Property Category Prediction Model using Random Forest Classifier to Improve Property Industry in Surabaya

Yosua Setyawan Soekamto(✉), Michelle Chandra, Trianggoro Wiradinata, Rinabi Tanamal, and Theresia Ratih Dewi Saputri

Universitas Ciputra Surabaya, Surabaya, Indonesia
yosua.soekamto@ciputra.ac.id

Abstract. Urban planning is done not only to regulate residential areas, offices, retail spaces, and green spaces but also to ensure that people (community) who live in cities have a decent quality of life. Surabaya is a city that was built in the beginning of Indonesian civilization, so the arrangement of the city of Surabaya is a bit difficult and has an impact on housing costs. In reality, housing development is influenced by businesses in the residential development sector. This causes uneven house types to be built in accordance with the expectations of the government, which could impact the sustainability of Surabaya. This study is crucial because, from the data of Bank Indonesia, in supply and demand index for the property sector in Surabaya has not increased since 2019. Although property price has decreased since the fourth quarter of 2019 because of the Covid 19 pandemic, the demand index has not increased that well. This study intends to assist the process of classifying house types, so the government can make a selection on the house that will be built by the developer. 14 input attributes and 490 data from Surabaya property agencies were used in this study. In this study, random forest is used as the classification technique. The result of the classification model obtained an accuracy value of 89% and F1 score of 89%. A classification prediction model that can be used to determine property classification was found through this study.

Keywords: Classification · Data Science · Predictions · Property Industry · Random Forest · Sustainable City

1 Introduction

In essence, living things reproduce to maintain their species and also to increase the number of species. This earth is filled with 3 categories of living things, that is humans, animals, and plants. That three living things mainly live on land, which land area on earth is about 149 million square kilometers. The human population data in 2021 will reach around 7,8 billion with a growth rate of 0.9%. The human population in Indonesia in 2021 is recorded at around 276 million people or about 28% of the total population in the world. Meanwhile, the land area owned by Indonesia is around 1.9 million square kilometers or about 78% of the total land area in the world [1].

S. Jahroh et al. (Eds.): BIEC 2022, AEBMR 236, pp. 256–265, 2023.
https://doi.org/10.2991/978-94-6463-144-9_24

Between the three categories of living things, human needs settlements and civilizations for shelter and social life, and plants need an area with good soil to thrive, while animals do not need a specific environment to survive. According to that needs, the land area needed a good strategy to regulate it.

Urban planning is done to ensure that people (community) who live in the city have a decent quality of life. It is also done to control residential areas, offices, retail stores, and green spaces [2].

Surabaya is one of the biggest cities in Indonesia located in the East Java province. Surabaya was built more than 700 years ago, and because of that the structuring Surabaya city has its complexity. Several strategies were done by the government, such as structuring the green lane in the city and activating several units of the Suroboyo Bus as mass transportation in the city [3] and many other initiatives.

Even if the strategies of urban planning were done, the population growth of Surabaya still give an impact on increasing the need for housing (house). The area of Surabaya and the high demand for housing have led to the increase in housing prices, so it is so common for some families to live together [4].

The government has set strategies to regulate the distribution of housing, that is determining the type or category of housing (house). In the regional government regulation, it is agreed that the type of house is divided into three categories, namely homely, moderate, and luxurious houses [5]. In general, the house categories are divided based on the building area. In practice, residential development is influenced by businesses in the residential development sector [6]. This causes housing type development is not well-distributed following the expectations of the government, which may have an impact on the sustainability of Surabaya city.

This study is important to be done because due to the data from the Bank of Indonesia, in supply and demand index for the property sector in Surabaya has not increased since 2019. Even though property prices have decreased since the fourth quartal of 2019 due to the covid-19 pandemic, the demand index is also not increasing well [7]. This will have a negative impact on the city and the nation's sustainability, especially because it may lead to a financial crisis [8, 9].

Based on the problems that have been mentioned, a system is needed to make classify house types that will be built by the developer. The purpose is to select and contribute to the sustainable city planning of Surabaya.

2 Literature Review

A city is an area within the country that contain various function, such as residential and center of the education activity, culture, economy, and regional government. The people (community) can develop if the environment or city they live in is also developing. On the other hand, the city can develop along with the increasing population and the community's economy. A city is considered sustainable if there is a balance between the growth of its population (settlement), the number of workers, the number of public facilities, and the transportation infrastructure [10]. If the population is increased, but there is also much unemployment, then the economy in that city will be affected. Furthermore, if a city lacks public facilities, like schools or education, shopping stores, and

recreation area, it is people will find it difficult to meet their basic needs. Especially if the infrastructure and transportation in a city are insufficient, people will have difficulty mobility.

