Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

Land Price Classification Map in Jakarta Using Random Forest and Ordinary Kriging

Naufal Alvin Chandrasa*, Sri Suryani Prasetyowati, Yuliant Sibaroni

School of Computing, Informatics, Telkom University, Bandung, Indonesia
Email:[1,*]naufalalvinchandrasa@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id,
[3]yuliant@telkomuniversity.ac.id
Email Penulis Korespondensi: naufalalvinchandrasa@student.telkomuniversity.ac.id
Submitted:**20/07/2022**; Accepted:**14/08/2022**; Published**: 30/09/2022**

Abstract−This research provides information about land prices in Jakarta by classifying using the Random Forest method. Where Random Forest is a data mining technique that is usually used to perform classification and regression. Random Forest is one of the best classification methods. It is found that classification accuracy will increase dramatically as a result of voting to select class types and ensemble tree growth. The method helps in providing information about the classification of land prices with the class of land prices per meter less than IDR 15 million, land prices per meter with a price range of IDR 15 to 25 million and land prices per meter more than IDR 25 million. With a fairly good accuracy of 82%, this method can classify where the permeter land price data that is tested will match the predicted classification accurately. Classification is performed on unbalanced data which is then oversampled using the ADASYN method. Assisted by doing spatial interpolation with the Ordinary Kriging method using Semivariogram, information about the classification of land prices can be seen on the distribution of the Jakarta area map. Ordinary Kriging can predict the estimated price per meter of land around the area of land that has a known price. The Root Mean Square Error (RMSE) results of the best Semivariogram model are obtained from the lowest RMSE value, namely the Spherical model with a value of 1.014896e7. The contribution of this research is to provide information about a reliable classification method, namely Random Forest and Ordinary Kriging performance as a spatial analysis method that can predict land prices per meter at unknown points so as to provide information about the distribution of land prices in Jakarta with each price class.

Keywords: Land Price; Jakarta; Classification; Ranfom Forest; Ordinary Kriging

1. INTRODUCTION

According to Rumah.com Indonesia Property Market Index (RIPMI) the results of the analysis of 700,000 property listings for sale and rent throughout Indonesia get more than 17 million visits and are accessed by more than 5.5 million property seekers each month [1]. Jakarta, the administrative center of Indonesia is currently one of the largest urban area in Southeast Asia and is constantly growing in size and population, posing a problem for the Jakarta government[2]. Looking at the UHI (Urban Heat Island) from 2008 to 2018, it can be seen that the city of Jakarta shows a warm area, this is in line with the increasing population growth rate and also urbanization that occurs in Jakarta and also land use in Jakarta[3]. As population levels rise, the need for land and property increases. Land use in cities are typically residential, commercial and services, industry and warehousing, and transportation facilities. According to the neoclassical approach, firms and households have specific demand for their location such as, proximity to urban services, availability of amenities, environmental quality, social factors, and transportation, among other attributes, influence the concept of hedonic land prices. Therefore, information about land price classifications and land price maps will help land seekers to get the land they need.

Several methods were used to classify land prices, namely Random Forest, BP neural network, and Support Vector Machine in research [4] by Chai Shousong *et al.* in 2019. From the comparison of the three methods, a good classification accuracy was obtained where the average accuracy was 90.28%. The author of the paper concludes that all three methods including random forest can perform classification and valuation appropriately. In research [5] 2018, Nejdet Dogru and Abdulhamit Subasi conclude that random forest can be used and get better performance results than ANN and SVM with an accuracy of 91.56%, the study uses Random Forest to solve the classification case, where in the case the classification of Traffic Accident Detection was carried out. Also found the use of Random Forest for classification in the research [6] by Rodriguez-Galiano *et al.* in 2012, the non-parametric nature of RF, their excellent classification accuracy, and their capacity to determine variable importance are some of RF key advantages. The RF algorithm produces reliable land cover classifications, with an overall accuracy of 92%, according to the results. In research [7], Abdulkareem and Abdulazeez in 2021 found that Random Forests are shown fast to build and even faster to predict, where the author collected several related studies and found the accuracy of Random Forests to be quite high. In paper [8], James Magidi *et al* use Random Forest to classify Irrigated Areas problem, which gets an overall classification accuracy of 88% in 2021. For spatial analysis, research in 2021[9], Massimiliano Molinari *et al.* obtained results on the spatial analysis method that Ordinary Kriging even in situations when measures are spatially non-uniformly dispersed and there is a relatively low measurement density, ordinary kriging typically captures the distinctive qualities of crowdsourced measurements well. In other research by Derya Ozturk and Fatmagul Kilic [10] in 2016, geostatistical approaches have been employed in recent years to determine the geographical distribution of meteorological data, and the Ordinary Kriging method is currently the preferred choice in the literature. In this study [11], Ghiasi and Nafisi in 2016 said that the results of the strain tensor elements interpolation using these three techniques which are kriging, spline, and linear show that, on average, the results of Ordinary Kriging are 70% better than those of the spline and linear interpolation techniques. Some of the papers and researches above reveal that

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

Random Forest is one of the machine learning methods that has good performance in classification, and for spatial analysis, several papers and studies have concluded that ordinary kriging is one of the reliable spatial analysis methods. Some of those aformentioned papers have not classified the land price in Jakarta and/or some of them have not done mapping.

