International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 6 No. 1, September 2024

Feature Engineering Optimization on the Performance of XGBoost, Random Forest, and Support Vector Regression Algoritms in House Price Prediction

Brahmantio Widyo Trenggono[1]
Department of Information Technology
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
brahmantio25@gmail.com

I Gede Susrama Mas Diyasa[2]
Department of Information Technology
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
igsusrama.if@upnjatim.ac.id

Ani Dijah Rahajoe[3]
Department of Information Technology
Faculty of Computer Science
Universitas Pembangunan Nasional Veteran Jawa Timur
Surabaya, Indonesia
adrahajoe.if@upnjatim.ac.id

***Abstract*— As the years go by, the ever-increasing movement of house prices has become an important factor in investment decisions and financial planning to curb inflation. However, fluctuations or increases in house prices can be caused by various factors that can affect the value of house price predictions. This study aims to analyze the influence of optimization and the relationship between feature engineering and modeling in house price predictions. The research stages include data preprocessing, logarithmic transformation, feature engineering, data splitting, and optimization in determining parameters during tuning. Model performance is evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Determination coefficient (R^2-Squared) metrics. The results show that the Support Vector Regression algorithm produces the best performance with a MAE value of 274 million, an RMSE of 780 million, a MAPE of 7%, and an R^2-Squared of 98%. This research is expected to serve as a reference for future studies on regression model optimization, particularly in decision-making for more accurate house price predictions.**

Keywords—Prediction Price House, Optimization, Feature Engineering

I. INTRODUCTION

The growing knowledge and insight among the younger generation, especially Generation Y or Generation Z who are just entering the workforce, has had a significant impact on their perceptions or views of property, particularly homes. This younger generation is beginning to realize the importance of owning a home as both a primary need and a future investment. However, increasingly complex economic challenges, such as rising living costs, market fluctuations, and limited income, mean that they must be more careful in planning their finances, including their plans to buy a home

(Wadani et al., 2023). According to Dewabrata et al. 2023, the number of housing units is becoming increasingly limited each year, making it difficult for people to achieve their desire to buy a home.

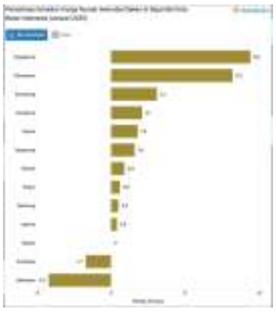

Figure I Source: databoks.katadata.co.id

Based on Figure 3 sourced from katadata.co.id, January 2025 shows several price increases at the beginning of the year, with a percentage of 9.4% in Yogyakarta, 8.2% in Denpasar, and 3.1% in Semarang. In addition, significant population growth, accompanied by numerous housing developments by the government and developers, has created various options for people to own homes (Fuadah et al. 2020).

The increase in house prices is driven by several factors, including population growth, increased public demand for housing, limited land availability, rising building material prices, and so on. Therefore, one solution that can help prospective home buyers is the use of machine learning algorithms to predict house prices by considering several factors that can affect those prices. According to Fuadah et al. (2020), the use of prediction algorithms not only helps determine home prices that are in line with consumers' financial capabilities, but also increases their confidence in investing in the property sector.

Therefore, careful analysis is needed to accurately predict house price movements so that prospective buyers can avoid potential losses and choose the most profitable property (Siregar et al., 2023). With the implementation of Random Forest, XGBoost, and Support Vector Regression algorithms, predicting house prices can be done more accurately based on available data (Nuzurialini, 2024).

Then, hyperparameter tuning in this study uses Random Search and Bayesian Optimization, where experiments on both were chosen because Random Search can explore more combinations in a short time, even though the combinations are made randomly, Random Search can often find combinations of hyperparameters with optimal results (Ramadhan et al. 2024). Bayesian Optimization was chosen because it can efficiently find the optimal hyperparameter combinations for Random Forest without having to try all possible combinations by building a probabilistic model, often referred to as Gaussian Process Regression, which predicts which hyperparameters can provide better results (Tuner et al. 2021).