The main task of the government, particularly regional governments, is to organize cities. The purpose to organize this city is to maintain its sustainability of the city and certainly to develop it into a better city. As the city planner, governments need to regulate the area distribution and the number of public facilities within a certain radius. The actions taken by the government are creating Regional Government Regulations that are regularly repaired or updated, build state educational facilities, hospitals, and markets [11].

The Surabaya city government can organize the city more effectively by utilizing technology. Through the Surabaya Command Center, the government can control the transportation and infrastructure systems [12]. This is an effort by the government to make Surabaya into a more sustainable city. Other efforts are required to regulate residential areas because the distribution of settlements is still not evenly distributed. This regulation cannot be done solely by relying on local government regulations, but also requires technological assistance such as the Surabaya Command Center [13, 14].

One technology that can support managing residential areas is utilizing machine learning on existing residential data. Machine learning itself is a method in information technology to enable computers to learn independently. In general, computers can learn from existing data or expert experience [15]. In this study, supervised learning methods are prioritized as machine learning. In the supervised learning method, the system will be given data that already has a target class label. This label has been pre-determined by property agents as experts in determining the types of houses. The computer will learn from and search through this labeled data for knowledge. Later, new data will be classified using this knowledge [16].

Until now supervised machine learning has become more frequently used to predict real estate prices. Models for predicting the type of a house from previous studies were constructed from popular classification algorithms such as Logistic Regression, Naïve Bayes, K-NN, Decision Tree, Random Forest, and others [17]. In this study, house type prediction Random Forest classification was selected due to its performance compared to other supervised machine learning algorithm.

Random forest is one of the machine learning algorithms that employ an ensemble method approach applied in the decision tree, which is a well-known method used by many academics and industries. Random forest is suitable to be used for classification and regression cases with a large dataset. This method is called forest because it trains on several decision trees [18] and the prediction result can be made by combining the prediction of the ensemble [19]. Because it consists of many decision trees, the accuracy results is expected to be higher than a decision tree and relatively more resistant to overfitting [20]. By using this machine learning method, it can also produce low bias trees 381038290500and high variance.

In this house types prediction research, is used classification algorithm because the result from the prediction is discrete values. In a random forest, the result of classification cases can be determined by the voting of each tree and taking the majority vote [21].

3 Data Structure and Preprocessing

The data used in this study is the data property (house) in Surabaya city. The process of collecting this dataset was done in collaboration with property agents from property companies in the Surabaya area. The total dataset collected from the sources is 490 data that are being listed or have been sold between 2019 and 2021.

3.1 Data Structure

As seen in Table 1, the dataset used in this study has 17 attributes that consist of 16 input features and 1 class target. Before further analyzing the data, it is necessary to perform data cleaning and to pre-process the data before it can be used in the classification method and to ensure that the data is of good quality.

The attribute that becomes the class target is classified into 3 classes, which are homely, moderate, and luxurious. The attribute community price is not used in this research due to the multicollinearity problem between community price and price. These highly correlated features will not need to be removed.

3.2 Data Preprocessing

Data preprocessing will perform the process of data cleaning, which includes checking for missing values. After checking, it was found that two attributes contain missing values, those attributes are the *building age* and *urgent*. SimpleImputer library was used to handle the missing values by replacing the missing values using attribute's central tendencies.

Table 1. Property Dataset

No	Attributes			No	Attributes		
	Name	Type	Data Type		Name	Type	Data Type
1	Cluster Name	Input	Object	10	Facing	Input	Object
2	Surface Area	Input	Integer	11	House Position	Input	Object
3	Building Area	Input	Integer	12	Road Width	Input	Integer
4	Bedrooms	Input	Integer	13	Urgent	Input	Object
5	Bathrooms	Input	Integer	14	Building Age	Input	Integer
6	Storey	Input	Integer	15	Ready to Use	Input	Object
7	Community Price	Input	Integer	16	Furnished	Input	Object
8	Price	Input	Integer	17	Category	Class Target	Object
9	Ownership Status	Input	Object				

Table 2. The Result of Encoded Category

Category	Encoding
Homely	1
Moderate	2
Luxurious	3

Table 3. Evaluation of Each Classifier

Classifier	Evaluation	
	Accuracy	F1-Score
K-Nearest Neighbour	78%	78%
Decision Tree	86%	86%
Support Vector Machine	88%	88%
Logistic Regression	88%	88%
Random Forest	89%	89%

Other than data cleaning, the existing data will go through the encoding process. The encoding process is divided into two ways, the Label Encoder for nominal scale data and the Map for ordinal scale data. The map method encodes the target class attribute, as shown in Table 2.

