Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

Comparative Analysis of Artificial Neural Network (ANN) and Random Forest in House Price Prediction

Yulia Retnowati, Abid Mafahim, Indira Syawanodya

Software Engineering, Universitas Pendidikan Indonesia, Indonesia
Correspondence: E-mail: yulia.retnowati@upi.edu

ABSTRACT

With the advancement of information technology, the application of machine learning in the property industry, particularly for house price prediction, has become increasingly important. Technology plays a crucial role in speeding up and enhancing the accuracy of property buying and selling processes. Therefore, the role of machine learning technology can be utilized to meet the need for improving the accuracy of house price predictions in major cities of developing countries, such as Bandung. This research aims to analyze the effectiveness of the Artificial Neural Network and Random Forest algorithms in predicting house prices in Bandung. The data used includes house sales data in Bandung, covering land area, building area, number of bedrooms, number of bathrooms, number of parking spaces, and the subdistrict location. The analysis of the algorithms is conducted by comparing the performance testing results of both algorithms using performance metrics for regression models such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Square (R2). Additionally, this research analyzes which data ratio among the training, validation, and test data yields the best results. The research findings indicate that the model with a data ratio of 60:20:20 produces the best performance for both algorithms. The Random Forest algorithm demonstrates superior performance with results of MAE: 0.0470; MSE: 0.0079; RMSE: 0.0888; and R2: 0.7085.

ARTICLE INFO

Article History:
Submitted/Received
15 August 2024
First Revised 08 September 2024
Accepted 10 November 2024
First Available online
15 December 2024
Publication Date
15 December 2024

Keyword:
Machine Learning,
House Price Prediction,
Neural Network,
Random Forest.

1. INTRODUCTION

The decision regarding house prices and buyers' decisions in purchasing a house are influenced by various factors. According to Kurniawan, et al. (2020) and Zulkifli and Ismail (2023), physical factors such as structure are among the most influential factors affecting buyers' influence and house prices. The structure, which includes the size of the house, the number of bedrooms, the number of bathrooms, and several other facilities, is commonly used as a benchmark that influences the selling price of a house (Musa, et al., 2023).

The importance of house price prediction data can impact both property sellers and buyers. For buyers, house price predictions can lead to better decision-making, helping them determine their decision in purchasing a house by providing an understanding of the value of a property Assudani and Wankhede (2022). House price predictions can also benefit property sellers by aiding in making decisions about the price they can offer to buyers (Kaushal and Shankar, 2021).

Machine learning technology can be applied in house price prediction research by utilizing various machine learning algorithms. This can help improve the accuracy of house price predictions in the market based on the characteristics of each house, resulting in more precise price assessments, better risk analysis, and more accurate lending decisions (Weng, 2022).

The use of machine learning has been applied with the Artificial Neural Network (ANN) algorithm, as it can be used to solve complex computational problems, such as house price prediction. According to Meghana, et al. (2024), the ANN algorithm can be an effective tool for performing regression tasks due to its ability to detect complex functions or relationships between input and output variables. Additionally, the ANN algorithm also has low computational costs (Rahman and Asadujjaman, 2021).

Other machine learning algorithms, such as the Random Forest algorithm, can also be used for prediction-related problems. Compared to other algorithms, Random Forest offers several advantages in regression problem analysis, such as minimal parameter tuning, making it very useful when a short development time is prioritized. Moreover, this algorithm can achieve high accuracy and low error rates (Gao, et al., 2023) (Truong, et al., 2020).

Research on house price prediction using machine learning algorithms remains relevant. The advantages of both algorithms can be leveraged to improve the accuracy of house price predictions. By comparing these two algorithms, insights into their similarities can be identified, which can enhance prediction accuracy (Harris and Grzes, 2019). Geerts, et al. (2023) mention that accurate house price predictions can provide better information about residential properties and improve housing policies as well as assessments of the housing market. Additionally, Li and Li (2024) emphasize that there is still a lack of research on house prices in developing countries, leaving room for quality improvement.

Given the research urgency mentioned earlier, further study on house price prediction using advanced technology is still needed. Moreover, research conducted in major cities in developing countries, such as Bandung, is an appropriate case study for this research. According to Savitri and Nasrudin (2023), the growth rate of Bandung, which creates favorable investment conditions, makes the city a promising area in the housing sector.

2. LITERATURE REVIEW

2.1. Artificial Neural Network (ANN)

One of the machine learning techniques for prediction is using the Artificial Neural Network (ANN) model. In short, an Artificial Neural Network is a computational model inspired by the workings of neural networks. The similarity between ANN and neural networks lies in their characteristics, such as adaptability and learning, generalization, massive parallelism (or the ability to work simultaneously in large quantities), robustness against noise, associative storage, and varied information processing. In the ANN algorithm, neurons represent computational processing elements that are interconnected with each other through weight coefficients. This enables the ANN algorithm to be used as a non-linear, multi-layer, and regression-based computational technique (Shanmuganathan, 2016).

Essentially, an ANN consists of neurons that are organized into an input layer and an output layer. These layers are interconnected, forming the ANN model. However, an ANN with multiple layers includes one or more hidden layers (Zhang, 2018). The depiction of the layers in an ANN is illustrated in **Figure 1**.

