BAREKENG: Journal of Mathematics and Its Applications
September 2024 Volume 18 Issue 3 Page 1649–1664
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss3pp1649-1664

DETECTING URBAN SLUMS IN DKI JAKARTA: A KOTAKU DATA APPROACH WITH ENSEMBLE METHODS

Muhammad Muawwad MS[1], Rani Nooraeni[2*], Ananda Galuh Intan Prasetya[3]

[1,2]Department of Computational Statistics, [3]Department of Statistics, Politeknik Statistika STIS
Jln. Otto Iskandardinata No.64C, Jakarta Timur, DKI Jakarta, 13330, Indonesia
*Corresponding author's e-mail: * **raninoor@stis.ac.id***

Article History:

Received: 1st January 2024
Revised: 5th March 2024
Accepted: 28th June 2024
Published: 1st September 2024

Keywords:

Ensemble Method;
KOTAKU Data;
Random Forest Algorithm;
Slum Indicators;
Urban Slums.

ABSTRACT

Slums are one of the problems that often occur in urban areas, especially in developing countries. Slum settlements cause various social, economic, and environmental problems, including social injustice, infrastructure inefficiency, and a decrease in the population's quality of life. The PUPR Ministry representing the Indonesian government is trying to overcome slum settlements in Indonesia by creating the Cities Without Slums (KOTAKU) program. The KOTAKU program provides relevant and detailed data on slum settlements in Indonesia. Challenges arise when analyzing and utilizing KOTAKU data to identify slum indicators and map slums broadly. The method used in detecting slums using KOTAKU data is still conventional. Machine learning can be used to model data and classify or predict data by applying the Ensemble Method. This modeling will look for patterns or structures from the data that has been provided so that the detection results become more objective. This study aims to model slum indicators from KOTAKU data and detect urban slum settlements in DKI Jakarta. Modeling is done using the Random Forest algorithm. Data sourced from the KOTAKU program website established by the Ministry of PUPR RI. The results of the study show that the indicators that contribute most to the modeling of urban slum indicators in DKI Jakarta are the availability of safe access to drinking water and not fulfilling needs for drinking water. The slum indicator model without additions has good performance after going through the parameter tuning process with parameters ntree = 500 and mtry = 6. In contrast, the slum indicator model with additions has good performance if it does not go through a parameter tuning process or retains its initial parameters namely ntree = 500 and mtry = 4.

How to cite this article:

M. Muawwad MS, R. Nooraeni and A. G. I. Prasetya., "DETECTING URBAN SLUMS IN DKI JAKARTA: A KOTAKU DATA APPROACH WITH ENSEMBLE METHODS," *BAREKENG: J. Math. & App.*, vol. 18, iss. 3, pp. 1649-1664, September, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

Slums that are characterized by substandard housing conditions are often found, especially in urban areas [1]. Slums are defined as settlements that are close together and whose inhabitants are characterized by inadequate housing and basic services [2]. Currently, there are an estimated one billion people worldwide living in slums [3]. This number is projected to increase to three billion people by 2050 if current trends continue [4]. This increase will have an impact on several problems such as poverty, health, and social problems. Poverty problems can be interrelated with health, while social problems can be caused by the bad stigma of the world community for residents of slum settlements so that they experience deprivation, displacement, and even denial of access to basic services [5].

In Indonesia, matters related to slum settlements have been regulated in Law no. 1 Article 1 Paragraph 13 of 2011 concerning Housing and Residential Areas. The law explains that slum settlements are settlements that are uninhabitable because of the irregularity of the buildings, the high level of building density, and the quality of the buildings, facilities, and infrastructures that do not meet the requirements [6]. The emergence of slum settlements in Indonesia can be caused by various factors, both internal and external factors. Factors from within a settlement may include its characteristics, population, facilities, and infrastructure, while external factors may be caused by urbanization [7]. In general, the growth of slum settlements in Indonesia occurs due to the high rate of urbanization. Urban life is an attraction for rural communities to get jobs. The impact of this growth in particular will create a bad paradigm for government administration. Discipline degradation and social order-disorder will also occur as a result of the low economic conditions of the slum population [8].

Slum settlements in Indonesia are often found in big cities, one of which is Jakarta. Jakarta is the capital city of Indonesia, which has a population of around 11 million people and an area of 661.23 km^2. The population density in Jakarta is around 17,000 people per km^2 [9]. Slum settlements in Jakarta are mostly inhabited by people with low incomes. They often face various problems related to health, education, and work. These problems can include several things such as access to clean water resources, access to healthy and nutritious food, as well as access to adequate health and education services. Communities in slum settlements in Jakarta also often experience high economic and social pressures, such as unemployment, limited access to decent jobs, as well as violence and discrimination [10].

Based on data from the Indonesian Central Statistics Agency (BPS RI), six provinces in Indonesia have a percentage of slum households of more than 10% in the 2020-2022 range, namely the provinces of DKI Jakarta, West Java, Bangka Belitung Islands, Riau Islands, East Nusa Tenggara (NTT), and Papua provinces. Over the past three years, the percentage of slum households in DKI Jakarta has fluctuated from 2020 to 2022. In 2020, the percentage of slum households in DKI Jakarta was 22.07%, decreased to 14.69% in 2021, and increased again by 18.82% in 2022 [11].