4 Data Analysis

Next, data analysis will be done using exploratory data analysis and model evaluation. It is hoped that by analyzing the data, this research can uncover helpful information that can be used to support prediction result as well as anticipating anomalies Table 3.

4.1 Exploratory Data Analysis

The exploratory data analysis process starts with analyzing the balance of the dataset. The analysis found that most of the houses listed for sale are homely houses by 54.5% (267 data). Meanwhile, moderate house and luxurious house that was marketed is about 35.1% (172 data) and 10.4% (51 data). Figure 1 shows these results.

Oversampling using Synthetic Minority Oversampling Technique (SMOTE) was used because of the imbalance in the data shown in the target class. The SMOTE can make the data balance by randomly replicating the minority data to be as much as the majority data. As a result, each class has the same number of data, 267 observations.

The next step is to observe the attributes that have a strong correlation with the class target, then take a closer look at the data characteristic and the spreading of the data in each attribute. From the result of the correlation in Fig. 2, can be found that the attribute

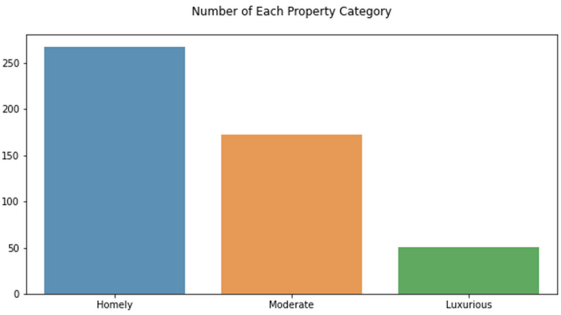

Fig. 1. Number of Each Property Category

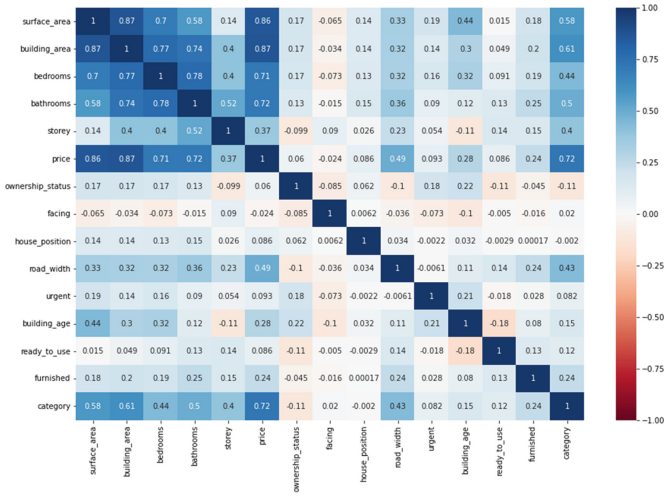

Fig. 2. Attributes Correlation

of price, building area, surface area, bathrooms, bedrooms, storey, and road width have a strong correlation to class target.

As an analysis result, the mean of the price attribute is 4.5 billions and the median is 2.5 billions. Then the median of the building area is 178.5 m^2, while the surface area has a median of 155 m^2. Further analysis is the attributes of storey, bathrooms, and bedrooms that have a strong correlation with the class target. From the dataset is found that most of the houses for sale had two stories, three bathrooms, and four bedrooms Figs. 3, 4 and 5.

The analysis was also done for the road width attribute by observing the spread in each class. It is found out that more of the houses have road width more than two cars and 1–2 car(s).