The author hopes that this research will contribute to the development of models with optimal accuracy in predicting house prices, given the fluctuating dynamics of house price increases, while also helping the younger generation make more informed financial decisions.

II. METHOD

A. *Optimization Flow*

This research began with careful data preprocessing to ensure optimal data quality. The initial process involved data cleaning, which consisted of detecting and correcting incorrect, inaccurate, or missing data so that the dataset was ready for analysis. Next, Exploratory Data Analysis (EDA) was conducted to understand the characteristics of the data, find patterns, detect outliers, and check assumptions through descriptive statistics and visualization.

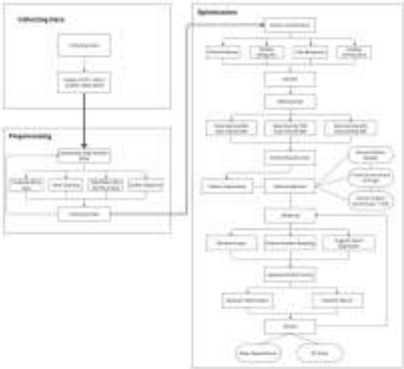

Figure II Block Diagram of Optimization

The next stage is feature transformation and feature combination. Transformation includes data standardization using Standard Scaler and conversion of categorical variables with Label Encoding. Feature combination is done by forming new variables such as space efficiency, interior quality, building age, and infrastructure quality to improve the predictive ability of the model. In addition, feature importance analysis is also performed to select the best features using three schemes: all features, the five features with the highest contribution, and features with an importance above 95%.

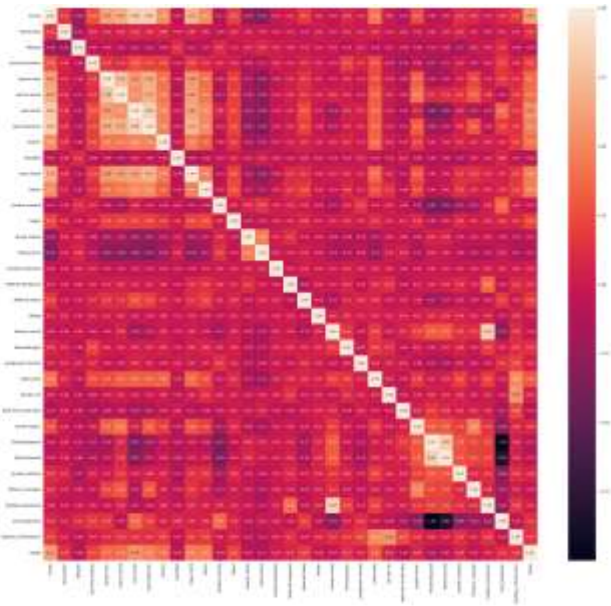

Figure III Multicolinarity Check

To improve performance, this study applies K-Fold Cross Validation as a validation method, so that the model can be tested consistently on various data subsets. Furthermore, hyperparameter tuning is performed using Random Search and Bayesian Optimization. Random Search explores parameter combinations randomly with computational efficiency, while Bayesian Optimization utilizes probabilistic models to adaptively search for optimal parameters.

Model evaluation was performed using three metrics, namely Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-Squared (R^2). These metrics were used to assess the accuracy of predictions and the model's ability to explain data variability. The evaluation results are expected to show the contribution of the preprocessing stage, feature selection, data partitioning scheme, and tuning method to improving the model's performance in predicting house prices.

B. *Feature Engineering*

In this study, feature combination was performed to create new features that are more representative based on the initial analysis results of the structure and correlation between variables. From the 11 initial features, five new features were formed, namely room efficiency, building quality, building age, and infrastructure quality. These features are designed to capture more relevant information and simplify data complexity, so that they are expected to improve the accuracy and efficiency of predictions.

The next stage was feature selection, which plays an important role in optimizing model performance by using only the features most relevant to the target variable. Feature selection was carried out using the feature importance values from the Random Forest and XGBoost algorithms. Three scenarios were tested, namely the use of all features, the selection of the five features with the highest contribution, and the selection of features with an importance value above 96%. This strategy allows the model to focus on important variables while reducing noise and computational load.