1) Input layer, it is the first layer in the ANN model that functions as the layer that receives input information.
2) Hidden layer, this layer functions as the processing unit in the neural network by applying weights based on the input layer. In this layer, the algorithm is enabled to extract high-level statistical features from its inputs.
3) Output layer, this layer consists of neurons that produce the decision or output signals of the algorithm. The output of this layer can be in the form of classifications or values for regression problems.

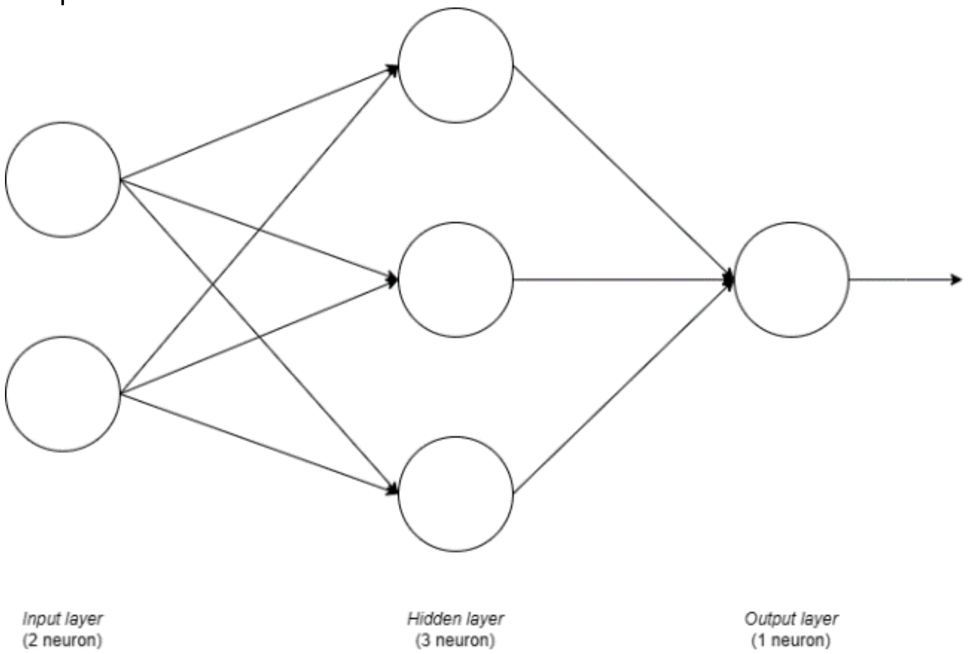

Figure 1. ANN algorithm illustration with its layers

2.2. Random Forest Regression

The Random Forest algorithm is a machine learning algorithm that can be used for prediction problems, including regression. One of the advantages of Random Forest is its ease of use. This algorithm only requires tuning a few parameters to produce a highly accurate model. Additionally, Random Forest excels in handling datasets with small sample sizes and features with high dimensionality (Biau and Scornet, 2016).

Random Forest is an extension of the decision trees algorithm, meaning that the Random Forest algorithm consists of a collection of predictor or decision trees. The way Random Forest works in solving regression problems is illustrated in **Figure 2**. Each tree is trained on different randomly selected data subsets, which are split accordingly. For regression cases, the final output, used as the prediction result from each tree in the Random Forest model, is aggregated or averaged across all the predictions from its trees (Biau and Scornet, 2016). This approach enhances prediction accuracy because aggregating the results from many trees reduces the variability and bias that may exist in individual predictions. In this way, Random Forest can produce a more accurate model compared to using a single decision tree (Harris and Grzes, 2019).

In the research by Rodriguez-Galiano et al. (2015), the error in the generalization of the Random Forest model decreases as the number of trees increases, thereby reducing overfitting in the model. However, Breiman, as cited by Biau and Scornet, suggests that the number of trees in a Random Forest should be limited because it can reduce the correlation between the trees.

