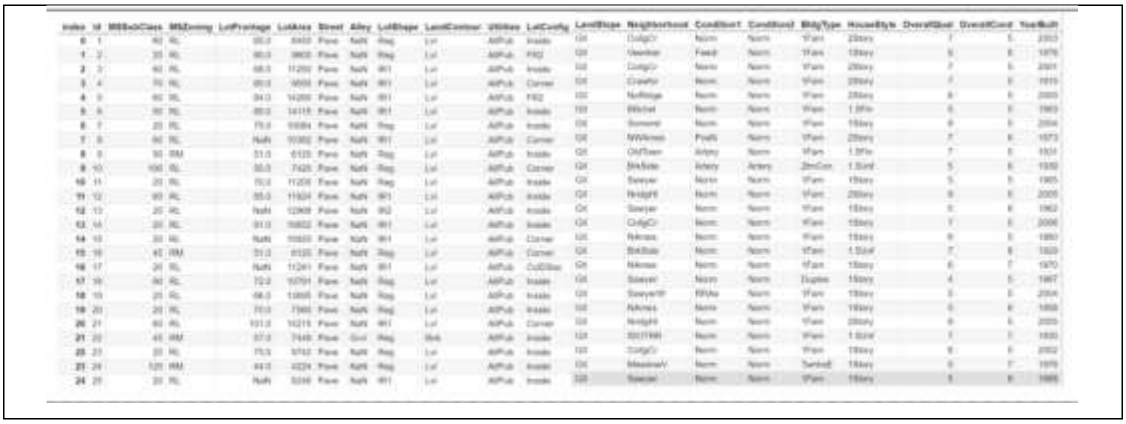

Figure 2 Data Before Preprocessing

p-ISSN: 2252-3006
e-ISSN: 2685-2411

Figure 2 shows data that has not been preprocessed; some of the data is not required for the research. The existing columns are as follows: ‘SalePrice’, ‘MSSubClass’, ‘MSZoning’, ‘LotFrontage’, ‘LotArea’, ‘Street’, ‘Alley’, ‘LotShape’, ‘LandContour’, ‘Utilities’, ‘LotConfig’, ‘LandSlope’, ‘Neighborhood’, ‘Condition1’, ‘Condition2:’, ‘BldgType’, ‘HouseStyle’, ‘OverallQual’, ‘OverallCond’, ‘YearBuilt’, ‘YearRemodAdd’, ‘RoofStyle’, ‘RoofMatl’, ‘Exterior1st’, ‘Exterior2nd’, ‘MasVnrType’, ‘MasVnrArea’, ‘ExterQual’, ‘ExterCond’, ‘Foundation’, ‘BsmtQual’, ‘BsmtCond’, ‘BsmtExposure’, ‘BsmtFinType1’, ‘BsmtFinSF1’, ‘BsmtFinType2’, ‘BsmtFinSF2’, ‘BsmtUnfSF’, ‘TotalBsmtSF’, ‘Heating’, ‘HeatingQC’, ‘CentralAir’, ‘Electrical’, ‘1stFlrSF’, ‘2ndFlrSF’, ‘LowQualFinSF’, ‘GrLivArea’, ‘BsmtFullBath’, ‘BsmtHalfBath’, ‘FullBath’, ‘HalfBath’, ‘Bedroom’, ‘Kitchen’, ‘KitchenQual’, ‘TotRmsAbvGrd’, ‘Functional’, ‘Fireplaces’, ‘FireplaceQu’, ‘GarageType’, ‘GarageYrBlt’, ‘GarageFinish’, ‘GarageCars’, ‘GarageArea’, ‘GarageQual’, ‘GarageCond’, ‘PavedDrive’ , ‘WoodDeckSF’, ‘OpenPorchSF’, ‘EnclosedPorch’, ‘3SsnPorch’, ‘ScreenPorch’, ‘PoolArea’, ‘PoolQC’, ‘Fence’, ‘MiscFeature’, ‘MiscVal’, ‘MoSold’, ‘YrSold’, ‘SaleType’, ‘SaleCondition’.