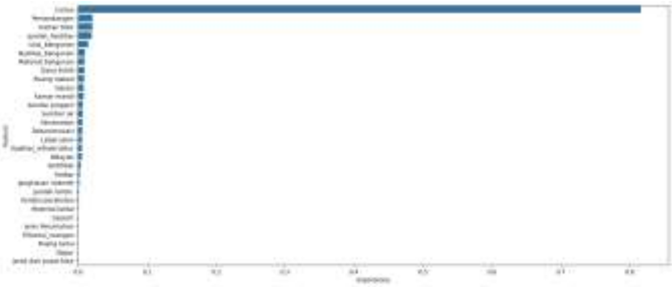

Figure IV Feature Importance

Next, feature importance analysis was conducted to identify the variables that most influenced the prediction results. Based on the calculation results, the features of installments, number of facilities, building area, view, and building age emerged as the dominant variables. In addition to the individual approach, this study also applied accumulative feature importance to measure the cumulative contribution of features. Thus, only features with a collective contribution above 94% were retained in the modeling to maintain a balance between complexity and prediction quality.

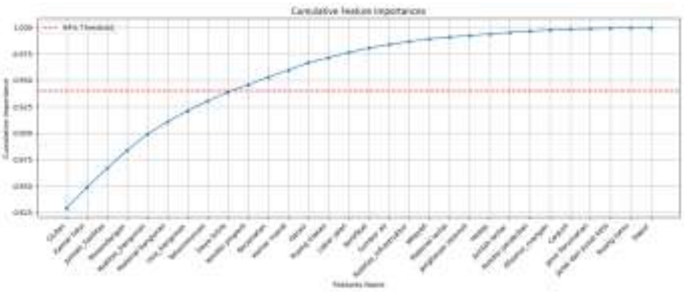

Figure V Accumulative Importance

Finally, at the feature transformation stage, data distribution normalization was performed through log transformation. This step was necessary because a number of numerical features showed uneven distributions and contained extreme outliers, for example in property prices, installments, land area, and electricity power. This transformation helped normalize the distribution, reduce the influence of extreme values, and improve the stability and accuracy of the model in making predictions.

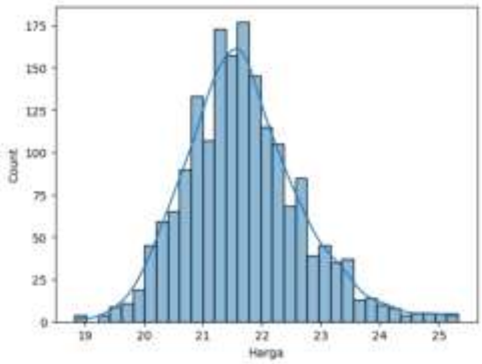

Figure VI Data training log transform

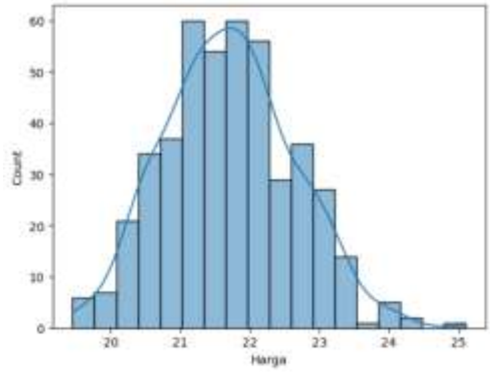

Figure VII Data testing log transform

III. RESULT

In this study, three machine learning algorithms, namely Random Forest Regression, Extreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR), were implemented through nine experiments for each algorithm. The experiments were designed based on a combination of data division schemes (80:20, 70:30, and 60:40) and three feature selection approaches, namely the use of all features, the five features with the highest importance, and features with an importance accumulation above 94%. This strategy allowed for an in-depth evaluation of the influence of variations in the amount of training data and feature quality on model performance.