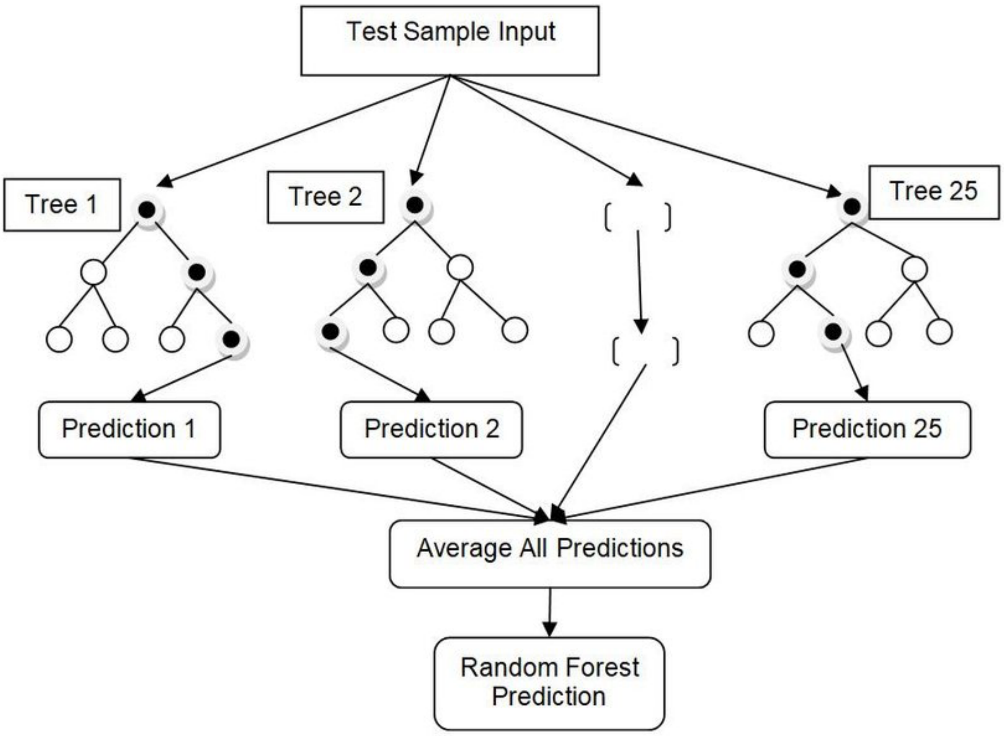

Figure 2. Illustration of how Random Forest algorithm works with 25 trees

3. RESEARCH METHOD

4.1. Clarification Study

At this stage, a literature review of previous studies related to machine learning and house price prediction is conducted to identify the objectives and core issues of the research. Additionally, the literature review aims to ensure that the upcoming research aligns with previous studies by exploring effective techniques, recommendations from researchers, and potential advancements. This stage is also undertaken to find references that can assist in this research.

4.2. Descriptive Study I

This stage is carried out to map the problems and objectives identified after conducting the literature review in the research clarification phase, ensuring that the research remains consistent with existing studies. The problems and objectives are determined based on the literature review previously conducted during the research clarification phase.

The first descriptive study phase is conducted to establish benchmarks for measuring the success of the research. This phase aims to provide an overview of the research context and the variables that will be used in the study.

4.3. Prescriptive Study

There are three main activities in this prescriptive study: data processing, model development, and result evaluation. Data processing involves preparing the house price dataset, data cleaning, and data splitting. Next is model development, which includes creating and optimizing the model, training the model, and testing the developed model. Finally, model evaluation is carried out by analyzing and comparing the results of each developed model. The model evaluation activity also considers the metrics used in this research.

1) Data Preparation

In this activity, the process includes searching, collecting, and improving the dataset to be used. Dataset collection involves downloading or extracting datasets that are already available on the internet. Dataset improvement addresses any data or features that may render the dataset unusable.

2) Data Cleaning

Data cleaning activities are performed on the prepared dataset. The data cleaning process may include common techniques such as selecting relevant columns from the dataset for model training, performing data standardization, and data normalization. The purpose of data cleaning is to improve data quality and optimize performance before processing by the machine learning model (Fatima et al., 2017). According to Assudani and Wankhede (2022), data cleaning techniques include handling missing data and removing extreme values. Additionally, Fatima et al. (2017) also mentions that data correction techniques and adjusting data to real-world cases provide consistency to the dataset.

Based on previous research, the data cleaning process in this study will involve several activities. This includes removing parts of the dataset such as duplicate data, irrelevant columns, data with missing values, house price data that only involves land sales, and some extreme values. Removing data that only involves land sales is done to maintain dataset consistency and focus the research on house price prediction. Removing duplicate data and data with missing values helps maintain data integrity, preventing bias (Fatima et al., 2017). Reducing extreme values or outliers is also performed because extreme values in the dataset can adversely affect model performance or accuracy (Tang et al., 2022). However, this does not mean outliers are removed entirely, as completely removing outliers can lead to data bias against group differences (Karch, 2022).

Additionally, several data transformation activities are carried out in the data cleaning process for this study, such as changing data types, encoding categorical data, and normalizing data. Changing data types for a feature or column and encoding categorical data (converting categorical data into numerical form) are intended so that house price data can be processed by the machine learning prediction models. Data normalization is performed to transform data into a more uniform format, ensuring that no data value dominates others. This also makes value interpretation more consistent and enhances model analysis performance (Firmansyah, 2024).

3) Data Splitting

Data splitting involves dividing the data into three parts: training data, test data, and validation data. Training data is used by the model to learn how the data are related. This data is also used by the machine learning model as a reference for its predictive capabilities, making it a crucial factor influencing the model's accuracy. Test data is used by the model to evaluate its predictive ability. With this data, the machine learning model can be assessed on how well it has learned from the training data. Validation data in this study is used for model parameter tuning and to check for overfitting. This aligns with the explanation provided by Firmansyah (2024), who observed the distance between training and validation results. Parameter tuning is conducted on the ANN algorithm model during the hyperparameter tuning process.

