Web-Based Property Value Prediction Utilizing Random Forest Algorithms

Dewi Nur Arifah[1*], Salamun Rohman Nudin[2], Dodik Arwin Dermawan[3], I Gde Agung Sri Sidhimantra[4], and Asmunin[5]

[1*, 2, 3, 4, 5] Universitas Negeri Surabaya, Surabaya, Indonesia

Keywords:
Property price prediction
Machine learning
Random Forest algorithm
Web-based system
Laravel frameworks

ABSTRACT

Property price prediction is a complex problem and has many influencing factors. The lack of applications that can facilitate property appraisers in predicting property values makes it difficult know prices quickly and accurately, so that it can disrupt the balance and efficiency of the property market. Research with the title Development of a Web-based Property Value Prediction System can provide a clearer picture or recommendation about the selling value of property in Surabaya based on physical characteristics, position and location. This research was conducted using one of the machine learning algorithm models, namely Random Forest for data processing and modelling. The data used includes information about land area, road width, designation zone, and indication of land value. The results of the prediction will be displayed on the website by utilising the Laravel frameworks. Evaluation of the developed model showed promising results, with an accuracy of 83% in predicting property values. This shows that the system has the potential to help property appraisers determine property values more effectively.

INTRODUCTION

The property market is a place where people buy, sell, or rent properties, such as houses, apartments, land, and commercial buildings. In the property market, sellers offer their properties at a certain price, and buyers look for properties that suit their needs and budget. The property market has evolved to be more than just a place to live or a place to do business [1]. Investment in property has become one of the most significant and far-reaching forms of investment, covering residential, commercial, and industrial segments. Property price prediction is one of the important issues in real estate and finance.

Property prices are influenced by various factors such as location, size, shape, and surrounding facilities [2]. Accuracy in predicting property prices is very important for homeowners, investors, and real estate agents to make the right decision in buying or selling property [3]. Therefore, the development of reliable prediction models is the focus of much research in this area. Machine learning algorithms are often used to predict property prices. Among the many existing algorithms, Random Forest is one of the most effective.

Random Forest uses multiple decision trees simultaneously to improve the accuracy and stability of predictions [4]. Previous research shows that Random Forest has significant advantages over traditional methods such as linear regression. Random Forest is able to produce more accurate predictions and is more resistant to overfitting than linear regression. In addition, Random Forest is more effective in handling data with many features and interactions between features. However, there are still challenges in applying Random Forest for property price prediction.

In addition, the selection of appropriate parameters for Random Forest, such as the number of trees and maximum depth, is crucial to achieve optimal performance. This research aims to address these challenges and evaluate the performance of Random Forest in the context of property price prediction. To achieve this goal, this research uses an extensive and diverse property transaction dataset, covering a wide range of relevant

features. This dataset is processed through several stages, including data cleaning, categorical feature coding, and normalization.
Random Forest models were then trained and evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The results of this study are expected to provide deeper insights into the effectiveness of Random Forest in predicting property prices and make important contributions to the development of more accurate prediction models.

1.1 Random Forest Model

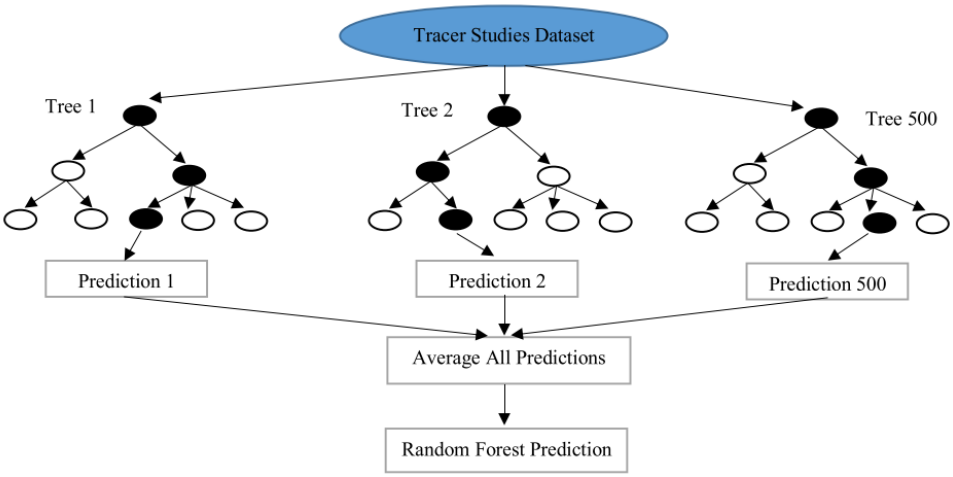

Fig. 1. Prediction of Random Forest Model
Source: (*S. R. Nudin et al*, 2022)

