Developing Machine Learning Algorithms for Predicting House Prices in Surabaya Using IBM SPSS Modeler

Hansel Davin Sugiarto[1][0009-0007-3670-6848], Doddy Prayogo[2][1111-2222-3333-4444], and Njo Anastasia[3][0000-0003-4480-9365]

[1, 2] Civil Engineering and Design, Petra Christian University, Surabaya, Indonesia
hanseldavins@gmail.com
[3] School of Business and Management, Petra Christian University, Surabaya, Indonesia

Abstract. Using PHP and IBM's SPSS Modeler for data mining and analytics, the authors collected and processed housing data from Rumah123 in Surabaya. A total of 10,336 data points were divided into two clusters. Three machine learning (ML) models were developed using SPSS Modeler for 2,460 data points. The results indicate that the Artificial Neural Network (ANN) model provided the most consistent correlation across all scenarios, while the Support Vector Machine (SVM) model performed the worst. The Classification and Regression Tree (CART) model showed good performance in both training and testing for the larger cluster but did not perform as well with the smaller cluster. Overall, ANN and CART models can be used to predict housing prices, with ANN offering higher accuracy.

Keywords: Data Analysis, Web Scraping, Artificial Neural Network, Support Vector Machine, Classification And Regression Tree, Linear Regression.

1 Introduction

The growing demand for housing in Indonesia underscores the essential nature of residential properties [1][2]. Surabaya, a key economic hub in East Java, is recognized for its significant real estate potential in the Asia-Pacific region [3][4]. The city's rapid economic growth and urbanization have led to rising housing prices, highlighting the need for accurate property valuation [5][6]. Surabaya has seen the highest property price increase in Indonesia, with some areas experiencing a 34.88% rise, necessitating careful consideration of various factors in property purchases [7][8][9]. Accurate assessments of property value and lifespan are crucial to avoid speculative pricing and unmet objectives [10][11].

Machine learning (ML) algorithms offer a solution by improving the accuracy of price predictions, addressing the complexities of unique locations [10]. Previous studies have successfully applied ML and statistical methods to predict property prices, such as using artificial neural networks (ANN) in Italy and Spain to assess environmental and location factors [12][13], and the Random Forest (RF) algorithm in China for accurate price predictions [14].

This study aims to utilize AI algorithms and Linear Regression (LR) methods to enhance property valuation techniques in Surabaya, contributing to more reliable and advanced predictive models.

2 Methodology

2.1 Data Collection

The data collection process begins with searching for publicly available online real estate platforms on the internet. Data collection was decided to be performed using web scraping on the "Rumah123" website (https://www.rumah123.com). Rumah123 is an Indonesia online real estate platform that provides information about various types of residential properties, including new homes, resale homes, and second-hand homes. Rumah123 offers extensive and additional information about each property, related to the attributes and characteristics of the properties, such as building age, listing description, certification (SHM, SHGB, etc.), and more. To perform scraping on the Rumah123 site, a web scraping process plan is required to assist in the automatic information retrieval process using PHP programming language (see Figure 1).