In the studies by Rahayuningtyas, et al. (2021) and Saiful et al. (2021), the data ratios used were 70:30 and 80:20 for training and test data. Additionally, other research by Xu and Zhang (2021) used data ratios of 80:10:10, 70:15:15, and 60:20:20 for training, test, and validation data. Similarly, Muneeb (2022) used a ratio of 50:25:25 for training, test, and validation data. However, this study will explore data ratios by adding different splitting ratios. The data ratios used in this study are divided into four: 80:10:10, 70:15:15, 60:20:20, and 50:25:25. Each data ratio will be tested on both the ANN and Random Forest algorithms to determine which data splitting ratio yields the best model performance.

4) Data Splitting

To determine the appropriate parameters and architecture for the ANN model, hyperparameter tuning techniques are used. This is to reduce overfitting in the model and optimize its performance level (Calugar, et al., 2022). The house price prediction models are created using the ANN and Random Forest algorithms. Additionally, each model will be developed with different implementations of training and test data ratios.

5) Model Training and Testing

After the data has been split and the model architecture has been set, the next step is to implement the training and test data into the model. The results of training and testing the model will be compared with other models to evaluate the performance of each. Additionally, an analysis of the validation results will be conducted to assess the extent of overfitting in the model.

4.4. Descriptive Study II

At this stage, the results from the trained and tested algorithm models are calculated based on the predetermined evaluation metrics. The outcomes of the models will be evaluated by comparing the predictions from each model. This comparison process is conducted to assess the performance of each model. Conclusions and hypotheses are formulated based on this comparison evaluation, providing an overview of the research results obtained.

4.5. Dataset

The data preparation process uses a dataset obtained from Kaggle titled "House Price Data in Bandung City." This dataset was collected from the website rumah123.com in March 2024 through web scraping by Al Faaath (2024). The dataset includes house price data for the Bandung City area, West Java, along with information such as house names, installment details, whether the listing is a premier or featured website, type, price, location by sub-district, and structural characteristics of the houses. The available dataset is in .csv format with 7,611 rows and 11 columns.

4.6. Metrics

1) Mean Absolute Errors (MAE)

MAE is defined as a model evaluation method that calculates the mean of the absolute differences between the actual data and the predicted data. This evaluation metric is used to assess how far the model's predictions are from the actual values on average (Hodson, 2022). In this research, MAE measures the difference between the actual house prices and the predicted house prices from the model. The smaller the MAE, the better the model's performance (Sharma et al., 2021).

$$MAE = \frac{\Sigma|Y' - Y|}{n}$$

Where:
Y’= Prediction Value
Y=Actual Value
n=Amount of data

2) Mean Absolute Errors (MAE)

MSE is a model evaluation metric that calculates the mean of the squared differences between the actual values or actual house prices and the predicted house prices from the model. This metric is sensitive to extreme values or outliers (Plevris et al., 2022). Therefore, MSE can be used to provide information on whether there are extreme values or outliers that significantly affect the model's performance.

$$MSE = \frac{\Sigma|Y' - Y|^2}{n}$$

3) Root Mean Squared Errors (RMSE)

The RMSE metric measures model performance by calculating the square root of the MSE. According to (Chai and Draxler, 2014). RMSE is useful for identifying whether a model has performance issues by giving more weight to larger errors. Additionally, RMSE avoids absolute values, which are often avoided in many mathematical calculations, especially when the data is likely to be Gaussian distributed.

$$RMSE = \sqrt{\frac{\Sigma|Y' - Y|^2}{n}}$$

4) R Squared (R^2)

The coefficient of determination, or R-squared, is a commonly used metric for evaluating the performance of a regression model. This metric measures the proportion of variability or dispersion that can be explained by the model, meaning it calculates the proportion of variance between the actual data values and the predicted values (Plevris et al., 2022). According to Chicco, et al. (2021), this metric has a clear upper limit, unlike MAE, MSE, and RMSE, which do not have an upper bound and can potentially yield positive infinity values. This makes R-squared more informative and easier to interpret. R-squared values range from 0 to 1. Unlike the previous metrics, a value of R-squared close to 1 indicates that the model explains most of the variability in the data.

$$R^2 = 1 - \frac{\sum(Y - Y')^2}{\sum(Y - \bar{Y})^2}$$

4. DISCUSSION AND EXPERIMENTAL ANALYSIS

4.1. Data Processing

1) Data Preparation

The dataset used in this study is in raw form, meaning it does not include column names or features in the file. Therefore, column names have been added, and the descriptions of these columns are provided in **Table 1**. This naming is based on the information presented on the Kaggle page.

Table 1. Units for magnetic properties.

Colum Name	Data Type	Description
type	String	Contains the type of property for each item. In the raw dataset, all rows have the value "house."
status	String	Contains values such as "premiere," "featured," or null.
price	String	The house price in Indonesian Rupiah, formatted as a string.
house_name	String	The name or title of the property for sale.
installments	String	Information on installments for each property for sale.
location	String	The location of the property for sale, formatted as Sub-district, City.
bedroom_count	Number	The number of bedrooms in the property for sale.
bathroom_count	Number	The number of bathrooms in the property for sale.
carport_count	Number	The number of garages or parking spaces available for the property for sale.
land_area	String	The land area of the property for sale, in meters (m).
building_area	String	The building area of the property for sale, in square meters (m^2)

2) Data Cleaning

After the data has been prepared by adding column names, further adjustments are needed through the data cleaning stage. Therefore, the data cleaning process in this study follows the techniques mentioned earlier. The data cleaning process carried out in this research is illustrated in Diagram **Figure 3**.