There are two techniques used to develop the Random Forest model: bagging and random subspace. Bagging involves taking bootstrap samples and combining the models learned from each sample. Bootstrapping is a statistical method of random sampling with replacement used to manage unbalanced data [5]. In the Random Forest model, the nodes in each decision tree are randomly assigned to select attributes from a random subspace. The use of bagging and random subspaces helps the Random Forest model manage overfitting more effectively than a single decision tree [6]. Generally, developing a Figure 1 involves four steps:

1. Using the bootstrap method to create a random sample that is the same size as the training dataset or the full dataset.
2. Employing the random subspace method to select K attributes from a total of M attributes, where K << M (typically, K is chosen to be equal to the square root of M).
3. Building a decision tree using the bootstrap sample and selecting attributes from steps 1 and 2.
4. Repeating steps 1 to 3 to construct multiple trees until the desired Random Forest is achieved.

The out-of-bag (OOB) error rate is used to determine the number of trees in the Random Forest.

1.2 Property Theories

As is well known, there are various bases used to assess the quality of a property. While there are several theories that can be applied, not all property theories are explained in this subchapter as they are too diverse. Therefore, only the basic theories used in the manufacturing process will be discussed. The following are the theories used:

- **Property Position**

There are 4 positions in this system. Interior position, this position is the most popular position because it can have two directions. Corner position, this position is still classified as a position that is quite popular, namely the position at the end. Kuldesak position, this position is a position that surrounds the park, or is at the entrance to a housing estate and the position of a skewer, this position is what prospective buyers usually avoid due to myths that exist in the community. But the position of the property cannot determine the price of the property because the quality of the property depends on the condition of the property [7].

- **Property Shape**

The shape and dimensions of a property are what determine a property the most. There are land shapes such as square, rectangle, triangle, trapezoid, and others that may occur. This is supported by the size of a property's land such as the length and width of the square metre.

- **Property Designation**

- **Residential zone / Housing:** Spatial designation consisting of groups of residential houses that accommodate the lives and livelihoods of people equipped with facilities.
- **Protected Forest**: Spatial designation which is part of a protected area that has the main function as protection of life support systems to regulate water systems, prevent flooding, control erosion, prevent seawater intrusion, and maintain soil fertility.
- **Green Open Space**: An elongated / striped and / or grouped area, the use of which is more open, where plants grow, both naturally growing plants and those that are deliberately planted.
- **Trade and Service Zones**: Spatial designations that are part of cultivation areas functioned for the development of commercial business activities, places of work, places of business, and places of entertainment and recreation, as well as supporting public/social facilities.
- **Industrial Zones Industry**: is an economic activity that processes raw materials, raw materials, semi-finished goods, and/or finished goods into goods with higher value for their use, including industrial design and engineering activities.
- **Other Designation Zones**: Spatial designations developed to accommodate activity functions in certain areas in the form of agriculture, mining, tourism, and other designations [8].

RESEARCH METHOD

Figure 2 is the section of Method explains how the research is conducted, research design, as well as techniques of data collecting and analysis. The following descriptions are guidance to set the page layout and text format of overall manuscript.