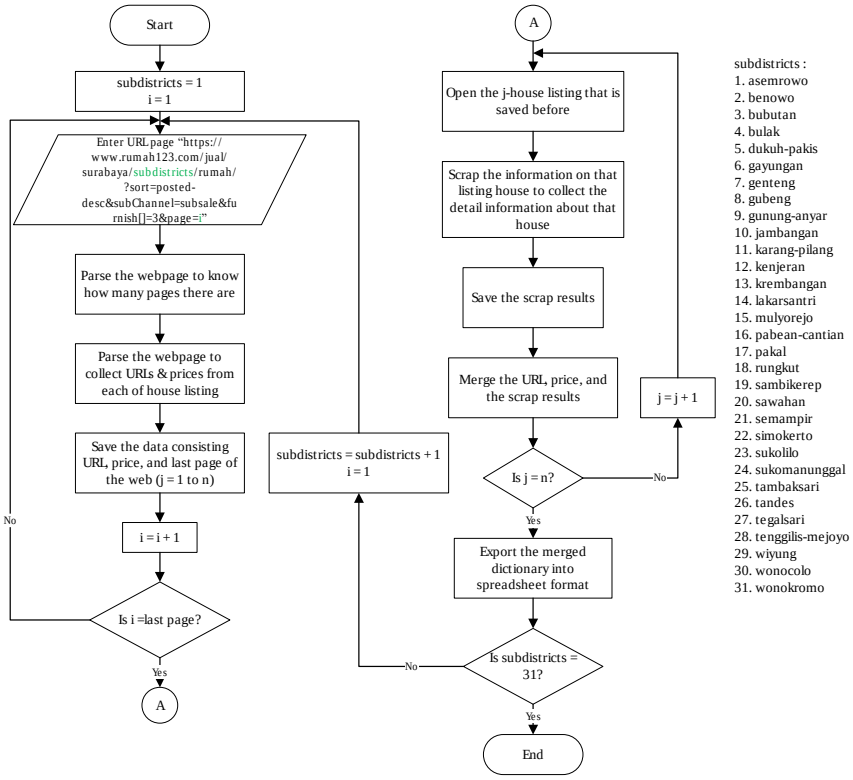

Fig. 1. Data collection planning flowchart

First, to automate the data collection process, it is necessary to understand the URL pattern of the Rumah123 site by entering the location "Surabaya" in the search box. It was found that the pattern for searching residential listings for sale in Surabaya on Rumah123 uses a URL format as illustrated in the flowchart. If "subdistricts" is filled with "asemrowo" and "i" is replaced with "1," the URL will display and filter the residential listings in the Asemrowo subdistrict of Surabaya on page 1. To move to the next page, simply replace "i" with "2," and so on until the last page of the listings.

Once the pattern is understood, the next step is to collect information about the URL links and prices of each house on each page, starting from page 1 to the last page of the listings in the Asemrowo subdistrict. The URL and price information will be stored, and the stored URLs will be automatically opened one by one using web scraping methods. The collected data will then be saved and exported into a spreadsheet file. This spreadsheet file will serve as the dataset for this research. After completing the Asemrowo subdistrict, the same steps will be repeated for the Benowo subdistrict, until for the Wonokromo subdistrict.

2.2 Dataset Merging

Spreadsheet files containing information from each house listing in each sub-district of Surabaya will be combined into one spreadsheet file. This aims to allow the next process to carry out data pre-processing.

2.3 Pre-processing Data

This stage aims to ensure that the data used for training and testing the machine learning model is in optimal condition, thereby improving the accuracy and performance of the model. The data preprocessing process includes several steps: 1) removing duplicate data, 2) handling missing data, 3) clustering data, and 4) evaluating outliers.

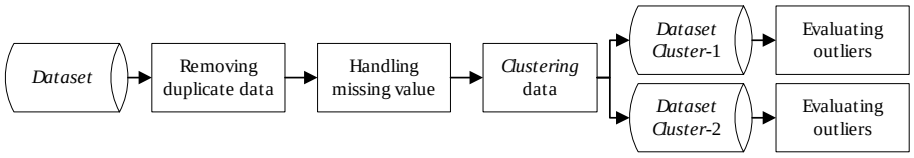

Fig. 2. Data pre-processing planning flowchart

Removing duplicate data. Removing duplicate data is accomplished by filtering the dataset to ensure that no two data entries have identical values across all variables. This process aims to eliminate any duplicate entries, thereby preventing bias during the training process.

Handling missing value. In the data collected through web scraping, not all entries have values for every variable. For independent variables deemed to be of insignificant

importance, variables with more than fifty percent (>50%) missing data will be removed [15]. The case-wise deletion method was chosen because the dataset is extensive and cannot be addressed using other methods such as imputation or predictive techniques.

Clustering data. The clustering process aims to identify and group data based on specific patterns or characteristics of each data point to support learning during the training process. This is necessary because the raw data obtained includes various types of houses, as the research covers not only new homes but also second-hand homes with diverse information and characteristics. Data clustering is performed using the k-means algorithm available in IBM SPSS Modeler 18.0 using 2 clusters and default settings.