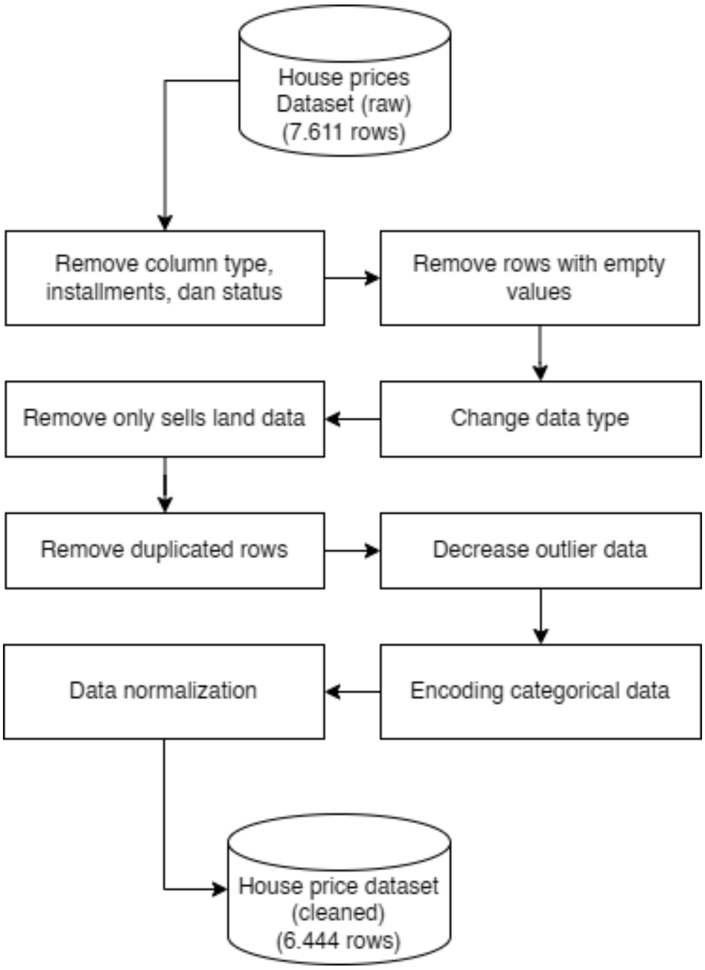

Figure 3. Data cleaning process flow diagram.

3) Data Splitting

The data splitting process was carried out using the train_test_split() function from the sklearn library. The splitting was conducted in two stages. The first stage involved dividing the entire dataset into training and test data, with the random_state parameter set to 42. The second stage involved splitting the test data into test and validation data by halving the number of test data points, which was done by setting the test_size parameter to 0.5. The number of data points resulting from the splitting process can be seen in **Table 2**.

Table 2. The amounts of data splitting on each data.

Ratio of *Training Data* (%)	**Amount of *Training Data***	**Amount of Test Data**	**Amount of Validation Data**
80	5.237	655	655
70	4.582	982	983
60	3.928	1.309	1.310
50	3.273	1.637	1.637

DOI: https://doi.org/10.17509/seict.v5i2.75945
p-ISSN 2774-1656 | e-ISSN 2774-1699

4.2. Development of House Price Prediction Models

A total of eight predictive models were developed. These models are divided into two categories based on the algorithm used, with each category consisting of four models, differentiated by the data ratio used. Overall, the models within each algorithm category share the same architecture. The structure of the ANN models was determined using hyperparameter tuning, while the Random Forest models were developed through several iterative experiments and by referencing previous studies.

1) ANN Model Architecture

For the ANN models developed in this research, the layers and their parameters were determined using the hyperparameter tuning process. This process involved creating an ANN model with adjustable configuration parameters to identify the optimal setup. During this process, 80% of the data was used for training, along with 20% of the training data reserved for validation.

The grid search technique was employed for hyperparameter tuning, with a maximum of 10 trials to find the optimal model configuration, and the process was conducted over 50 epochs. In each trial, the model was executed or trained twice, with the best RMSE metric being monitored. The hyperparameter tuning also monitored parameters such as the number of dense layers, the optimal learning rate, and the activation function used.

The hyperparameter tuning process resulted in the best RMSE value of 0.1186441220343113, with the optimal learning rate found to be 0.000316227766016838. The activation function and the number of dense layers used are detailed in **Table 3**.

Once the architecture was determined, the next step was to create the ANN models based on the architecture identified through hyperparameter tuning. All four models developed share the same architecture as determined by the hyperparameter tuning process. The models were built using the TensorFlow library. Each model was trained over 500 epochs using the Adam optimizer function with the learning rate set according to the results of the hyperparameter tuning. Early stopping regularization was also added to the models to reduce overfitting (Li, et al., 2020). This was implemented using the EarlyStopping function from the TensorFlow library, which works by saving the best training performance of the model and halting training if no further accuracy improvements are observed.