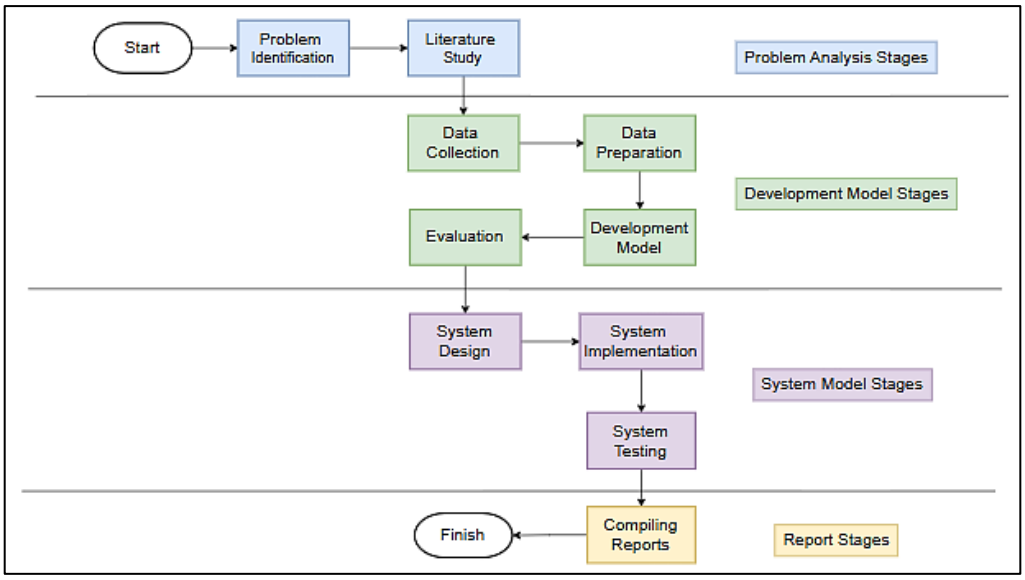

Fig. 2. Research Flow

2.1 Model Development

There are several stages of model development carried out to process the required data, such as data collection, data pre-pocessing.

2.1.1 Data Collection:

The data used in this study is property transaction data obtained from various sources, including appraiser. The dataset includes features such as land area, number of rooms, location, and selling price. The data is collected in CSV format to facilitate further analysis.

2.2.1 Data Pre-processing

Data pre-processing is done to improve the quality of the dataset and ensure the model can work optimally. As shown in table 1, which displays the features of the dataset we used consisting of Longitude, Latitude, Road Row, Designation, Shape, Property Location, Land Area, Building Area, Quantity of Bathrooms and Quantity of Bedrooms.

Table 1. Property dataset features

No	Fitur	Type Data
1	Longitude	Integer
2	Latitude	Integer
3	Road Row (m)	Integer
4	Designation	Text
5	Shape	Text
6	Property location	Text
7	Land Area	Integer
8	Building Area	Integer
9	Bathrooms	Integer
10	Bedrooms	Integer

The pre-processing steps include:

a. **Data Cleaning**: Removing missing values and dealing with data duplication [9].
b. **Categorical Feature Coding**: Converting categorical data (such as location) into numerical data using one-hot encoding techniques.
c. **Data Normalization**: Scaling numerical features to ensure that all features are in the same range, using methods such as Min-Max Scaling.
d. **Random Forest Model Training**: The Random Forest model is trained using the training data. The training process involves:
   - Initial Parameter Selection: Setting initial parameters such as number of trees (n_estimators) and maximum depth of trees (max_depth).
   - Parameter Optimization: Using Grid Search, Random Search and Halving Grid Search techniques to find the combination of parameters that gives the best performance. The optimized parameters include n_estimators, max_depth, min_samples_split, and min_samples_leaf.
e. **Model Evaluation**: The trained model was evaluated using test data [10]. Several evaluation metrics are used to assess model performance, including:
f. **Mean Absolute Error (MAE)**: Measures the average absolute error between the predicted price and the actual price. The formula 1 for calculating MAE.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} | yi - yi | \qquad (1)$$

g. **Root Mean Squared Error (RMSE)**: Formula 2 is used to measure the average squared error between predicted price and actual price

$$\text{RMSE} = \sqrt{\frac{\sum_{(Actual-predict)^2}}{n}} \qquad (2)$$

h. **R-squared (R^2)**: Formula 3 is used to measures how well the model explains the variability in the data.

$$R2 = 1 - \frac{\sum(yi-y)^2}{\sum(yi-y)^2} \qquad (3)$$

2.2 Website Development

This system is built in the form of a website, so that users can easily access and use this prediction system from anywhere. The system design includes the user interface and the process behind the website screen, how the data is inputted and processed, and how the prediction results are displayed to the user.

- **System Architecture**

  To develop a website-based property value prediction system that implements Random Forest prediction, there are several parts of the system. Property appraisers in this case act as users will access the website-based property value prediction application to input data and view prediction results with the flask framework for the frontend. Then the history of the data that has been inputted will be stored in the MySQL database which will later be used in the backend for the prediction, evaluation, and graphics process using Python and google collab as well as for the implementation of the random forest model.