The purpose of the following research is to provide information about land prices, especially in the city of Jakarta, where currently the level of land demand is quite high due to the parallel high population growth. With this research the author makes a classification of land prices and also mapping the distribution of land prices to make it easier for land seekers to see differences in land prices in each area in the city of Jakarta. Using the Random Forest classification machine learning model is expected to provide information on land prices from each class. In addition, the mapping containing the distribution of land prices can show the distribution of land prices with different colors on the visualization of the land made according to their respective prices. For the distribution of land prices can be obtained from spatial analysis with the Ordinary Kriging method. Ordinary Kriging can predict unknown land prices around points with known prices.

2. RESEARCH METHODOLOGY

2.1 Research Phases

The initial phase in this research is to scrape data from the land sales website, after which it is necessary to label the attributes on the preset dataset in order to obtain classification results with good accuracy. Preprocessing is done once the dataset is prepared to allow the machine learning model to properly process the data. Data splitting is done next, and the train data is used for Random Forest classification modeling. The modeling outputs are then compared with the labels from the test dataset. The review is then conducted to determine accuracy. Taking action is done if imbalanced data is still discovered. The data is then spatially examined using the best Semivariogram model and Ordinary Kriging, which is based on the modeling results, then visualizing the spatial analysis results to the Jakarta map with a heatmap. The experiment conducted can be seen from Figure 1 Experiment Flowchart.

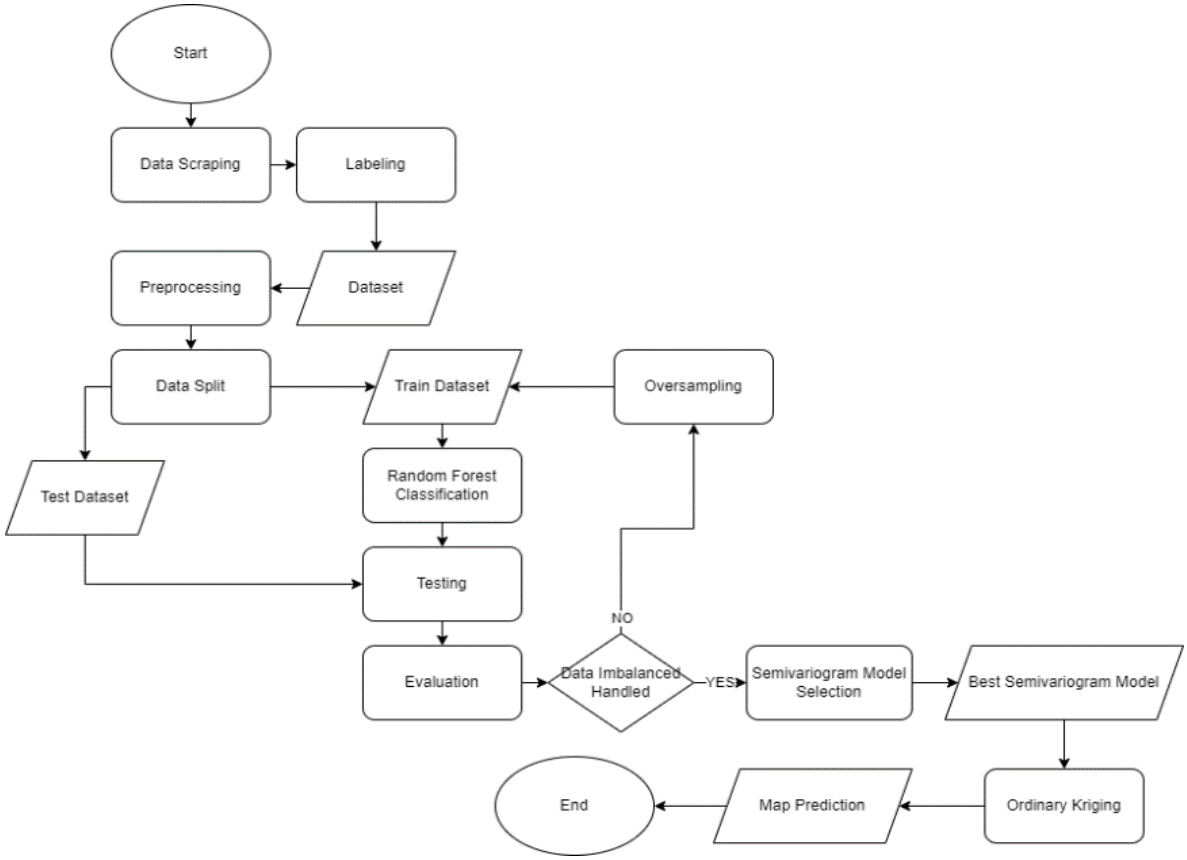

Figure 1. Experiment Flowchart

2.2 Data Scraping

The data taken is land sales data which includes land price, land area, and also the location of the land. Data by scraping through the land sales website. The amount of data obtained is around 500 land price data in the city of Jakarta. After that, the data was completed according to the attributes needed for this research through observations on Google Maps