Table 3. Results of The Hyperparameter Tuning

Layer	Activation Function	Besaran Dense
Layer 1	Relu	50
Layer 2	*None*	1
Layer	**Activation Function**	**Besaran Dense**
Layer 1	Relu	50

2) Random Forest Model Architecture

The Random Forest model implemented in this study is an ensemble learning method that uses decision trees for regression. Following the research conducted by Kurniawan, et al. (2020), this model comprises 512 decision trees, or n_estimators. The maximum depth (max_depth) of each tree is limited to 5 to ensure that the decision trees do not become overly complex. Additionally, based on the research by Togatorop, et al. (2022), each split of the tree only considers the square root of the total features using the max_features="sqrt" parameter.

3) Testing Results

After completing the training and validation process, the model undergoes testing. This testing phase uses the test dataset that was previously split. According to Table VII, the ANN model with a 60:20:20 data ratio achieved the best results across all metrics compared to the other four ANN models. This model has low MAE, MSE, and RMSE values, and a fairly good R² result. This means that the model can predict with a small margin of error and can understand the variability in house prices fairly well, at 64.05%. The 60:20:20 data ratio is an optimal data split, meaning that the model with a 20% test data ratio provides the best results during the testing phase without causing the model to suffer from overfitting or underfitting. The model is able to learn patterns in the data with 60% of the data used for training, which is 3,928 data points, and 20% of the data used for testing, which is 1,309 data points.

The MAE values obtained for the ANN models range from 0.0488 to 0.0513, indicating that the models produce a low and fairly consistent average prediction error. The lowest MAE of 0.0488 means that the model has the most accurate average prediction capability, closely matching the actual values. This consistency is also observed during the training phase of the ANN models, where the MAE ranges from 0.0492 to 0.0501. Additionally, the MSE, used to assess the impact of outliers on the model, ranges from 0.0097 to 0.0106. This indicates that the MSE values have a larger difference than the MAE values. Some outliers may influence the model's prediction error rate, but the lowest MSE value of 0.0097 suggests that these outliers do not have a significant impact. The RMSE values across all ANN models, which range from 0.0987 to 0.1048, indicate that the average prediction error deviates only slightly from the scale of the actual data. The model with the lowest RMSE of 0.0987 achieves a smaller and more accurate prediction error relative to the actual data.

Referring to the scatter plot in Figure 6, the diagram illustrates the distribution of predicted values against the actual data. In this diagram, the actual data is represented by a trend line shown in red. From the diagram, it can be observed that the predicted values with higher amounts tend to fall below the trend line of the actual data. This suggests that the model struggles to accurately learn patterns with extreme values and the variability of the data. As a result, even though the model's R² metric is relatively good, it fails to capture the variability in data with higher values effectively.

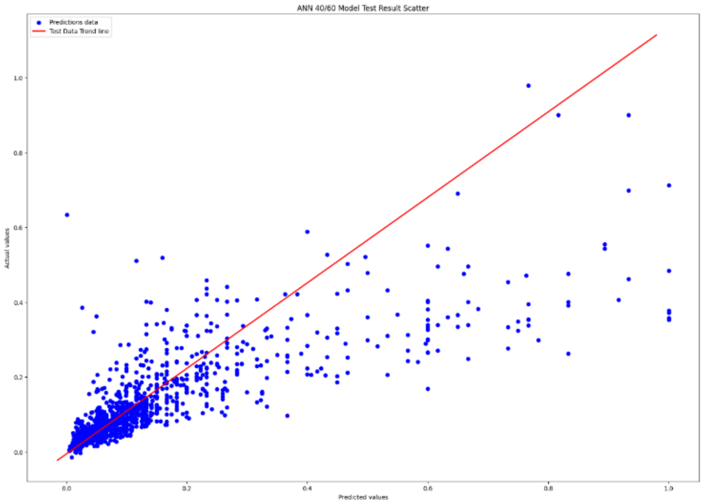

Figure 4. Scatter plot of ANN model result with data ratio of 60:20:20

DOI: https://doi.org/10.17509/seict.v5i2.75945
p-ISSN 2774-1656 | e-ISSN 2774-1699

Next is the evaluation of the Random Forest model through the model testing process. As shown in Table 4.16, the model with the same data ratio as the ANN model, namely the 60:20:20 model, performs better than the other four Random Forest models. This optimal data ratio indicates that the 60:20:20 ratio provides a more balanced distribution of data between training, validation, and testing, giving the model sufficient data to learn and generalize new data. The MAE range, from 0.0480 to 0.0484, is very narrow, indicating a small average error. The minimal difference between the highest and lowest MAE values among the Random Forest models suggests that each model with this algorithm consistently produces average predictions with low errors.

The MSE values for the Random Forest models show a smaller gap compared to the MAE. Compared to the ANN models, the Random Forest models tend to have lower MSE values, indicating that outliers have a lesser impact on the average prediction error of the model. The RMSE values for the Random Forest models range from 0.0888 to 0.0920, meaning the Random Forest models have a smaller average error relative to the scale of the original data compared to the ANN models. With lower RMSE values, the Random Forest model demonstrates better generalization ability, especially in the model with the 60:20:20 data ratio.