- **Use case :**

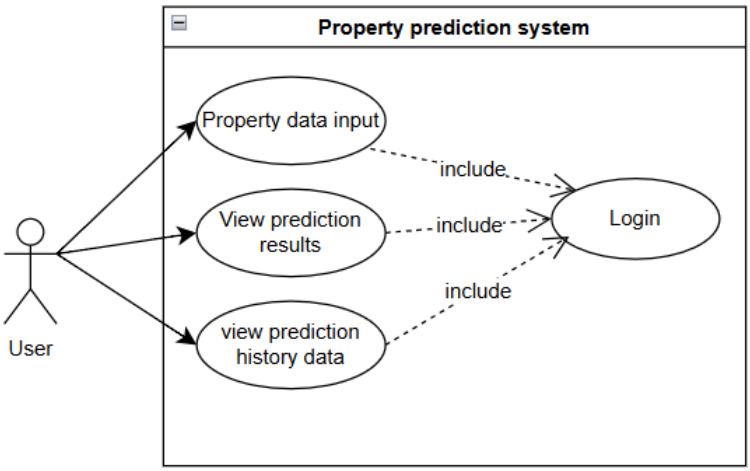

Fig. 3. Use Case System

Use case on Figure 3 diagram shows the relationship that occurs between actors and use cases that have each of their respective functions in a system. Where a user performs several stages such as login, dataset input, viewing data and graphs, and viewing prediction results as well as seeing the evaluation of the model that has been made.

- **User Interface Web:**

- Design of prediction feature page

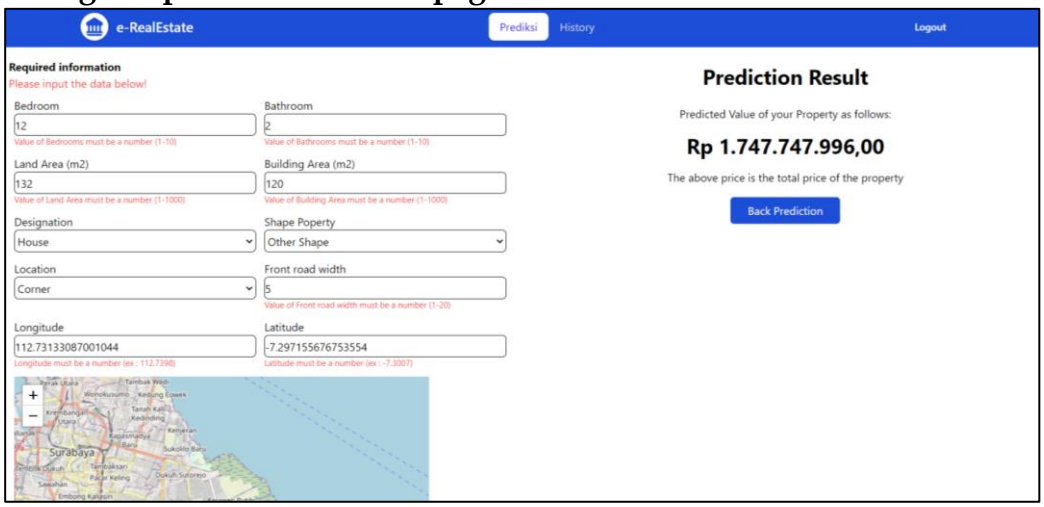

Fig. 4. Prediction Page

Design of prediction feature display On Figure 4. there are features consisting of several inputs to create a new dataset that will be processed and features of property value prediction results. The required inputs are coordinate points, address, property location, property shape, zone designation indication of permeter value and property area.

- **Design of overview feature page**

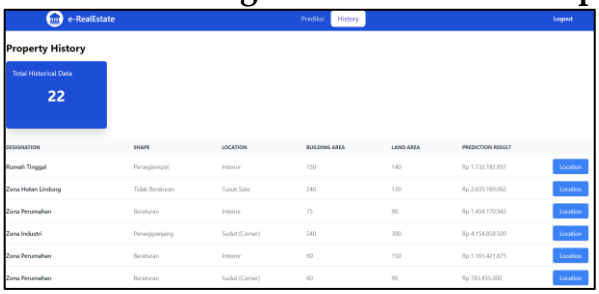

Fig. 5 History page

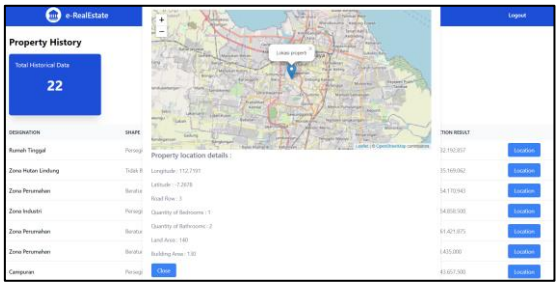

Fig. 6 Location Page

On Figure 5 there is a history of data that has been predicted before, which is displayed on this history page in the form of shape, designation, land area, building area and location details such as maps, longitude, latitude points.