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

2.3 Labeling

After scraping data from one of the land sales websites in Jakarta, the next stage is data labeling, the original land price is divided by the land area to get the land price per meter, then the land price is divided into 3 classes which can be seen in Table 1 below

Table 1. Land Price Class

Label	Description
0	the price of each meter of land <= IDR 15 million
1	IDR 15 Million < the price of each meter of land <= IDR 25 million
2	the price of each meter of land > IDR 25 million

2.4 Dataset

The dataset obtained is taken from rumah123.com and 99.co with information taken from land price, region, luasTanah, and Kota. For other attributes, the data is obtained from observation through google maps according to the location available from the data obtained from web scraping. Other attributes are cityCategory, address, landWidth, totallandprice, landPrice, publicTransportation, river, roadWidth, mainRoad, residentialArea, education, healthCenters, serviceTrade, tourismSpot, distancetotollroad.

Table 2. Dataset Description

Attribute	Description
cityCategory	Object, contains districts
address	Object, contains street address
landWidth	Float, contains land area
totallandprice	Integer, contains the land price before dividing by the land area
landPrice	Integer, contains land price per meter
publicTransportation.	Integer, contains availability of public transportation {0,1,2}
river	Integer, contains flood level {0,1,2}
roadWidth	Integer, contains road width in meters
mainRoad	Integer, contains the distance of the land to the main road
residentialArea	Integer, contains the settlement type {0,1,2,3}
education	Integer, contains the availability of education places{0,1,2,3,4,5}
healthCenters	Integer, contains the availability of health centers{0,1}
serviceTrade	Integer, contains the availability of health centers
tourismSpot	Integer, contains the availability of tourism spots
distancetotollroad	Integer, contains distance to toll road

2.5 Preprocessing

One of the most common data mining activities is data preprocessing, which involves getting the data ready and transforming it so that it can be used in the mining process. Data preprocessing techniques that will be performed are as follows:

a. Data Cleaning, Data cleaning is a phase in data preprocessing techniques that is used to remove unnecessary data and look for missing values since these types of data will affect machine learning processing and result in inaccurate unreliable output [12]. The data cleaning that must be done is to delete unnecessary data and check whether there are missing values in the dataset. The data cleaning that must be done is to delete data that is unnecessary in later processing such as the “landWidth” , "totallandprice" and "address" attributes and check whether there are missing values in the dataset.

b. Data Transformation, The next preprocessing technique performed is data transformation, by changing the form of data with integer type to float on the attribute "landPrice".

c. Label Encoding, The next step is to perform label encoding on categorical data into numeric. In this process, the data attribute that is label encoding is "cityCategory".

Table 3. Label Encoding Result

Label	cityCategory
0	West Jakarta
1	Central Jakarta
2	South Jakarta
3	East Jakarta
4	North Jakarta

d. Normalization, The normalization technique performed is StandardScaling. The Standard Scaler (SS) method for scaling features removes the mean and scales the variation to one to normalize each feature. Because the mean

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

and variance are the only two factors that determine the normalized result, it has the advantages of being linear, reversible, quick, and highly scalable. The Standard Scaler, on the other hand, exhibits a strong sensitivity to outliers and is better suited for normally distributed data [13] . Its normalized version can be determined from the following equation :

$$\hat{X} = \frac{X_i - \bar{X}}{\sigma} \quad (1)$$

Where $\hat{X}$ is Value Normalized, X_i is Sample, $\bar{X}$ = Mean, σ = Standard Deviation**.**

2.6 Data Split

The next step after preprocessing is data splitting where the data will be divided into training data and testing data. The ratio of the two data is 75% for train data and 25% for test data.

2.7 Random Forest Classification

Random Forest is one of data mining techniques that is used to address issues with classification and regression. Classification accuracy has increased dramatically as a result of voting to choose the class type and ensemble tree growth. These ensembles are grown by creating random vectors[5]. The use of randomization to promote variety has proven to be particularly effective in many methods that try to construct ensembles of diverse classifiers, such as bagging or random subspace methods. The Random Forest algorithm creates numerous decision trees because it uses an ensemble technique. An input vector is passed down each tree in the forest in order to classify a new object from it. Every tree casts a vote for the class that they believe best fits the situation. The categorization with the most votes (out of all the trees in the forest) is chosen by the forest[14]. Random Forests are appealing because they can be trained and predicted reasonably quickly, depend on just one or two tuning parameters, have generalization error estimates built in, and are simple to build in parallel. They can also be utilized directly for high-dimensional applications[15]. The Gini Index, which assesses an element's impurity in relation to the other classes, is frequently used by RFs to determine the best split selection. The Gini Index can be written as follows for a particular training dataset T[6]:

$$\sum \sum_{j \neq i} (f(C_i, T)/|T|)(f(C_j, T)/|T|) \quad (2)$$

$f(C_i, T)/|T|$ represents the probability that a selected case belongs to class C_i. Random forest has several advantages, including The non-parametric nature of RF; high classification accuracy; and capacity to determine variable importance[14].

2.8 Testing

In this testing stage, the author compares the test data containing labels from the dataset with the predicted data of the classification results that have been done in the random forest method before to get the accuracy of the model.