index	OverallQual	YearBuilt	YearRemodAdd	TotalBsmtSF	1stFlrSF	GrLivArea	FullBath	TotRmsAbvGrd	GarageCars	GarageArea	SalePric
0	7	2003	2003	856	856	1710	2	8	2	548	2085
1	6	1976	1976	1262	1262	1262	2	6	2	460	1815
2	7	2001	2002	920	920	1786	2	6	2	608	2235
3	7	1915	1970	756	961	1717	1	7	3	642	1400
4	8	2000	2000	1145	1145	2198	2	9	3	836	2500
5	5	1993	1995	796	796	1362	1	5	2	480	1430
6	8	2004	2005	1686	1694	1694	2	7	2	636	3070
7	7	1973	1973	1107	1107	2090	2	7	2	484	2000
8	7	1931	1950	952	1022	1774	2	8	2	468	1299
9	5	1939	1950	991	1077	1077	1	5	1	205	1180
10	5	1965	1965	1040	1040	1040	1	5	1	384	1295
11	9	2005	2006	1175	1182	2324	3	11	3	736	3450
12	5	1962	1962	912	912	912	1	4	1	352	1440
13	7	2006	2007	1494	1494	1494	2	7	3	840	2795
14	6	1960	1960	1253	1253	1253	1	5	1	352	1570
15	7	1929	2001	832	854	854	1	5	2	576	1320
16	6	1970	1970	1004	1004	1004	1	5	2	480	1490
17	4	1967	1967	0	1296	1296	2	6	2	516	900
18	5	2004	2004	1114	1114	1114	1	6	2	576	1590
19	5	1958	1965	1029	1339	1339	1	6	1	294	1390
20	8	2005	2006	1158	1158	2376	3	9	3	853	3253
21	7	1930	1950	637	1108	1108	1	6	1	280	1394
22	8	2002	2002	1777	1795	1795	2	7	2	534	2300
23	5	1976	1976	1040	1060	1060	1	6	2	572	1299
24	5	1968	2001	1060	1060	1060	1	6	1	270	1540

Figure 3 Data after preprocessing

Figure 3 shows the outcomes of the data selection process. The information used in this research includes: 'OverallQual', 'YearBuilt', 'YearRemodAdd', 'TotalBsmtSF', '1stFlrSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'GarageCars', 'GarageArea', 'SalePrice'.

The following data was not used in this study: ‘MSSubClass’, ‘MSZoning’, ‘LotFrontage’, ‘LotArea’, ‘Street’, ‘Alley’, ‘LotShape’, ‘LandContour’, ‘Utilities’, ‘LotConfig’, ‘LandSlope’, ‘Neighborhood’, ‘Condition1’, ‘Condition2:’, ‘BldgType’, ‘HouseStyle’, ‘OverallCond’, ‘RoofStyle’, ‘RoofMatl’, ‘Exterior1st’, ‘Exterior2nd’, ‘MasVnrType’, ‘MasVnrArea’, ‘ExterQual’, ‘ExterCond’, ‘Foundation’, ‘BsmtQual’, ‘BsmtCond’, ‘BsmtExposure’, ‘BsmtFinType1’, ‘BsmtFinSF1’, ‘BsmtFinType2’, ‘BsmtFinSF2’, ‘BsmtUnfSF’, ‘Heating’, ‘HeatingQC’, ‘CentralAir’, ‘Electrical’, ‘2ndFlrSF’, ‘LowQualFinSF’, ‘BsmtFullBath’, ‘BsmtHalfBath’, ‘HalfBath’, ‘Bedroom’, ‘Kitchen’, ‘KitchenQual’, ‘Functional’, ‘Fireplaces’, ‘FireplaceQu’, ‘GarageType’, ‘GarageYrBlt’, ‘GarageFinish’, ‘GarageQual’, ‘GarageCond’, ‘PavedDrive’ , ‘WoodDeckSF’, ‘OpenPorchSF’, ‘EnclosedPorch’, ‘3SsnPorch’, ‘ScreenPorch’, ‘PoolArea’, ‘PoolQC’, ‘Fence’, ‘MiscFeature’, ‘MiscVal’, ‘MoSold’, ‘YrSold’, ‘SaleType’, ‘SaleCondition’.