- **System Implementation**

  The next stage is the system implementation stage, which is a continuation of the system design stage, namely translating the system design into a form that can be implemented, namely building a system that has been designed and applying a prediction calculation model to it. The system implementation process involves developing an information system to predict property price data in Surabaya, which is then applied to the system. This system development is done usinLaravel and flask for API.

- **System Implementation**

  The next stage is the system implementation stage, which is a continuation of the system design stage, namely translating the system design into a form that can be implemented, namely building a system that has been designed and applying a prediction calculation model to it. The system implementation process involves developing an information system to predict property price data in Surabaya, which is then applied to the system. This system development is done usinLaravel and flask for API.

- **System Testing**

  System testing focuses on the software in terms of logic and functionality to ensure that all parts have been tested. This is done to minimise errors and ensure that the resulting output is as desired. In this study, system testing was carried out using black-box testing techniques. Black-box testing is a software testing method that is carried out without paying attention to the internal details of the software. In the black-box testing process, the programme is tested by trying to input data on each form. This test is to find out the programme runs as needed.

RESULTS AND DISCUSSION

The initial step of model development is data collection, which is obtained from a database on the Radata website platform. The data taken consists of property data in Surabaya in the range of years 2019 - 2024 with 21 features and 1,990 rows of data. Next,

namely by doing Data Preparation which contains the stages of Data Cleaning, combined data and Data Encoder.

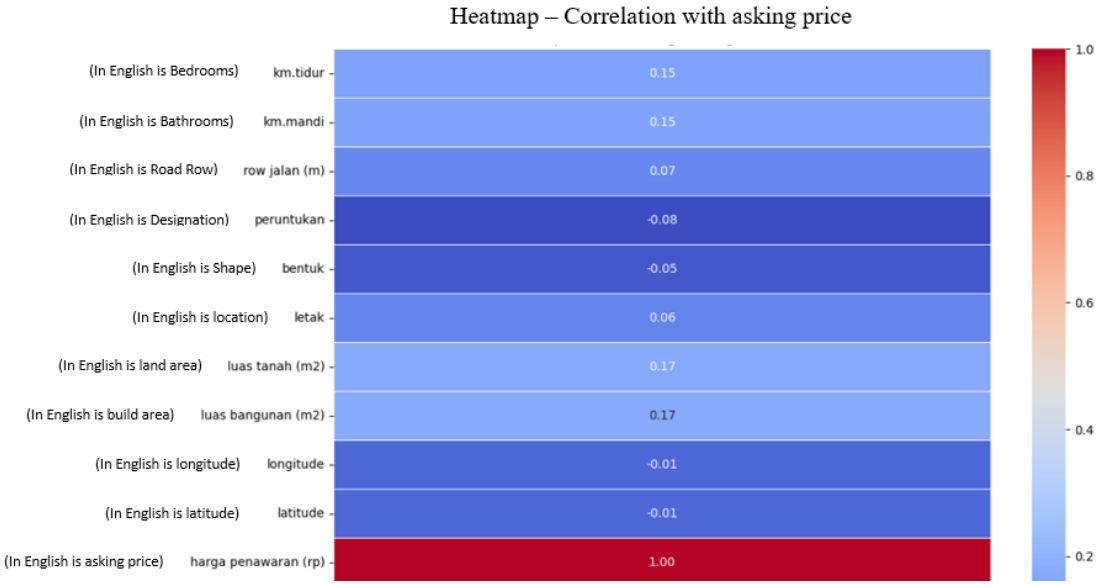

Fig. 7. Heatmap Correlation features

The Figure 6 diagram is a heatmap that shows the correlation between various property features and the asking price (harga penawaran in Indonesian). Correlation measures the relationship between two variables, with values ranging from -1 to 1. Here's how to read the diagram:

a. **Color Scale**:
Dark red (1.0): Very strong positive correlation, dark blue (-1.0): Very strong negative correlation, and lighter colors indicate weaker correlations, either positive or negative.