2.9 Evaluation

At this stage the evaluation metrics is used as a determinant of the score of the data classification using the Random Forest model, The table below are some of the variables of the metric:

Table 4. Confusion Matrix

	Actual Positive Class	Actual Negative Class
Predicted Positive Class	True Positive (*TP*)	False Negative (*FN*)
Predicted Negative Class	False Positive (*FP*)	True Negative (*TN*)

Some of the metrics used are :

a. Accuracy, It represents the percentage of observations that the classifier successfully predicted using the provided dataset [16].
It can be determined by the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

b. Precision, It represents the measure of the positive patterns in a positive class that are correctly predicted from the total anticipated patterns.
It can be determined by the following equation:

$$Precision\ (p) = \frac{TP}{TP + FP} \quad (4)$$

c. Recall[17]**,** It is the measure of positive patterns that are correctly categorized.
It can be determined by the following equation:

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

$$Recall\ (r)\ = \frac{TP}{TP + TN} \tag{5}$$

d. F1-score[18]**,** The harmonic mean of the recall and precision values is represented by this metric. It can be determined by the following equation:

$$F1 - Score\ = 2\ \frac{p * \ r}{p + r} \tag{6}$$

2.10 Oversampling

Because the weight of the data for each class in the author's dataset is not balanced, it is possible that the accuracy of the machine learning model performed is still not optimal, therefore by oversampling the accuracy of the machine learning model will increase. The oversampling method used is *Adaptive Synthetic Sampling Approach* (ADASYN) where this technique is based on adaptively producing minority data samples in accordance with its distributions. Minority class samples that are more difficult to learn are utilized to generate more synthetic data than minority class samples that are easier to learn, which in turn helps lessen the learning bias that was initially established owing to unequal data distribution[19]. And the second oversampling method that we used is *Synthetic Minority Oversampling Technique* (SMOTE). The study found that the SMOTE oversampling strategy generates bias towards minority classes since it uses SMOTE samples at the decision border. According to cost-sensitive learning, which imposes various costs on minority and majority classes, the researchers conducted this investigation. The best class separator is then determined using risk minimization[20].

2.11 Semivariogram Model Selection

We can distinguish between models with and without a sill when describing the semivariogram using the most prevalent fitting models. Power, linear, parabola, and logarithmic models are examples of models without a sill, whereas spherical, exponential, and gaussian models have a sill. The sill models among them show a steady-state roughness. The appropriate formulas are[21]:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left[1 - \exp\left(-\frac{h^2}{a^2}\right)\right] & h > 0 \end{cases} \text{ (Gaussian Model)} \tag{7}$$

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left(\frac{3h}{2a} - \frac{1h^3}{2a^3}\right) & 0 < h \leq a \\ C_0 + C & h > a \end{cases} \text{ (Spherical Model)} \tag{8}$$

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ C_0 + C\left[1 - \exp\left(-\frac{h}{a}\right)\right] & h > 0 \end{cases} \text{ (Exponential Model)} \tag{9}$$

Where C represent Partial Sill and C_0 represent Nugget are undetermined coefficients, where interval h is an independent variable , a represent range[21]. The model that will be selected from the three models is the model that has the smallest root mean squared error level.

2.12 Ordinary Kriging

One of the most popular kriging methods is Ordinary Kriging. Kriging is a weighted average of levels in the immediate vicinity and a local interpolation technique for such spatial interpolation issues. It is a form of optimal linear estimator[22]. By forecasting the value $Z^*(x_0)$, which is equal to the line sum of the known measured points, the spatial prediction of the unmeasured point x_0 is provided. Some researchers such as Isaaks, Srivastava, and Cressie and several other researchers provide a formula that can describe Ordinary Kriging simply with the following formula[23]:

$$Z^*(x_0)\ = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{10}$$

represents the measured value at position x_i, $Z^*(x_0)$ represents the predicted value at the unmeasured point x_0, λ_i represents the weighting coefficient from the measured position to x_0, and n represents the number of positions in the neighborhood search. To represent the spatial continuity of the data and display the spatial relationship between the pairs of points, a fitted model based on the input data distribution is required[23].

3. RESULT AND DISCUSSION

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

In this research, the authors split the data by dividing the data into 2 data, data train and data test with a ratio of 75% train data and 25% test data. The method used in the prediction model for classification is Random Forest Classifier. The classified data is identified with 3 classes, namely data with land prices per meter range of less than IDR 15 million, then data with range of IDR 15 to 25 million, and data with price range of more than IDR 25 million.