The model with the 60:20:20 data ratio shows fewer errors in terms of MAE, MSE, and RMSE compared to the other three Random Forest models. However, the differences between each Random Forest model are not significant. Unlike the ANN algorithm models, the Random Forest models generally have higher R^2 values than the ANN models. The average R^2 for the ANN models is 0.6179 or 61.79%, while the Random Forest models have a higher average R^2 of 0.69102 or 69.102%. The difference of 7.312% is substantial enough to influence the model's accuracy. This can be confirmed by comparing the scatter plot diagrams between the ANN model testing results **(Figure 4)** and the Random Forest model **(Figure 5)**.

Unlike the scatter plot of the ANN model test results, the scatter plot of the Random Forest model has a pattern that is more closely aligned with the actual data trend line. This explains the higher R^2 results of this model compared to the best ANN model. However, this model exhibits more extreme values above the actual data trend line, and its distribution is wider.

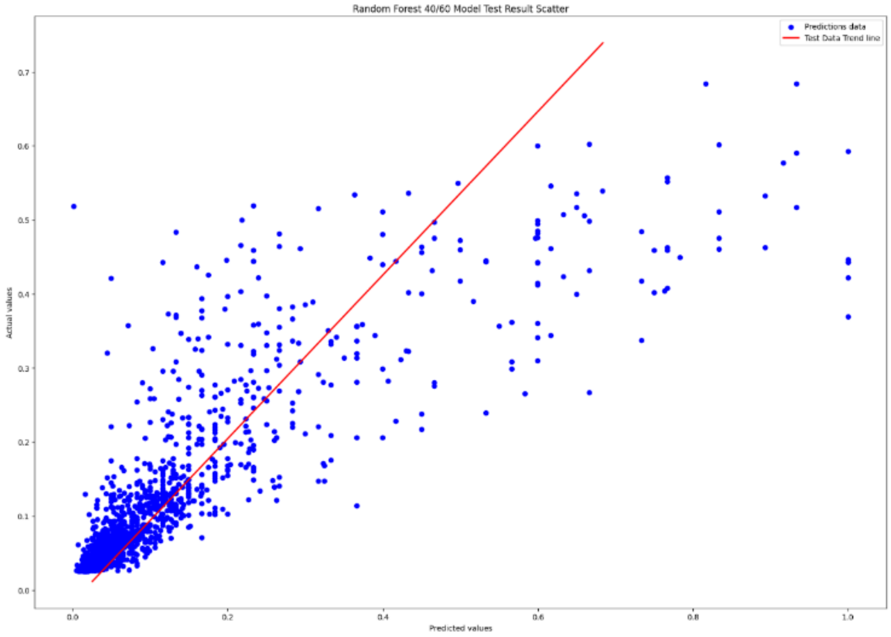

Figure 5. Scatter plot of random forest model result with data ratio of 60:20:20

Based on the evaluation results of both models, the Random Forest model demonstrates superior predictive performance, particularly in terms of predicted error (MAE, MSE, and RMSE) and better data variability (R^2). The difference in predicted error between the two models is not substantial, but the difference is more pronounced in the R^2 metric. The Random Forest model achieves a higher R^2 value, reaching 70%. Both the R^2 metric and the scatter plots describe the variability or distribution of data predicted by the models. When comparing the scatter plots, the Random Forest model's predictions are more closely aligned with the actual data trend compared to the ANN model. This comparison correlates with the R^2 values of both models. However, both models still struggle to capture patterns in data with high price values. The comparison of the evaluation results also reveals that both algorithms achieve their best performance with the same data ratio. This indicates that a training data ratio of 60% and a validation data ratio of 20% is optimal for both models in learning the data for evaluation or testing purposes. Additionally, a test data ratio of 20% suggests that the models have good generalization capabilities for new data.

5. CONCLUSION

The application of machine learning technology for house price prediction has proven to be effective, as demonstrated by the research findings. The analysis of both ANN and Random Forest models on the house price dataset from Bandung shows promising results. Among the models, the Random Forest algorithm outperforms with evaluation metrics of MAE: 0.0470, MSE: 0.0079, RMSE: 0.0888, and R^2: 0.7085, and training metrics of MAE: 0.0445, MSE: 0.0065, RMSE: 0.0806, and R^2: 0.7487. For both ANN and Random Forest, the optimal data ratio of 60:20:20 for training, validation, and testing data yielded the best results. Random Forest consistently showed better performance compared to ANN, which had evaluation metrics of MAE: 0.0488, MSE: 0.0097, RMSE: 0.0987, and R^2: 0.6405.

For future research, it is recommended to expand the dataset to include a wider range of house characteristics and price variables, exploring factors such as environmental influences and strategic locations. Additionally, employing other generalization techniques in ANN, such as L1 or L2 Regularization and data augmentation, could enhance the model's ability to generalize and understand patterns. Finally, investigating other machine learning algorithms could provide insights into how different approaches address house price prediction and reduce prediction bias.

6. REFERENCES

Al Faaath, K. (2024). Daftar harga rumah di Kota Bandung. Retrieved from https://www.kaggle.com/datasets/khaleeel347/harga-rumah-seluruh-kecamatan-di-kota-bandung

Assudani, P., and Wankhede, C. (2022). Analysing the factors influencing the house prices and studying house price prediction methods. International Journal of Next-Generation Computing.