Table I Accuracy Results MAE, RMSE

Algoritma	Splitting Data	Feature Selection	Sumber	Parameter	MAE	RMSE
Random Forest	70:30	Skema 3	Zamzam et al (2024)	Bayesian Optimization n_estimators=37, max_depth=16, min_samples_split=6, min_samples_leaf=5, max_features=0.999	321,960,456.42	937,630,669.13
XGBoost	70:30	Skema 1	Hakyemez et al (2024)	Random Search colsample_bytree: 0.78, learning_rate: 0.034, max_depth: 3, min_child_weight: 3, n_estimators: 741, subsample: 0.91	392,757,034.97	935,432,180.53
Support Vector Regression	70:30	Skema 1	Chaibi et al (2022)	Bayesian Optimization C=7, Epsilon=0.09, gamma = 0.004	274,749,625.03	780,591,247.15

Table II Accuracy Results MAPE, R-Squared, Cross Validation

Algoritma	Splitting Data	Feature Selection	Sumber	MAPE	R-Squared	Cross Validation
Random Forest	70:30	Skema 3	Zamzam et al (2024)	7.83	0.97	0.90
XGBoost	70:30	Skema 1	Hakyemez et al (2024)	12.68	0.97	0.89
Support Vector Regression	70:30	Skema 1	Chaibi et al (2022)	7.09	0.98	0.90

The optimization results show that Random Forest performs best in a 70:30 data split scenario with an accumulative feature scheme >94% and an error percentage of 7.83%. Under these conditions, the model produces an R^2 value of 0.97 with a validation of 0.90, with an error value obtained at MAE of 321 million and RMSE of 937 million.

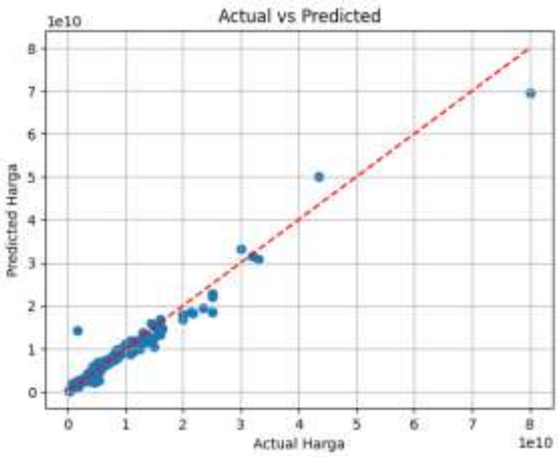

Figure VIII Scatterplot Accuracy Results Random Forest

The XGBoost algorithm also showed high performance, especially in the 70:30 data split scenario with the use of all features. Optimization through a combination of the colsample_bytree,learning_rate,max_depth,min_child_weight, n_estimators, and subsample parameters resulted in an R^2 of 0.97 with a validation of 0.89. However, the MSE value of 392 million and RMSE of 935 million indicate that there are a number of predictions with large errors, even though the model's accuracy remains competitive with a MAPE of 12.68%.

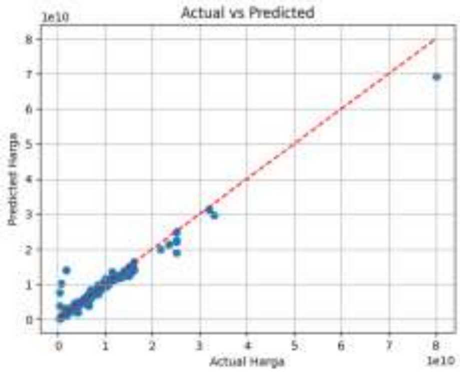

Figure IX Scatterplot Accuracy XGBoost

Meanwhile, Support Vector Regression (SVR) is recorded as the algorithm with the best performance, as seen in the error percentage in MAPE of 7.09%, which has the smallest error value among other algorithms. In the 70:30 scenario with all features, SVR produced an R^2 of 0.98 with a validation of 0.90, as well as consistency in accuracy, indicating excellent generalization capabilities. Although there was a significant gap between the MAE of 274 million and the RMSE of 780 million, these results still showed model stability without any indication of significant overfitting.