4.3 Exploratory Data Analysis

Exploratory data analysis is a method of examining data and deciding how to process it. This stage consists of checking for empty data, eliminating duplicate data, and transforming it.

 p-ISSN: 2252-3006
e-ISSN: 2685-2411

```
missing_values = df.isnull().sum()
print(missing_values)

OverallQual     0
YearBuilt       0
YearRemodAdd    0
TotalBsmtSF     0
1stFlrSF        0
GrLivArea       0
FullBath        0
TotRmsAbvGrd    0
GarageCars      0
GarageArea      0
SalePrice       0
dtype: int64
```

Figure 4 Check for Missing Values

Figure 4 shows the procedure of determining missing values. Because there is no empty data in the house price dataset, it can proceed to the next level.

```
[13] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   OverallQual   1460 non-null   int64
 1   YearBuilt     1460 non-null   int64
 2   YearRemodAdd  1460 non-null   int64
 3   TotalBsmtSF   1460 non-null   int64
 4   1stFlrSF      1460 non-null   int64
 5   GrLivArea     1460 non-null   int64
 6   FullBath      1460 non-null   int64
 7   TotRmsAbvGrd  1460 non-null   int64
 8   GarageCars    1460 non-null   int64
 9   GarageArea    1460 non-null   int64
 10  SalePrice     1460 non-null   int64
dtypes: int64(11)
memory usage: 125.6 KB
```

Figure 5 Data Type Information

Figure 5 shows data type information for 1460 records. The data type int64 is used in the columns OverallQual, YearBuilt, YearRemodAdd, TotalBsmtSF, 1stFlrSF, GrLivArea, FullBath, TotRmsAbvGrd, GarageCars, GarageArea, and SalePrice. There is no need to change the data type because the data that will be processed is already numeric.

4.4 Representation

The visualization stage of the housing price dataset follows the representation stage. The distribution of the data is shown in Figure 4.

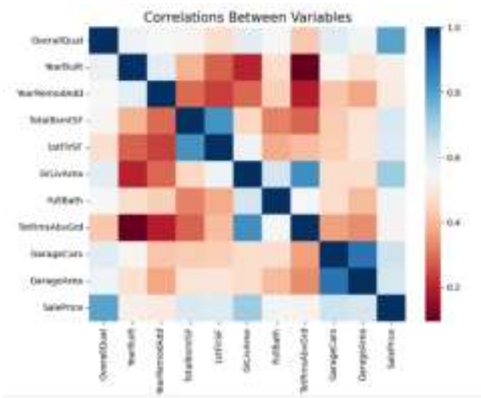

Figure 6 Correlation Matrix

The dark color of the correlation coefficient in Figure 6 suggests a strong link. Bright hues, on the other hand, suggest a negative linear relationship or no correlation between variables.

4.5 Decision Tree

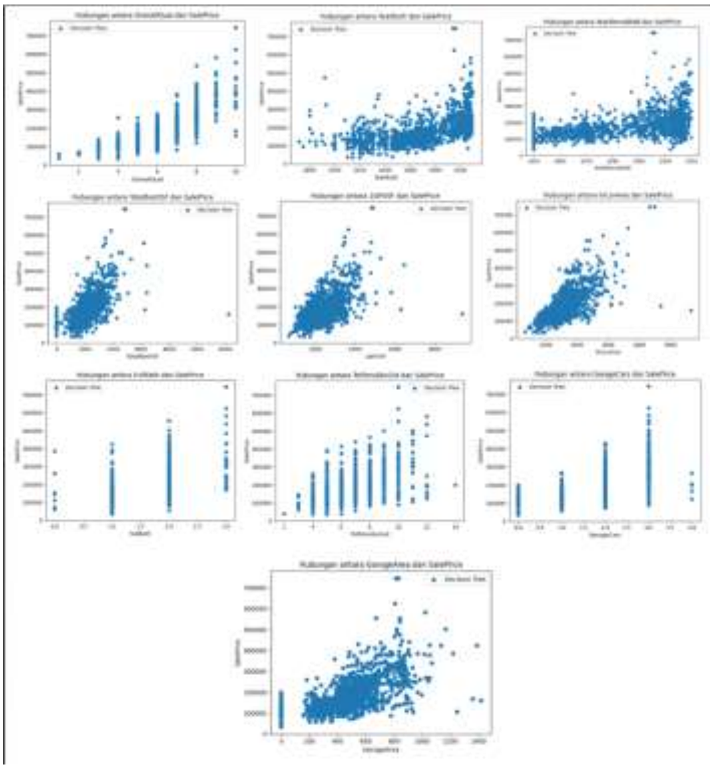

Figure 7 Decision Tree Model Visualization

Figure 7 shows the results of testing the decision tree model. The mean absolute error in the test was 24709.119863013697, and the root mean square error (RMSE) was 36610.36020458363.