Evaluating outliers. IBM SPSS Modeler 18.0 has features to help detect and visualize outliers when the standard deviation value is 3 standard deviations from the mean. The app can show the relation between each variable to price of the house in scatter plot figure. By evaluating these outliers, it is hoped that the performance of statistical and predictive models can be improved.

2.4 Model Tuning

The development of the predictive model is conducted using IBM SPSS Modeler 18.0 (see Figure 3). The first step involves importing the dataset, which has been divided into two clusters, into the IBM SPSS Modeler 18.0 program. The uploaded dataset is then randomly split with a 70:30 ratio [16].

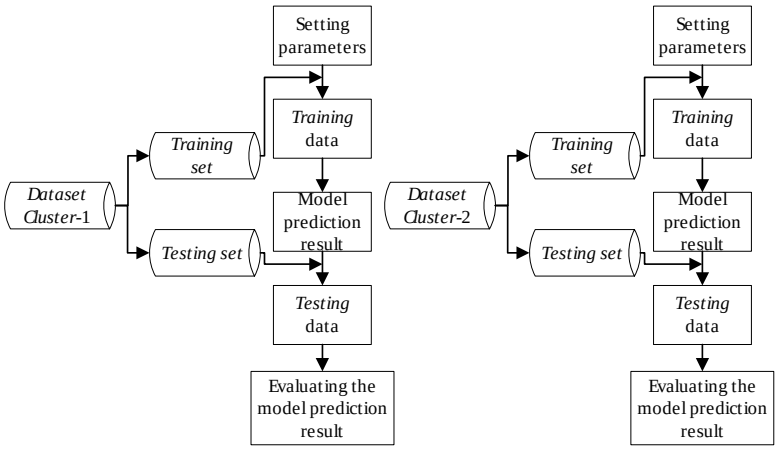

Fig. 3. Data analysis and prediction model development flowchart

After dividing the training and testing sets, parameters are set to find the predictive model for each ML algorithm: 1) Artificial Neural Network (ANN), 2) Support Vector

Machine (SVM), 3) Classification And Regression Tree (CART) and 4) Linear Regression (LR), in IBM SPSS Modeler 18.0 using standard model settings. After setting the parameters, the training process is conducted on 70% of the data, resulting in a predictive model. This predictive model is then tested using the testing set. The testing process involves applying the predictive model generated during training to the remaining 30% of the data.

2.5 Evaluation of Prediction Model Results

After testing, the results will be compared using predictive evaluation to determine which algorithm and parameters produce the most accurate outcomes. The evaluation will use four measurement methods: linear correlation, mean absolute error, mean absolute percentage error, and root mean squared error.

Linear correlation (R) measures the strength and direction of the linear relationship between two variables, and in the context of AI model evaluation, R quantifies how well the model's predictions correlate with actual values. Mean Absolute Error (MAE) is an evaluation metric that measures the average of all absolute errors between predicted values and actual values, providing an indication of how much error the model makes in its predictions. Mean Absolute Percentage Error (MAPE) measures the average absolute error as a percentage of the actual values, facilitating the comparison of prediction errors across different data scales, with a lower MAPE value indicating more accurate predictions. Root Mean Squared Error (RMSE) measures the square root of the average squared errors between predicted values and actual values, offering an indication of how large the model's prediction errors are on average, in the same units as the dependent variable.

3 Results

3.1 Dataset

The total dataset obtained through web scraping from Rumah123 comprises 10,336 data points with 32 independent variables and 1 dependent variable (house price). A detailed explanation and the percentage of missing data for each variable are presented in table 1.

Variables with more than 50% missing data were removed. Subsequently, for variables still containing missing data, entire rows with missing data were deleted, resulting in a dataset with no missing data. After this process, the dataset consisted of 2,562 data points with 16 input variables and 1 target variable (house price).

After data pre-processing and clustering, cluster-1 contains 2,155 data points, and Cluster-2 contains 407 data points. The ratio difference between the clusters is 5.29, with the smallest cluster representing 15.9% and the largest cluster representing 84.1% of the total dataset. After handling the outliers in each cluster, cluster-1 contained 2,070 data points and cluster-2 contained 390 data points, resulting in a total of 2,460 house data points. These data were then randomly divided with a 70:30 ratio for training and testing purposes.