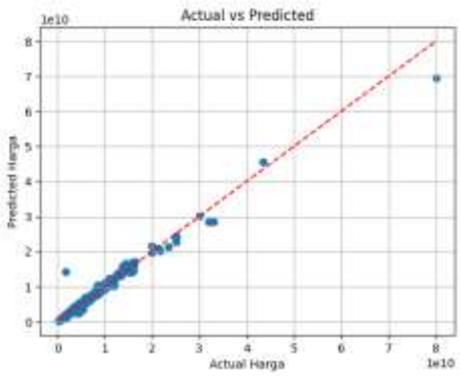

Figure X Scatterplot Accuracy Results SVR

Overall, the comparison of the three algorithms' performance indicates that SVR is the leading model for predicting house prices in the 2024–2025 Surabaya dataset. This advantage is supported by the application of Bayesian Optimization in hyperparameter tuning, which allows SVR to capture data patterns with greater precision while maintaining a balance between accuracy and generalization capabilities.

IV. Conclusion

This study confirms that optimization through feature engineering plays a crucial role in improving the accuracy of house price predictions. Techniques such as feature encoding, feature combination, and log transformation have been proven to improve model performance by capturing more relevant information and reducing data complexity. In addition, testing on datasets with a wide price range shows that model performance is not only determined by algorithms and parameters, but is also greatly influenced by the characteristics of the dataset relative to the algorithm used.

More specifically, Random Forest has proven to be quite robust against outliers and provides interpretability through feature importance, but its performance is still influenced by data distribution and sample splitting strategies. XGBoost has the strength to handle complex datasets by utilizing residual error, although it requires more careful parameter tuning and longer computation time. Meanwhile, Support Vector Regression (SVR) is highly sensitive to scaling, splitting, and kernel selection processes, but when properly optimized, it can deliver simple yet highly effective results, especially on medium-sized datasets.

Compared to the baseline without optimization, the combination of feature engineering, feature selection, and hyperparameter tuning techniques consistently provides significant performance improvements. Thus, this study successfully answers the research question and provides recommendations for regression model optimization strategies, where SVR proves to be the most effective algorithm compared to Random Forest and XGBoost for supporting house price decision making.

Acknowledgment

The author would like to thank colleagues who have supported the development of this research, especially those involved in data collection, as well as the academic advisor who provided invaluable knowledge throughout this research. This work would not have been possible without their support and input.

References

[1] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, “House Price Prediction using Random Forest Machine Learning Technique,” *Procedia Computer Science*, vol. 199, pp. 806–813, Jan. 2022, doi: 10.1016/j.procs.2022.01.100.

[2] Jain, Rajput, Garg, and Chawla, Eds., *Prediction of house pricing using machine learning with Python*. Proceedings of the International Conference on Electronics and Sustainable Communication Systems, 2020.

[3] ., N. Rais, A. Robi, W. Hadi, and Kurniawan, “ANALISA PREDIKSI HARGA JUAL RUMAH MENGGUNAKAN ALGORITMA RANDOM FOREST MACHINE LEARNING,” JURSISTEKNI (Jurnal Sistem Informasi dan Teknologi Informasi), May 2024.

[4] E. Fitri, *Analisis perbandingan metode regresi linier, random forest regression dan gradient boosted trees regression Method untuk prediksi harga rumah*, vol. Vol. 4. Journal of Applied Computer Science and Technology (JACOST), 2023. [Online]. Available: http://journal.isas.or.id/index.php/JACOST

[5] C. Haryanto, N. Rahaningsih, and F. M. Basysyar, *KOMPARASI ALGORITMA MACHINE LEARNING DALAM MEMPREDIKSI HARGA RUMAH*, vol. Vol. 7. JATI (Jurnal Mahasiswa Teknik Informatika), 2023.

[6] J. Hong, H. Choi, and W.-S. Kim, “A HOUSE PRICE VALUATION BASED ON THE

RANDOM FOREST APPROACH: THE MASS APPRAISAL OF RESIDENTIAL PROPERTY IN SOUTH KOREA," *International Journal of Strategic Property Management*, vol. 24, no. 3, pp. 140–152, Feb. 2020, doi: 10.3846/ijspm.2020.11544.