4.6 Linear Regression

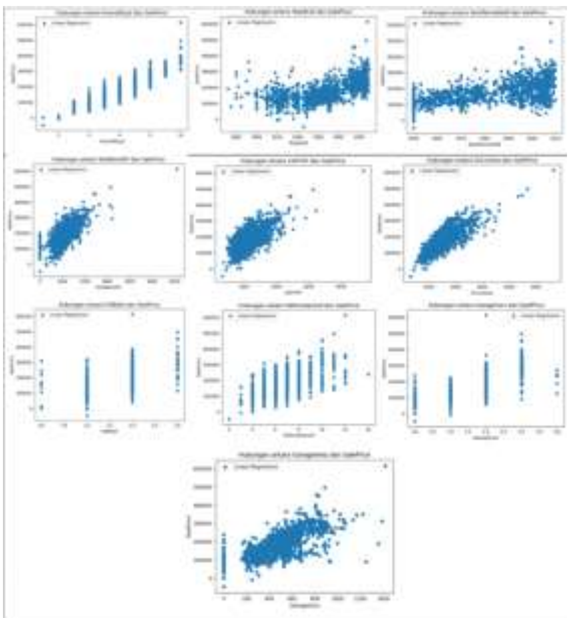

Figure 8 Linear Regression Model Visualization

Figure 8 shows the results of testing the linear regression model. The mean absolute error in the test was 24774.219519604765, and the root mean square error (RMSE) was 39474.54338116007.

4.7 Random Forest Regressor

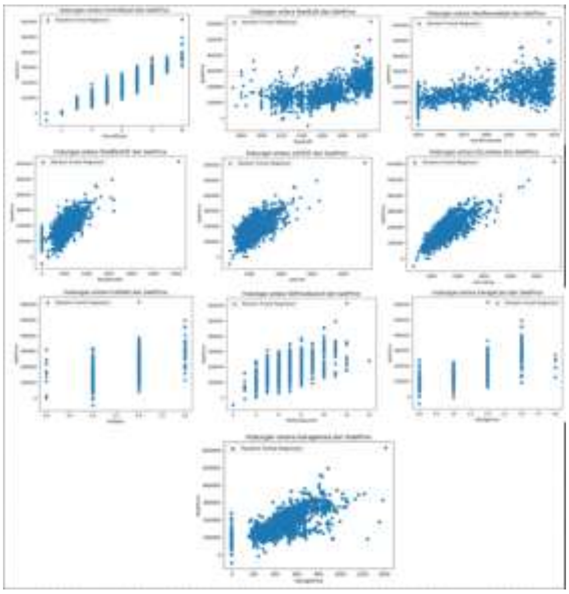

Figure 9 Random Forest Regressor Model Visualization

Figure 9 shows the results of testing the random forest regressor model. The mean absolute error in the test was 18907.850671885193, and the root mean square error (RMSE) was 29362.523197914874.

4.8 Model Comparison

Table 1 Comparison of Three Model Predictions

Model	MAE	MSE	RMSE	R2 Score	RMSE (Cross-Validation)
Random Forest Regressor	18672.378190	8.482698e+08	29125.071965	0.889409	32000.207212
Linear Regression	24774.219520	1.558240e+09	39474.543381	0.796848	38573.182562
Decision Tree	24750.934361	1.420956+09	37695.575371	0.814746	42159.674232

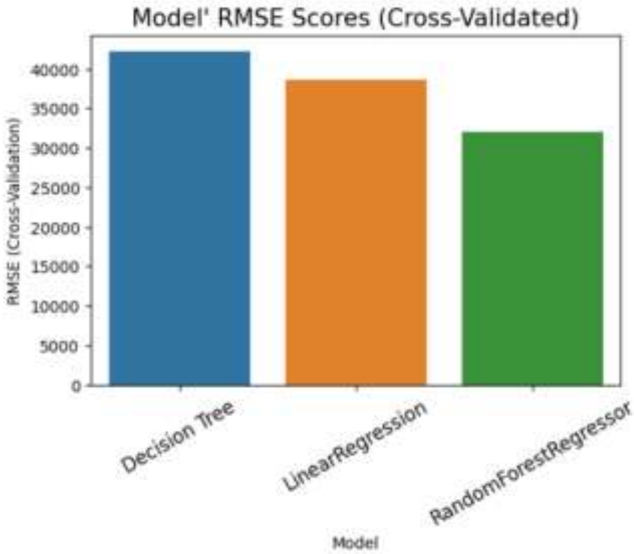

Figure 10 Comparative visualization of three model predictions

Table 1 and Figure 10 show the comparison results from the three models. Each model is explained in greater depth below:

a. The MAE value of the Random Forest Regressor model is 18672.378190, the MSE value is 8.482698e+08, the RMSE value is 29125.071965, and the R2 value is 0.889409.
b. The linear regression model's MAE is 24774.219520, its MSE is 1.558240e+09, its RMSE is 39474.543381, and its R2 is 0.796848.
c. The MAE value of the Decision Tree model is 24978.369863, the MSE value is 1.420956e+09, the RMSE value is 37695.575371, and the R2 value is 0.814746.