Table 1. Variables explanation and percentage of missing data in raw dataset

No.	Variable	Explanation	%Missing
1	Price	House listing price	0.00
2	Kecamatan	House listing subdistrict	0.00
3	URL	House listing URL	0.00
4	Kamar Tidur	Number of bedrooms	1.45
5	Kamar Mandi	Number of bathrooms	0.97
6	Luas Tanah	Surface area	0.03
7	Luas Bangunan	Surface area of building	0.14
8	Carport	Number of carports	41.92
9	Tipe Properti	Landed House	0.01
10	Sertifikat	House certificate	0.07
11	Daya Listrik	Electric power	23.43
12	Dapur	Number of kitchens	51.98
13	Ruang Makan	Dining room	26.69
14	Ruang Tamu	Living room	6.47
15	Kondisi Perabotan	Furnished/semi-furnished/unfurnished	0.08
16	Material Bangunan	Building material	74.89
17	Jumlah Lantai	Number of floors	0.61
18	Hadap	House orientation	34.79
19	Konsep dan Gaya Rumah	House concept and style	58.55
20	Terjangkau Internet	Internet covered	6.46
21	Lebar Jalan	Road width (measure in number of cars)	48.08
22	Tahun Dibangun	Year of built	68.19
23	Sumber Air	Water source	33.32
24	Hook	Is the house positioning on hook?	6.45
25	Kondisi Properti	Property condition	4.28
26	ID Iklan	Listing ID	0.01
27	Material Lantai	Floor material	72.04
28	Pemandangan	House view	57.21
29	Tahun di Renovasi	Year of renovation	87.50
30	Garasi	Number of cars that can fit in garage	75.77
31	Kamar Pembantu	Number of maid's bedrooms	58.67
32	Kamar Mandi Pembantu	Number of maid's bathrooms	60.19
33	Periode Sewa	Leasing period	99.98

3.2 Model Comparison

The evaluation of training results was performed on 70% of the data selected randomly, and the evaluation of testing results was performed on 30% of the data selected randomly.

Table 2. Models Performance

Model	Eval. Method		Cluster-1		Cluster-2	
			Training	Testing	Training	Testing
ANN	R		0.887	**0.842**	0.828	0.716
	R^2		0.787	**0.709**	0.686	0.513
	MAE	Rp	634,118,868	**694,004,310**	530,969,736	1,004,693,267
	MAPE		28.139%	**29.102%**	25.811%	28.917%
	RMSE	Rp	1,125,303,772	**1,207,978,827**	940,286,118	2,831,965,786
SVM	R		0.467	0.492	0.354	0.256
	R^2		0.218	0.242	0.125	0.066
	MAE	Rp	1,327,826,130	1,295,049,219	1,041,501,881	1,659,387,582
	MAPE		47.613%	49.224%	44.953%	28.917%
	RMSE	Rp	2,559,713,915	2,302,852,491	1,716,927,994	3,568,222,439
CART	R		0.724	0.760	0.855	0.695
	R^2		0.524	0.578	0.731	0.483
	MAE	Rp	804,371,767	782,796,106	502,416,068	1,015,937,131
	MAPE		31.748%	31.440%	25.014%	30.646%
	RMSE	Rp	1,681,792,437	1,422,176,803	859,302,441	2,505,443,044
LR	R		0.825	0.831	0.820	**0.847**
	R^2		0.680	0.691	0.672	**0.717**
	MAE	Rp	752,856,799	721,174,520	542,948,202	**948,439,392**
	MAPE		31.484%	30.534%	24.315%	**28.032%**
	RMSE	Rp	1,378,492,121	1,221,587,390	946,226,624	**2,137,342,585**

4 Discussion and Conclusion

Currently, with the rapid economic development, housing prices in Surabaya are steadily rising. Various factors influence these prices, which can be categorized into internal and external factors. External factors such as population, policies, and other macroeconomic reasons are beyond our control and judgment, so we can only analyze the aspects that are within our knowledge. In this paper, we take Surabaya as an example. We use PHP to scrap the relevant housing information from Rumah123. After cleaning and filtering, the data are analyzed and different machine learning models are established for predictive analysis. Three different types of Machine Learning methods including ANN, SVM, and CART and one traditional technique, LR, are compared and analyzed for optimal solutions.