Table 5. Dataset Sample

cityCategory	Central Jakarta	South Jakarta
landPrice	16666667	38851351
publicTransportation	1	1
river	1	1
roadWidth	6	8
mainRoad	2300	50
residentialArea	1	3
education	1	3
healthCenters	1	1
serviceTrade	1	4
tourismSpot	0	0
distancetotollroad	1500	3500
label	2	3

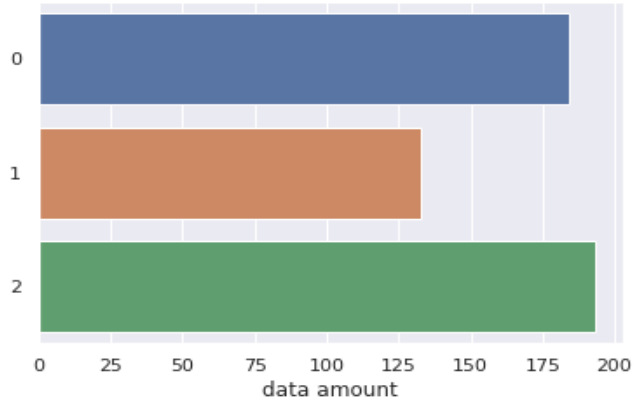

Figure 2. Dataset Amount for Each Class

in Figure 2, obtained imbalanced data in each class which will affect the accuracy of the classification, where the amount of data in class 0 is 184 data, class 1 is 133 data, and class 2 is 193 data. To overcome the imbalanced data, we use ADASYN and SMOTE oversampling methods on the train data.

3.1 Random Forest Classification Result

The Random Forest method is used to classify land price data per meter in each class, where the model will be assessed whether the predicted data class matches the respective price class. The features in the data that are modeled are, kotaCategory, transportasiPublik, sungai, lebarjalan, jalanUtama, kawasanPemukiman, pendidikan, kesehatan, perdaganganjasa, pariwisata, dan jarakkeJalantol. Classification is done with train data that is oversampled with the ADASYN and SMOTE methods.

Table 6. Evaluation Table

Oversampled	**Accuracy**	**Precision**	**Recall**	**F1-score**
Not Oversampled	79%	79%	79%	79%
With ADASYN	82%	84%	82%	82%
With SMOTE	81%	80%	80%	80%

Table 4 shows Evaluation Metrics of each dataset type, the difference in classification accuracy between the non-oversampled and oversampled train data. On data that is not oversampled, an accuracy of 79% is obtained, for data that has been oversampled with the ADASYN method, an accuracy of 82% is obtained, while data that is oversampled by SMOTE gets an accuracy of 81%. For precision score differences where the data train that is not oversampled gets a precision score of 79%, then the ADASYN oversampling gets a precision score of 84%, and for data that is oversampled with SMOTE gets a precision score of 80%. For recall score, train data recall score that is not oversampled is 79%, then 82% for the recall score of train data that is oversampled with the ADASYN method, and 80% for the recall score of data that is oversampled using SMOTE. For f1-score, train data that is not oversampled gets f1-score of 79%, then the train data that is oversampled using ADASYN gets an f1-score of 82% and for train

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

data that is oversampled using the SMOTE method gets an f1-score of 80%. We can see that oversampling method can help Random Forest to better recognize the class with the smallest amount by increasing the sample data in that class. In general, oversampling using ADASYN can improve the classification performance of Random Forest by 1% more than the data oversampled with SMOTE, and 3% more than the data that is not oversampled. In general, oversampling using ADASYN can improve the classification accuracy of Random Forest by 1% more than the data oversampled with SMOTE, and 3% more than the data that is not oversampled. For f1-score, it is found that ADASYN has a score that is 2% greater than the data oversampled with SMOTE and 3% greater than the data that is not oversampled. It can be seen that in this data classification case, oversampling using ADASYN is more effective in improving accuracy than SMOTE.

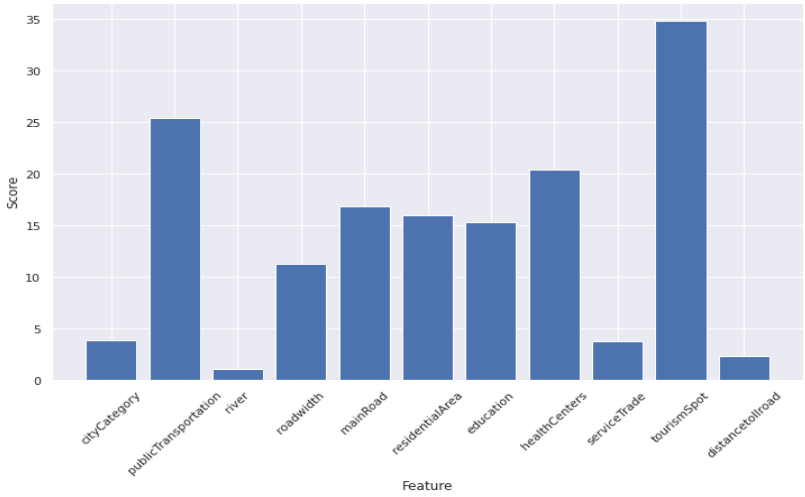

Figure 3. Feature Optimization Rate

In Figure 7**,** the optimization level of the features is obtained, where the most important features in the dataset are Pariwisata attributes with the first most optimal level, then transportasiPublik for the second most optimal level, and kesehatan for the third most optimal level. Of the 11 attributes, experiments were conducted with different attribute combinations, where in the first experiment using attributes (0,3,4,5,6,8,10) got an accuracy of 68%, then with a combination of attributes (0,1,2,3,4,5,6,8,10) got an accuracy of 67%, then using a combination of attributes (0,3,4,5,6,8,9) gets an accuracy of 77%, with a combination of attributes (0,1,4,5,8,9) getting an accuracy of 70%, and the last with a combination of attributes (0,1,2,3,4,5,6,7,8,9,10) getting the greatest accuracy with a value of 82%.