b. **Correlation Values**:
- **Positive Correlation**: Positive values (e.g., 0.17) indicate that as the feature value increases, the asking price tends to increase.
- **Negative Correlation**: Negative values (e.g., -0.08) indicate that as the feature value increases, the asking price tends to decrease.
- **Zero Correlation**: Values close to 0 (e.g., -0.01) indicate no clear linear relationship between the feature and the asking price.

c. **Interpretation of Each Feature**:
- **km.tidur (bedrooms)**: Positive correlation of 0.15, indicating that more bedrooms tend to increase the asking price.
- **km.mandi (bathrooms)**: Positive correlation of 0.15, indicating that more bathrooms tend to increase the asking price.
- **row jalan (road width in meters)**: Positive correlation of 0.07, showing a weak but positive relationship with the asking price.
- **peruntukan (designation)**: Negative correlation of -0.08, indicating that certain land uses tend to decrease the asking price.
- **bentuk (shape)**: Negative correlation of -0.05, showing a weak but negative relationship with the asking price.

- **letak (location)**: Positive correlation of 0.06, showing a weak but positive relationship with the asking price.
- **luas tanah (land area in m2)**: Positive correlation of 0.17, indicating that larger land areas tend to increase the asking price.
- **luas bangunan (building area in m2)**: Positive correlation of 0.17, indicating that larger building areas tend to increase the asking price.
- **longitude**: Negative correlation of -0.01, showing a very weak and negative relationship with the asking price.
- **latitude**: Negative correlation of -0.01, showing a very weak and negative relationship with the asking price.
- **harga penawaran (asking price)**: Correlation of 1.00, as this is the target variable compared with itself.

The next step is model development, in this step X and Y data will be defined, then divide the amount of data into 2, namely for Test and Train. Next, data training is carried out by implementing the Random Forest method, then using the Hyperparameter Optimization algorithm GridSearchCV method. In the modeling that has been made, the accuracy results will be obtained as in Figure 7 in the form of R2, MAE, and RMSE values.

```
R-squared (R2): 0.8367700789342328
Mean Absolute Error (MAE): 1512363644.942205
Root Mean Squared Error (RMSE): 3563481006.742458
```

Fig. 8. Result of R2 score, MAE, RMSE

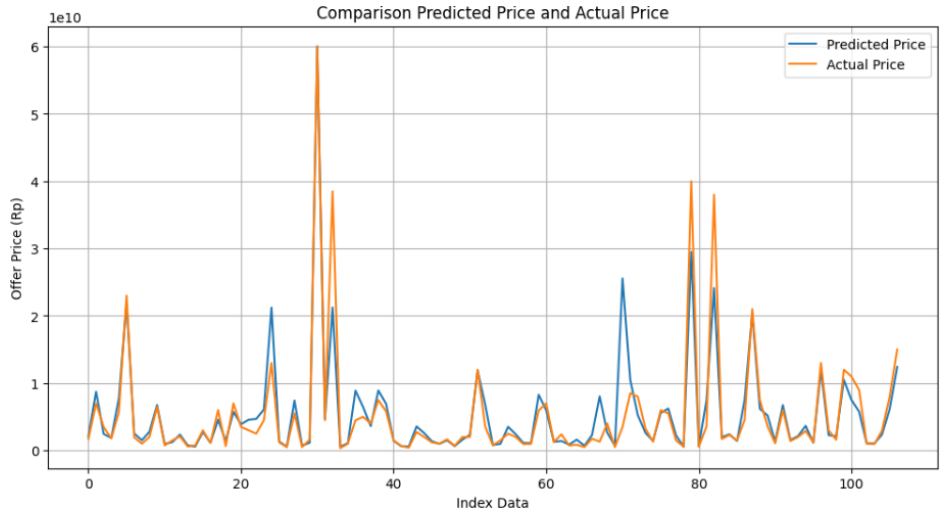

Fig. 9. Comparison of Predict Price and Actual Price

The figure 8 above shows a comparison chart between the predicted price (labelled "Predicted Price" and the blue line) and the actual price (labelled "Actual Price" and the orange line). The X-axis represents the order of the data by index. Each point on this axis indicates a specific time sequence or event of the tested data. The Y-axis shows the offer price value in rupiah (Rp). The numbers on the Y-axis indicate the magnitude of the price at a particular time or index. At some points, especially around indices 20, 40, and 80, there are large spikes in the actual price that are not always well followed by the predicted price, suggesting the model had difficulty predicting these spikes. In contrast,

at many other points, the predicted prices are quite close to the actual prices, suggesting the model works well with the data.