[7] L. Breiman, "Random forests," *Kluwer Academic Publishers*, vol. 45, pp. 5–32, 2001.

[8] M. Ahtesham, N. Zakaria Bawany, and K. Fatima, "House price prediction using machine learning algorithm - the case of Karachi City, Pakistan," *International Arab Conference on Information Technology (ACIT)*, 2021.

[9] Zhang, Huang, Zhang, Liu, and Shorman, "Analysis and prediction of second-hand house price based on random forest," *Applied Mathematics and Nonlinear Sciences*, 2022, doi: 10.2478/AMNS.2022.1.00052.

[10] C. Fan, Z. Cui, and X. Zhong, "House prices prediction with machine learning algorithms," *Association for Computing Machinery.*, pp. 6–10, Feb. 2018, doi: 10.1145/3195106.3195133.

[11] N. H. Zulkifley, S. A. Rahman, N. H. Ubaidullah, and I. Ibrahim, "House Price Prediction using a Machine Learning Model: A Survey of Literature," *International Journal of Modern Education and Computer Science*, vol. 12, no. 6, pp. 46–54, Dec. 2020, doi: 10.5815/ijmecs.2020.06.04.

[12] Tanamal, Minoque, Wiradinata, Soekamto, and Ratih, "House price prediction model using random forest in Surabaya City," *TEM Journal*, vol. 12, no. 1, pp. 126–132, 2023, doi: 10.18421/TEM121-17.

[13] J. Bergstra and Y. Bengio, "Random search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, Dec. 2012, [Online]. Available: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

[14] X. Ouyang and Shanghai Normal University, "House price prediction based on machine learning models," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 870–871, 2024.

[15] M. Modi, A. Sharma, and P. Madhavan, "Applied research on house price prediction using diverse machine learning techniques," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, no. 04, pp. 371–372, Apr. 2020, [Online]. Available: https://www.ijstr.org/final-print/apr2020/Applied-Research-On-House-Price-Prediction-Using-Diverse-Machine-Learning-Techniques.pdf

[16] Sarjana, A. Meitriana, and W. Suwendra, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI TINGKAT HARGA PERUMAHAN DI KABUPATEN BULELENG," *Jurnal Pendidikan Ekonomi Undiksha*, vol. 10, no. 2, 2018.

[17] N. V. Setia, P. A. Bella, and L. J. Tjung, "FAKTOR – FAKTOR LOKASI YANG MEMPENGARUHI HARGA JUAL RUMAH DI KECAMATAN SAWANGAN KOTA DEPOK," *Jurnal Sains Teknologi Urban Perancangan Arsitektur (Stupa)*, vol. 6, no. 1, pp. 669–684, Apr. 2024, doi: 10.24912/stupa.v6i1.27509.

[18] A. Fuadah, A. M. Siregar, and Y. Cahyana, "Model Prediksi Harga Rumah Di Kabupaten Bandung Menggunakan Multiple Linear Regression Dan Support Vector Regression," *Scientific Student Journal for Information, Technology and Science*, p. 10, Jul. 2024.

[19] A. A. Bastian, H. H. Handayani, D. Wahiddin, T. Rohana, and Teknik Informatika, Universitas Buana Perjuangan, Karawang, Indonesia, "Implementasi Algoritma Support Vector Regression dan Linear Regression Untuk Prediksi Harga Rumah," *Progresif: Jurnal Ilmiah Komputer*, pp. 961–972, Aug. 2024.

[20] D. Riando, A. Afiyati, Universitas Mercu Buana, and Indonesia, "IMPLEMENTATION OF XGBOOST ALGORITHM TO PREDICT THE SELLING PRICE OF CAYENNE PEPPERS IN DKI JAKARTA," *Eduvest – Journal of Universal Studies*, vol. 4–4, no. 9, pp. 7878–7879, Sep. 2024, [Online]. Available: http://eduvest.greenvest.co.id