Biau, G., and Scornet, E. (2016). A random forest guided tour. Journal TEST, 25, 197–227.

Bilmes, J. (2020). Underfitting and overfitting in machine learning.

Calugar, A. N., Meng, W., and Zhang, H. (2022). Towards artificial neural network-based intrusion detection with enhanced hyperparameter tuning.

Chai, T., and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250.

Chicco, D., Warrens, M. J., and Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, 1–24.

Fatima, A., Nazir, N., and Khan, M. G. (2017, March). Data cleaning in data warehouse: A survey of data pre-processing techniques and tools. International Journal of Information Technology and Computer Science, 9(3), 50–61.

Firmansyah, M. R. (2024). Stroke classification comparison with KNN through standardization and normalization techniques. Advance Sustainable Science, Engineering and Technology, 6(1), 02401012.

Gao, J., Wang, K., Kang, X., Li, H., and Chen, S. (2023, June). Ultra-short-term electricity load forecasting based on improved random forest algorithm. AIP Advances, 13(6).

Geerts, M., vanden Broucke, S., and De Weerdt, J. (2023, May 1). A survey of methods and input data types for house price prediction. MDPI.

Harris, L., and Grzes, M. (2019). Comparing explanations between random forests and artificial neural networks. In 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), 2978–2985.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. Copernicus GmbH.

Karch, J. D. (2022). Outliers may not be automatically removed. https://orcid.org/0000-0002-1625-2822

Kaushal, A., and Shankar, A. (2021). House price prediction using multiple linear regression.

Kurniawan, C., Dewi, L. C., Maulatsih, W., and Gunadi, W. (2020). Factors influencing housing purchase decisions of millennial generation in Indonesia. International Journal of Management (IJM), 11(4), 350–365.

Li, M., Soltanolkotabi, M., and Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In S. Chiappa & R. Calandra (Eds.), Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (Vol. 108, pp. 4313–4324). PMLR.

Li, N., and Li, R. Y. M. (2024, February). A bibliometric analysis of six decades of academic research on housing prices. International Journal of Housing Markets and Analysis, 17(2), 307–328.

Meghana, P., et al. (2024). Analysis of neural network algorithm in comparison to multiple linear regression and random forest algorithm. In 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems, ICETSIS 2024, 437–443.

Muneeb, M. (2022). LSTM input timestep optimization using simulated annealing for wind power predictions. PLoS ONE, 17(10). https://doi.org/10.1371/journal.pone.0275649

Musa, U., Zahari, W., and Yusoff, W. (2015). The influence of housing components on prices of residential houses: A review of literature.

Plevris, V., et al. (2022). Investigation of performance metrics in regression analysis and machine learning-based prediction models. In The 8th European Congress on Computational Methods in Applied Sciences and Engineering.

Rahayuningtyas, F. E., Rahayu, F. N., Azhar, Y., and Artikel, I. (2021). Prediksi harga rumah menggunakan general regression neural network. Jurnal Informatika, 8(1).

Rahman, M., and Asadujjaman, M. (2021, September). Implementation of artificial neural network on regression analysis. In 2021 5th Annual Systems Modelling Conference, SMC 2021.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., and Chica-Rivas, M. (2015, August). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees, and support vector machines. Ore Geology Reviews, 71, 804–818.

Saiful, A., Andryana, S., and Gunaryati, A. (2021). Prediksi harga rumah menggunakan web scrapping dan machine learning dengan algoritma linear regression. Jurnal Teknik Informatika dan Sistem Informasi, 8(1).

Savitri, N. F., and Nasrudin, N. (2023, November). Peramalan indeks harga properti residensial di kota Bandung tahun 2023. Jurnal Kebijakan Pembangunan Daerah, 7(2), 140–157.

Shanmuganathan, S. (2016). Studies in Computational Intelligence 628 Artificial Neural Network Modelling.

Sharma, S., Jhaketiya, V., Kaul, A., Raza, A. A., Ahmed, S., and Naseem, M. (2021). Automatic prediction of road angles using deep learning-based transfer learning models. IOP Conference Series: Materials Science and Engineering, 1099(1), 012060.

Tang, J., et al. (2022). Joint modeling strategy for using electronic medical records data to build machine learning models: An example of intracerebral hemorrhage. BMC Medical Informatics and Decision Making, 22(1).

Togatorop, P. R., Sianturi, M., Simamora, D., and Silaen, D. (2022). Optimizing random forest using genetic algorithm for heart disease classification. Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, 13(1), 60.

Truong, Q., Nguyen, M., Dang, H., and Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433–442.

Weng, W. (2022). Research on the house price forecast based on machine learning algorithm.

Xu, X., and Zhang, Y. (2021). House price forecasting with neural networks. Intelligent Systems with Applications, 12, 52.

Zhang, Z. (2018). Artificial neural network. In Multivariate Time Series Analysis in Climate and Environmental Research, 1–35.

Zulkifli, F., and Ismail, H. (2023). Factors influencing house buyer's decision in Malaysia. Case study: Sepang, Selangor.