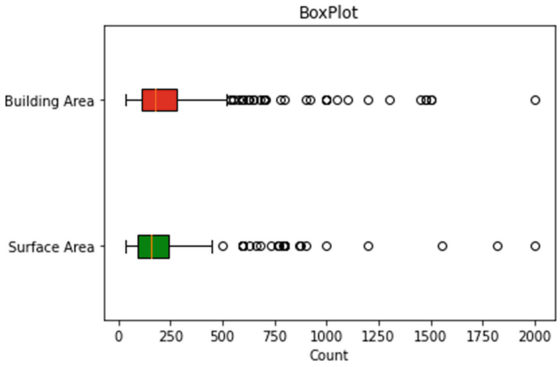

Fig. 3. Boxplot of Building Area and Surface Area

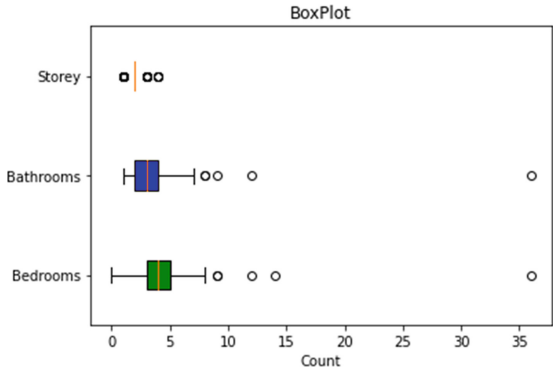

Fig. 4. Boxplot of Storey, Bathrooms, and Bedrooms

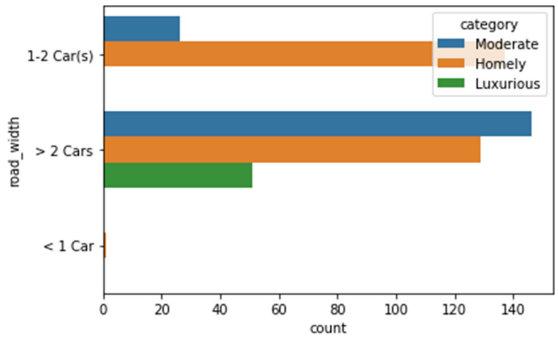

Fig. 5. Boxplot of Storey, Bathrooms, Bedrooms

4.2 Model Evaluation

In this study, several classification models were compared, and calculate the evaluation model. K-Nearest Neighbor, Logistic Regression, Support Vector Machine, Decision

Tree, and Random Forest were used as model classifiers. Random forest shown to be the best classifier with the highest accuracy and f1-score, which is 89%.

Accuracy is the ratio of predictions, both true positive and true negative, from all data. While F1-score is a harmonic mean from the comparison between recall and precision.

5 Findings and Discussion

According to the data analysis, the dataset contains more homely than moderate or luxurious class target. Because this study discovered imbalanced data in each category, the SMOTE method was used for oversampling. SMOTE was chosen because it can increase the minority class to be the same as the majority class based on the K-nearest neighbor [22]. According to the analysis, the price, building area, surface area, bathrooms, bedrooms, and storey attributes strongly correlate to the class target category.

Another finding is that most of the properties that have been listed have a mean price that is 4.5 billion, the median attribute of building area is 178.5 m^2, and the median attribute of the surface area is 155 m^2. However, Outlier data with the attribute of building area greater than 500 m^2 and surface area greater than 400 m^2 were also discovered. This is because most of the target classes of the dataset are homely. Moreover, most houses for sale have 2 storeys, 3 bathrooms, and 4 bedrooms.

Evaluation of the random forest model resulted in higher accuracy and F1-score than other classifiers, so it was chosen as the classifier in this study. Random forest generates an accuracy of 89% and an F1-score of 89%. Even if the data analysis contains many outliers, the accuracy results can still be good because, as previously stated, random forest is resistant to overfitting.

6 Conclusion

Surabaya city is developing and has become home to many families. One of the problems that arise in the property field is the development of property types that are not evenly distributed in the Surabaya area. This is because residential development is influenced by businesses in the residential development sector and the problems of financial crises that often occur.

To help the governments to plan the city to achieve a sustainable city, this study performs a classification prediction test using random forest as machine learning algorithms and it was discovered that the accuracy score and F1 score of 89% were satisfactory.

Acknowledgments. This research article and research study are funded by the Indonesian Ministry of Research, Technology, and Community Service, Indonesian Ministry of Higher Education, Research, and Technology, and Indonesian Ministry of Education, Culture, Research, and Technology.