In the training phase, ANN demonstrated the best correlation and accuracy in cluster-1, closely followed by LR, with minimal differences in MAPE values between CART and LR. For cluster-2, CART showed the highest correlation, though not significantly different from ANN, while SVM performed the worst in both clusters, with high error rates.

Testing results for cluster-1 mirrored the training outcomes, with ANN outperforming other models and SVM showing poor performance. In cluster-2, LR excelled in testing with the highest correlation and lowest MAPE, while SVM continued to underperform with low correlation values.

Future research can focus on developing a more comprehensive dataset by expanding data collection from multiple sources, not limited to just one website but also incorporating various platforms and information. This approach would create a more robust dataset. Furthermore, in addition to analyzing internal factors, future studies could integrate external factors such as macroeconomic conditions and changes in government policies, which can then be applied to the predictive model to identify an AI model that is suitable for further analysis.

References

1. Universitas Airlangga Repository, http://repository.unair.ac.id/id/eprint/62883, last accessed 2024/06/01.
2. Halim, K. C., Rahardjo, J., and Utomo, C. Factors in shaping property prices in Surabaya. Dimensi Utama Teknik Sipil 11(1), 19-36 (2024).
3. Sari, Y. Empirical study of the application of value management in real estate development in Surabaya. NALARs 16(1), 85-90 (2017).
4. Yasmin, S. Cengriani, J. and Ariyah, M. R. A. The potential of Surabaya as a MICE tourist destination. Nusantara Journal of Multidisciplinary Science 1(5), 1368-1378 (2023).
5. Putra, H. Economic growth and residensial property demand in Surabaya. Journal of Urban Development 12(1), 78-90 (2021).
6. Sari, M. Urbanization and housing price trends in Surabaya. Journal of Real Estate Research 11(2), 112-125 (2020).
7. Yusuf, A. The role of Surabaya as a business and educational hub in property market dynamics. Jurnal Ekonomi dan Pembangunan 9(3), 45-57.
8. Gunawan, K. Analysis of the desire to purchase middle-class residential property in the city of Surabaya. Petra Business and Management Review 4(1), 89-100 (2018).
9. Kompas Webpage, https://www.kompas.com/properti/read/2022/02/18/090000421/walau-harga-rumah-terus-naik-bubble-properti-diprediksi-tak-akan?page=all, last accessed 2024/06/14.
10. Pratama, I. W. Implementation of support vector machine in house price prediction. Jurnal Akademisi Vokasi 2(2), 101-113 (2023).
11. Sujono, B. Asset valuation in the property sector. MODUL 11(1), 37-40 (2011).
12. Chiarazzo, V. Caggiani, L., Marinelli, M., and Ottomanelli, M. A neural network based model for real estate price estimation considering environmental quality of property location. Transportation Research Procedia 3, 810-817 (2014).
13. Duran, L. F., Valero, S., Llorca, A., Botti, V.: The impact of location on housing prices: Applying the artificial neural network model as an analytical tool. In: 51st European Regional Science Association Conference, pp. 1595-1620. European Regional Science Association, Barcelona (2011).
14. Zhang, Y., Huang, J., Zhang, J., Liu, S., Shorman, S. Analysis and prediction of second-hand house price based on random forest. Applied Mathematics and Nonlinear Sciences 7(1), 27-42 (2022).
15. Quang, T., Nguyen, M., Dang, H., Mei, B. Housing price prediction via improved machine learning techniques. Procedia Computer Science 174, 433-442 (2020).
16. Gholamy, A., Kreinovich, V., Kosheleva, O. Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation. International Journal of Intelligent Technologies and Applied Statistics 11(2), 105-111 (2018).