3.2 Ordinary Kriging Using Semivariogram

In Table 4 below, the prediction of the average land price per meter in each sub-district in Jakarta is obtained, then based on the location and price of land, spatial analysis is carried out with Ordinary Kriging using Semivariograms and the RMSE value of the data used is obtained.

Table 7. Land Price Prediction Result per Sub-district

No.	Sub-district	Price	No	Sub-district	Price
1.	Cempaka Putih	18750000	22.	Tambora	15000000
2.	Gambir	21556656	23.	Cilandak	17983974
3.	Johar Baru	26580933	24.	Jagakarsa	9000000
4.	Kemayoran	28002586	25.	Kebayoran Baru	56198338
5.	Menteng	62703459	26.	Kebayoran Lama	32694594
6.	Sawah Besar	22167120	27.	Mampang Prapatan	19910925
7.	Senen	28568670	28.	Pancoran	20000000
8.	Tanah Abang	39612469	29.	Pasar Minggu	17875000
9.	Cilincing	12201554	30.	Pesanggrahan	9625000
10.	Kelapa Gading	25509237	31.	Setiabudi	71335692
11.	Koja	8500000	32.	Tebet	35959384
12.	Pademangan	15000000	33.	Cakung	12074900
13.	Penjaringan	28119362	34.	Cipayung	8750000
14.	Tanjung Priok	18000000	35.	Ciracas	5933333
15.	Cengkareng	11003876	36.	Duren Sawit	13631194
16.	Grogol Petamburan	13000000	37.	Jatinegara	35946660
17.	Kali Deres	12475000	38.	Kramat Jati	21214025
18.	Kebon Jeruk	19000000	39.	Makasar	18000000

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

19.	Kembangan	16250833	40.	Matraman	14485125
20.	Palmerah	21050725	41.	Pasar Rebo	5600000
21.	Taman Sari	25000000	42.	Pulo Gadung	21291964

Table 8. RMSE Result for Semivariogram Model

Model	Root Mean Square Error
Exponential	1,131274e7
Gaussian	1,147808e7
Spherical	1,014896e7

The Root Mean Squared Error of the exponential, gaussian, and spherical semivariogram models between the simulated and measured values of the selected verification points is obtained and can be seen in Table 5. The Spherical model obtained an RMSE of 1.014896e7 where the RMSE value is smaller than the exponential with RMSE 1,131274e7 and gaussian model with RMSE 1,147808e7.

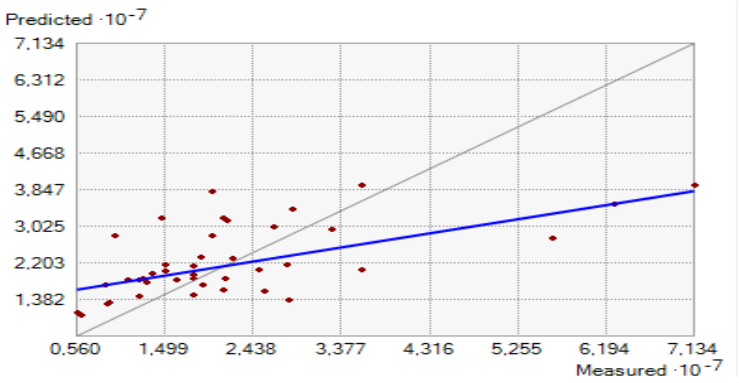

Figure 4. Prediction Result of Semovariogram Spherical Model

In Figure 8, it can be seen that the scenario line formed which has an angle of 45 degrees is different from the regression line, where in the model it can be seen that the blue line intersects with the grey line. This results in the predicted and measured values being quite different. The difference illustrates the RMSE value obtained by the Spherical model.