References

1. World Data.info by eglitis-media, "World Population Growth 2012 - 2021," Egilitis-Media. .
2. H. Ahvenniemi and A. Huovila, "How do cities promote urban sustainability and smartness? An evaluation of the city strategies of six largest Finnish cities," Environ. Dev. Sustain., vol. 23, no. 3, pp. 4174–4200, 2021, doi: https://doi.org/10.1007/s10668-020-00765-3.
3. A. ARY KURNIAWAN and I. PRABAWATI, "Implementasi Suroboyo Bus Di Dinas Perhubungan Kota Surabaya," Publika, vol. 6, no. 9, 2018.
4. E. Hutapea, "Surabaya Catat Kenaikan Harga Rumah Tertinggi di Indonesia," Kompas.com, 2018. .
5. E. Fitriana, "Implementasi Kebijakan Tata Ruang Wilayah Dalam Mewujudkan Pembangunan Kota Berkelanjutan (Studi Di Kabupaten Magetan)," J. Adm. Publik Mhs. Univ. Brawijaya, vol. 2, no. 2, pp. 217–223, 2014.
6. S. Aminah, "Konflik dan Kontestasi Penataan Ruang Kota Surabaya," Masy. J. Sosiol., vol. 20, no. 1, pp. 59–79, 2016, doi: https://doi.org/10.7454/mjs.v20i1.4751.
7. "Laporan Perkembangan Property Komersial," Bank Indonesia, 2022. .
8. Putri Setyaningsih, "Pasar Properti Residensial Di Tengah Pandemi Covid-19," Kementrian Keuangan Republik Indonesia, 2021. .
9. Mahdaniar Maulidini Muhyi and Joko Adianto, "Literature Review: The Effects of Covid-19 Pandemic-Driven Home Behavior in Housing Preference," Smart City, vol. 1, no. 1, pp. 0–15, Singapore 43 59https://doi.org/10.1007/978-981-16-1357-9_3
10. H. Ahvenniemi, A. Huovila, I. Pinto-Seppä, and M. Airaksinen, "What are the differences between sustainable and smart cities?," Cities, vol. 60, pp. 234–245, 2017, doi: https://doi.org/10.1016/j.cities.2016.09.009.
11. A. Olivier et al., "Data analytics for improved closest hospital suggestion for EMS operations in New York City," Sustain. Cities Soc., vol. 86, no. February, p. 104104, 2022, doi: https://doi.org/10.1016/j.scs.2022.104104.
12. "Command Center-Pelayanan Terpadu di Siola Diplot jadi Percontohan Nasional," Pemerintah Kota Surabaya, 2022. .
13. M. Castells, "Urban sustainability in the information age," City, vol. 4, no. 1, pp. 118–122, 2000, doi: https://doi.org/10.1080/713656995.
14. L. Yang, Y. Chen, N. Xu, R. Zhao, K. W. Chau, and S. Hong, "Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture," Sustain. Cities Soc., vol. 58, no. March, p. 102140, 2020, doi: https://doi.org/10.1016/j.scs.2020.102140.
15. G. Giray, "A software engineering perspective on engineering machine learning systems: State of the art and challenges," J. Syst. Softw., vol. 180, p. 111031, 2021, doi: https://doi.org/10.1016/j.jss.2021.111031.
16. N. Wijaya, W. Ordiyasa, and A. F. Rachman, "Evaluation of Naïve Bayes and chi-square performance for classification of occupancy house," Int. J. Informatics Comput., vol. 1, no. 2, 2019.
17. Y. Widiastuti, S. W. Sihwi, and M. E. Sulistyo, "Decision Support System for House Purchasing Using Knn (K-Nearest Neighbor) Method," J. Itsmart, vol. 5, no. 1, pp. 43–49, 2016.
18. M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 13, pp. 6308–6325, 2020, doi: https://doi.org/10.1109/JSTARS.2020.3026724.
19. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," IJCSI Int. J. Comput. Sci. Issues, vol. 9, no. 5, pp. 272–278, 2012.

20. G. Teles, J. J. P. C. Rodrigues, R. A. L. Rabêlo, and S. A. Kozlov, "Comparative study of support vector machines and random forests machine learning algorithms on credit operation," Softw. - Pract. Exp., vol. 51, no. 12, pp. 2492–2500, 2021, doi: https://doi.org/10.1002/spe.2842.
21. M. Čeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments," ISPRS Int. J. Geo-Information, vol. 7, no. 5, 2018, doi: https://doi.org/10.3390/ijgi7050168.
22. S. S. Patil, "Enhanced SMOTE Algorithm for Classification of ImbalancedBig-Data using Random Forest," pp. 403–408, 2015, [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7154739.
23. A.- Amrin and O.- Pahlevi, "Implementation of Logistic Regression Classification Algorithm and Support Vector Machine for Credit Eligibility Prediction," *J. Informatics Telecommun. Eng.*, vol. 5, no. 2, pp. 433–441, 2022, doi: https://doi.org/10.31289/jite.v5i2.6220.
24. 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 Comput. Sci.*, vol. 199, pp. 806–813, 2021, doi: https://doi.org/10.1016/j.procs.2022.01.100.
25. S. Amershi *et al.*, "Software Engineering for Machine Learning: A Case Study," *Proc. - 2019 IEEE/ACM 41st Int. Conf. Softw. Eng. Softw. Eng. Pract. ICSE-SEIP 2019*, no. 1, pp. 291–300, 2019, doi: https://doi.org/10.1109/ICSE-SEIP.2019.00042.
26. Sinha, A. (n.d.). Utilization Of Machine Learning Models In Real Estate House Price Prediction. *Amity Journal of Computational Sciences (AJCS)*, *4*(1), 18–23. www.amity.edu/ajcs.
27. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.