5. Conclusion

Several data mining algorithms, including decision trees, linear regression, and random forest regressors, are used to predict house prices. According to the findings of the tests performed on the three models, the random forest regressor model had the best predictive performance for predicting house prices. This is evident from the lowest values of numerous assessment metrics, such as MAE, MSE, and RMSE in the random forest regressor model. Meanwhile, for the R2 Score matrix, the random forest regressor has the highest value, which is near to 1, indicating that it can explain more variances in the data.

References

[1] E. Febrion Rahayuningtyas, F. Novia Rahayu, Y. Azhar, and I. Artikel, "Prediksi Harga Rumah Menggunakan General Regression Neural Network," *JURNAL INFORMATIKA*, vol. 8, no. 1, 2021, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Real

[2] D. N. Batubara, A. P. Windarto, and E. Irawan, "Analisis Prediksi Keterlambatan Pembayaran Listrik Menggunakan Komparasi Metode Klasifikasi Decision Tree dan Support Vector Machine," *JURIKOM (Jurnal Riset Komputer)*, vol. 9, no. 1, p. 102, Feb. 2022, doi: 10.30865/jurikom.v9i1.3833.

[3] W. Apriliah *et al.*, "Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest," 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[4] B. Prasojo and E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 2, pp. 79–89, Sep. 2021, doi: 10.25077/teknosi.v7i2.2021.79-89.

[5] H. W. Herwanto, T. Widiyaningtyas, and P. Indriana, "Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi," 2019.

[6] C. Haryanto, N. Rahaningsih, and F. Muhammad Basysyar, "KOMPARASI ALGORITMA MACHINE LEARNING DALAM MEMPREDIKSI HARGA RUMAH," 2023.

[7] M. A. Rahman, “Penerapan Metode Rough Set Dalam Memprediksi Penjualan Perumahan (Studi Kasus Di Pt. Anugerah Pasadena Pekanbaru),” *Jurnal Warta Dharmawangsa*, vol. 14, no. 2, pp. 342–355, 2020, doi: https://doi.org/10.46576/wdw.v14i2.632.

[8] Y. Mahena, M. Rusli, and E. Winarso, “Prediksi Harga Emas Dunia Sebagai Pendukung Keputusan Investasi Saham Emas Menggunakan Teknik Data Mining,” *KALBI SCIENTIA*, vol. 2, no. 1, 2015.

[9] F. Yulian Pamuji, V. Puspaning Ramadhan, and R. Artikel, “Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy,” vol. 7, pp. 46–50, 2021, [Online]. Available: http://http://jurnal.unmer.ac.id/index.php/jtmi

[10] A. Saiful, S. Andryana, and A. Gunaryati, “Prediksi Harga Rumah Menggunakan Web Scrapping dan Machine Learning dengan Algoritma Linear Regression,” 2012, [Online]. Available: http://jurnal.mdp.ac.id

[11] I. L. Mulyahati, “Implementasi Machine Learning Prediksi Harga Sewa Apartemen Menggunakan Algoritma Random Forest Melalui Framework Website Flask Python (Studi Kasus: Apartemen di DKI Jakarta Pada Website mamikos. com),” Undergraduate Thesis, Universitas Islam Indonesia, 2020.

[12] M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, “Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments,” ISPRS International Journal of Geo-Information, vol. 7, no. 5, p. 168, May 2018

[13] H. Tyralis and G. Papacharalampous, “Variable Selection in Time Series Forecasting Using Random Forests,” Algorithms, vol. 10, no. 4, p. 114, Oct. 2017

[14] C. Iwendi et al., “COVID-19 patient health prediction using boosted random forest algorithm,” Frontiers in Public Health, vol. 8, Jul. 2020.

[15] P. J. Moore, T. J. Lyons, and J. Gallacher, “Random forest prediction of Alzheimer’s disease using pairwise selection from time series data,” PLoS ONE, vol. 14, no. 2, Feb. 2019.