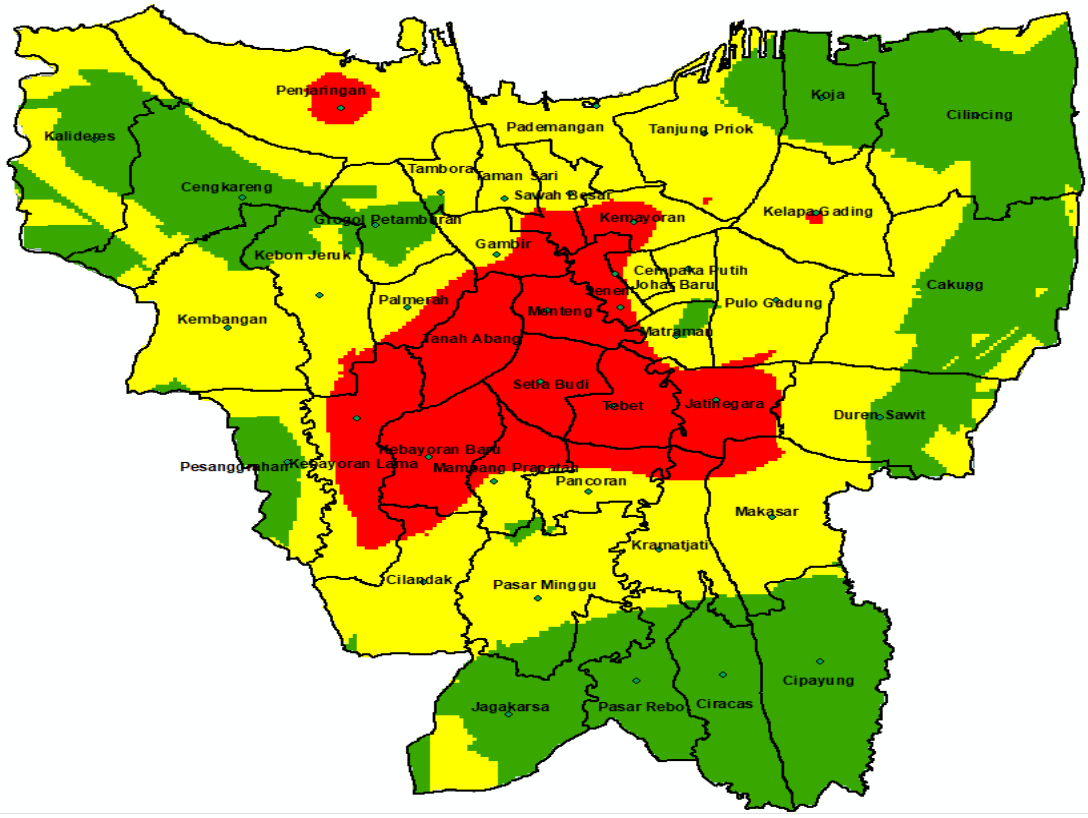

Figure 5. Land Price Prediction Map

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

In Figure 9, illustrates the map of land price prediction results with the Ordinary Kriging method using the spherical semivariogram model. The green part (33.93% of total area) illustrates areas that have land prices per meter with prices less than IDR 15 million, usually influenced by the fact that there are still few tourism sites in the area, and transportation is not as complete as the city center, where public transportation is only KRL Commuter Line and small inter-regional public transport along with a few elite residences. The yellow part (50.06% of total area) illustrates areas that have land prices per meter with prices above IDR 15 million but less than IDR 25 million, the part where city facilities, public transportation such as KRL Commuter Line, Transjakarta Bus, LRT in some areas, and health centers are sufficient, this area has more elite residences than the green part. The red part (16.01% of total area) is an area with a price of more than 25 million per meter of land, where there are city facilities, public transportation such as MRT, Transjakartabus, KRL Commuterline and a very complete health center so that land prices in the area are expensive, this part contains many elite residences and office centers.

The classification of land prices carried out with Random Forest and spatial interpolation using Ordinary Kriging with the Spherical Semivariogram model is good enough with a classification accuracy of 82% and an Interpolation RMSE of 1.014896e7. Compared to a similar study that discusses land price classification [4] , which has the highest Random Forest accuracy of 90.28%, the accuracy we get is indeed smaller, but the study did not predict prices using Ordinary Kriging and no map of the distribution of land prices was found. Compared to the study [24][25][26] , where the study predicted land prices using Ordinary Kriging but the land prices used were not the result of price classification from Random Forest but with existing price data.

4. CONCLUSION

In this research, land price per meter in Jakarta is classified into 3 classes using Random Forest. One experiment was conducted on the train data by oversampling because there was an imbalance of weights between the 3 existing classes. The most superior experiment was obtained on oversampled train data using the ADASYN method with an accuracy of 82%. Overall, the method outperformed the train data oversampled using SMOTE and the train data that was not oversampled. For spatial analysis, interpolation was performed using the Ordinary Kriging method using Semivariogram. The best model obtained is the Spherical model with the smallest RMSE value compared to the Exponential and Gaussian models. From the interpolation method, we generate a map based on the price of land per meter in each class. Where 33.93% of areas with green color are in the price range of less than IDR 15 million, then 50.06% of areas with yellow color are in the price range of IDR 15 to 25 million, and 16.01% of areas are in the price range of more than IDR 25 million.

REFERENCES

[1] M. Novita, "Kondisi Properti Masih Wait and See, Permintaan Apartemen Alami Recovery," *Rumah.com*, 2022. https://www.rumah.com/informasi-perusahaan/kondisi-properti-masih-wait-and-see-permintaan-apartemen-alami-recovery-64169 (accessed Jun. 09, 2022).

[2] I. Fata Robbany, A. Gharghi, and K.-P. Traub, "Land Use Change Detection and Urban Sprawl Monitoring in Metropolitan Area of Jakarta (Jabodetabek) from 2001 to 2015," *KnE Eng.*, vol. 2019, pp. 257–268, 2019, doi: 10.18502/keg.v4i3.5862.

[3] C. D. Putra, A. Ramadhani, and E. Fatimah, "Increasing Urban Heat Island area in Jakarta and it's relation to land use changes," in *IOP Conference Series: Earth and Environmental Science*, Apr. 2021, vol. 737, no. 1. doi: 10.1088/1755-1315/737/1/012002.