CONCLUSION

After the results of the stages of problem analysis, model development, web development, and system testing on the Development of a Property Value Prediction System in Surabaya Using the Web-Based Random Forest Method, a conclusion is obtained, among others: The implementation of the Random Forest algorithm to predict property values in Surabaya has been successfully carried out. Through the use of this algorithm, the system can analyse property data and produce fairly accurate predictions. The algorithm implementation steps involve data collection, data processing, model training, and testing. The model testing can be proven through the R2 Score evaluation results by showing the accuracy obtained by 83%. This shows that the amount of bid price prediction results can be influenced by the value of each feature such as location (longitude and latitude), road row, property formation, designation, number of bedrooms and bathrooms, and also the land and building area of the property. Web-based applications developed using the Laravel framework have been successfully implemented. This application produces output in the form of property value predictions that can be accessed and used by property appraisers. This can provide value recommendations and make it easier for users to determine the value or price of property in Surabaya.

ACKNOWLEDGEMENTS

The researcher would like to thank KJPP GEAR for providing the dataset so that it can help the authors to conduct his research and thanks also to parents, lecturers, friends and all those who have helped the authors to complete this research.

REFERENCES

Byman, R. (2005). Curiosity and sensation seeking: A conceptual and empirical examination. *Personality and Individual Differences, 38*(6), 1365-1379. https://doi.org/10.1016/j.paid.2004.09.004

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52(4), 281-302. https://psycnet.apa.org/doi/10.1037/h0040957*

Geddis, A. N. (1993). Transforming subject-matter knowledge: The role of pedagogical content knowledge in learning to reflect on teaching. *International Journal of Science Education, 15*(6), 673-683. https://doi.org/10.1080/0950069930150605

Herráez, A. (2006). Biomolecules in the computer: Jmol to the rescue. *Biochemistry & Molecular Biology Education, 34* (4), 255-261. https://doi.org/10.1002/bmb.2006.494034042644

Johnson, J. A. (1997). Units of analysis for the description and explanation of personality. In R. Hogan, J. Johnson, & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 73-93). Academic Press.

Kennedy, M. (2018, October 15). To prevent wildfires, PG&E pre-emptively cuts power to thousands in

California. *NPR*. https://www.npr.org/2018/10/15/657468903/to-prevent-wildfires-pg-e-preemptively-cuts-power-to-thousands-in-california

Lamanauskas, V. (2019). 3rd international Baltic symposium on science and technology education "Science and technology education: Current challenges and possible solutions (BalticSTE2019)": Symposium review. *Švietimas: politika, vadyba, kokybė / Education Policy, Management and Quality*, *11*(1), 42-48. http://oaji.net/articles/2019/513-1567660630.pdf

Nasledov, A. (2005). *SPSS: komp'juternyj analiz dannyh v psihologii i social'nyh naukah* [SPSS: Computer analysis of data in psychology and social sciences]. Piter.

Novák, M., & Langerová, P. (2006). Raising efficiency in teaching mathematics in non-English speaking countries: An electronic bilingual dictionary of mathematical terminology. In: *Proceedings of 3rd international conference on the teaching of mathematics at the undergraduate level.* Istanbul: TMD (Turkish Mathematical Society), 2006. [CD-ROM].

Posner, M. (2004). Neural systems and individual differences. *TC Record.* http://www.tcrecord.org/PrintContent.asp?ContentID=11663

Šlekienė, V., & Lamanauskas, V. (2019). Sisteminis „judėjimo" sąvokos turinio integravimas, kaip viena iš visuminio gamtamokslinio ugdymo prieigų [Systematic integration of the content of "Movement" concept as one of the approaches to comprehensive natural science education]. *Gamtamokslinis ugdymas / Natural Science Education*, *16*(1), 43-53. http://oaji.net/articles/2019/514-1563213127.pdf

Thurstone, L. L. (1959). *The measurement of attitude: A psycho-social method and some experiments.* Univerity of Chicago.

Vaitkevičius, J. (1995). *Socialinės pedagogikos pagrindai* [Basics of social pedagogy]. Egalda.

Walker, J., Halliday, D., & Resnick, R. (2008). *Fundamentals of physics*. Washington: Wiley.