[4] C. Shousong, G. Xiaomin, W. Xiaoguang, and C. Ying, "Research on Urban Land Price Assessment Based on Artificial Neural Network Model," *IEEE Access*, vol. 7, pp. 180738–180748, 2019, doi: 10.1109/ACCESS.2019.2958978.

[5] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," *2018 15th Learn. Technol. Conf. L T 2018*, pp. 40–45, 2018, doi: 10.1109/LT.2018.8368509.

[6] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 67, no. 1, pp. 93–104, 2012, doi: 10.1016/j.isprsjprs.2011.11.002.

[7] N. M. Abdulkareem and A. M. Abdulazeez, "Machine learning classification based on Random Forest Algorithm: A review," *J. Sci. Bus.*, vol. 27, no. January, pp. 128–142, 2021, doi: 10.5281/zenodo.4471118.

[8] J. Magidi, L. Nhamo, S. Mpandeli, and T. Mabhaudhi, "Application of the random forest classifier to map irrigated areas using google earth engine," *Remote Sens.*, vol. 13, no. 5, pp. 1–15, 2021, doi: 10.3390/rs13050876.

[9] M. Molinari, M. R. Fida, M. K. Marina, and A. Pescape, "Spatial interpolation based cellular coverage prediction with crowdsourced measurements," *C2B(I)D 2015 - Proc. 2015 ACM SIGCOMM Work. Crowdsourcing Crowdsharing Big Data, Part SIGCOMM 2015*, pp. 33–38, 2015, doi: 10.1145/2787394.2787395.

[10] D. Ozturk and F. Kilic, "Geostatistical approach for spatial interpolation of meteorological data," *An. Acad. Bras. Cienc.*, vol. 88, no. 4, pp. 2121–2136, 2016, doi: 10.1590/0001-3765201620150103.

[11] Y. Ghiasi and V. Nafisi, "Strain estimation using ordinary Kriging interpolation," *Surv. Rev.*, vol. 48, no. 350, pp. 361–366, 2016, doi: 10.1080/00396265.2015.1116155.

[12] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[13] P. Ferreira, D. C. Le, and N. Zincir-Heywood, "Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection," *15th Int. Conf. Netw. Serv. Manag. CNSM 2019*, no. Cnsm, 2019, doi: 10.23919/CNSM46954.2019.9012708.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 674−683
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.1896

[14] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," *Proc. - 2012 Int. Conf. Data Sci. Eng. ICDSE 2012*, pp. 64–68, 2012, doi: 10.1109/ICDSE.2012.6282329.
[15] A. Cutler, D. R. Cutler, and J. R. Stevens, "Ensemble Machine Learning," *Ensemble Mach. Learn.*, no. February 2014, 2012, doi: 10.1007/978-1-4419-9326-7.
[16] A. K. Sandhu and R. S. Batth, "Software reuse analytics using integrated random forest and gradient boosting machine learning algorithm," in *Software - Practice and Experience*, Apr. 2021, vol. 51, no. 4, pp. 735–747. doi: 10.1002/spe.2921.
[17] M. Hossin and M. . Sulaiman, "a Review on Evaluation Metrics for Data," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, no. 2, pp. 1–11, 2015.
[18] D. Laffly, "PART 2 Basic Mathematical, Statistical and Computational Tools," 2020.
[19] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017-Janua, pp. 79–85, 2017, doi: 10.1109/ICACCI.2017.8125820.
[20] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.
[21] Z. Lianheng, Z. Shuaihao, H. Dongliang, Z. Shi, and L. Dejian, "Quantitative characterization of joint roughness based on semivariogram parameters," *Int. J. Rock Mech. Min. Sci.*, vol. 109, no. May, pp. 1–8, 2018, doi: 10.1016/j.ijrmms.2018.06.008.
[22] J. Ibrahim, M.-H. Chen, and D. Sinha, *Springer Series in Statistics*, vol. 27, no. 2. 2009. [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf
[23] T. G. Pham, M. Kappas, C. Van Huynh, and L. H. K. Nguyen, "Application of ordinary kriging and regression kriging method for soil properties mapping in hilly region of central Vietnam," *ISPRS Int. J. Geo-Information*, vol. 8, no. 3, 2019, doi: 10.3390/ijgi8030147.
[24] L. Jackson, J. Zuo, Z. Zhao, G. Zillante, and Y. Feng, *Critical Success Factors for Refurbishment Projects*. 2015. doi: 10.1007/978-3-662-46994-1_98.
[25] R. Cellmer and S. Zrobek, "The Cokriging Method in the Process of Developing Land Value Maps," *Proc. - 2017 Balt. Geod. Congr. (Geomatics), BGC Geomatics 2017*, pp. 364–368, 2017, doi: 10.1109/BGC.Geomatics.2017.14.
[26] H. Crosby, T. Damoulas, A. Caton, P. Davis, J. Porto de Albuquerque, and S. A. Jarvis, "Road distance and travel time for an improved house price Kriging predictor," *Geo-Spatial Inf. Sci.*, vol. 21, no. 3, pp. 185–194, 2018, doi: 10.1080/10095020.2018.1503775.