Information about slum settlements and their conditions needs to be collected continuously so that the UN's Sustainable Development Goals can be achieved. This information can be used to determine whether the implemented policies, programs, and resources are effective in improving the conditions and quality of life of residents in slum settlements [3]. One of the information that can be obtained from slum settlements is their characteristics. According to UN-Habitat, the characteristics of slum settlements consist of five measurable characteristics that have been supported by substantial research globally, namely Durable Housing, Adequate Living Space, Access to Safe Water, Access to Adequate Sanitation, and Security of Tenure [12]. The availability of data on these characteristics needs to be ensured so that information and conditions of slum settlements in every country in the world can be known.

The Indonesian state has collected data related to slum settlements through the KOTAKU (Cities Without Slums) program. The KOTAKU program is one of the strategic efforts of the Directorate General of Cipta Karya, Ministry of Public Works and Public Housing (PUPR), to accelerate the handling of urban slums and support the "100-0-100 Movement", namely 100 percent access to proper drinking water, 0 percent of slums, and 100 percent access to proper sanitation [13]. Based on PUPR Ministerial Regulation No. 14 of 2018 concerning the Prevention and Improvement of the Quality of Slum Housing and Slums, 7 aspects become benchmarks for settlements that can be said to be slums, namely the Condition of Buildings, Conditions of Environmental Roads, Conditions of Drinking Water Provision, Conditions of Environmental Drainage, Conditions of Wastewater Management, Conditions Waste Management, and Availability of Public Open Space. From this aspect, there are 16 criteria or indicators for slum settlements in Indonesia, but these indicators do not include the characteristics of slum settlements as defined by UN-Habitat (2006) [12].

The PUPR Ministry has its slum characteristics, although other characteristics can identify slum settlements based on UN standard criteria. For example, the Building Permit Ownership (IMB) indicator contained in the KOTAKU data is not used to identify slum settlements, even though this indicator is included in the characteristics of slum settlements according to the UN.

The detection of slum settlements in Indonesian territory is based on indicators that have been defined by the government through the PUPR ministry. The detection process is carried out using the accumulation of scores obtained from each indicator. This score will determine whether the detected area is included in the category of light, medium, heavy slums, or not slums. Categorizing using scores has several drawbacks, one of which is that the assessment or measurement seems very subjective [14]. This can be corrected by utilizing modeling using Machine Learning. Machine learning is a branch of artificial intelligence that allows machines to learn from data without explicit programming [15]. Machine learning can be used to model data and classify or predict data by applying the Ensemble Method. The ensemble method is a technique in machine learning that combines several basic models to produce an optimal predictive model [16]. This modeling will look for patterns or structures from the data that has been provided so that the detection results become more objective.

Ensemble methods aim to reduce bias or variance that may exist in individual models, as well as increase generalization and robustness to overfitting. By combining various models that have different strengths and weaknesses, ensemble methods can produce more accurate and stable predictions [17]. In Yadav and Pal's (2020) research, they used the Random Forest Ensemble Method to predict heart disease. They collected a dataset of heart disease-related information and applied various tree-based classification algorithms, such as M5P, Random Tree, and Reduced Error Pruning to predict the presence of heart disease. The results of this research show that the Random Forest Ensemble Method provides the best results with an accuracy of 99%. They concluded that the Random Forest Ensemble Method outperformed other algorithms in previous studies [18]. Detection using the Ensemble Method was also carried out by Baradaran Rezaei et al. (2023). They used the Ensemble Method to predict the probability of Gastric Cancer occurrence and associated mortality. The Ensemble Method is also used to reduce prediction errors on a large number of patient features. The results of its application show that the designed Ensemble Method produces more precise predictions with an accuracy of 97.9% and 76.3% for predicting Gastric Cancer and related deaths, respectively [19].

Based on previous research, this research tries to apply the ensemble method in detecting slum settlements in DKI Jakarta. In its application, the slum indicators used by the PUPR ministry are the basis for forming the model. Improving the performance of the model was also carried out by adding several slum indicators other than those set by the PUPR ministry. This research aims to conduct a descriptive analysis of slum settlements in DKI Jakarta urban areas. In addition, this research also aims to model slum indicators from KOTAKU data, detect urban slum settlements in DKI Jakarta, and compare the KOTAKU slum indicator model without adding indicators and with adding indicators. This research is expected to provide several benefits, namely an in-depth understanding of the characteristics and indicators that contribute to slum settlements in urban areas, as well as more accurate detection of slums so that they can assist the government in taking appropriate actions to tackle slum settlements. Other benefits provided from this research can also be in the form of insight into the effectiveness of adding indicators in detecting slum settlements and an important source of information for policymakers, decision-makers, and other related parties to design appropriate policies and strategies in efforts to address slum settlement problems.

2. RESEARCH METHODS

In this study, the data was sourced from the KOTAKU program website which was formed by the Ministry of Public Works and Housing of the Republic of Indonesia. The KOTAKU website contains slum profile data for every sub-district in Indonesia. From this website, profile data on urban village slums in DKI Jakarta province are used as research data. This data consists of physical information and non-physical information for each village. The data was last updated in 2019. Urban areas in the province of DKI Jakarta were used as the location for this study consisting of the cities of West Jakarta, North Jakarta, East Jakarta, South Jakarta, and Central Jakarta. Village in urban DKI Jakarta is used as the unit of analysis in this study.

The research data consisted of 115 observations and 23 variables (including the target variable, namely the status of slum settlements). This study contains variables based on indicators of slum settlements in the KOTAKU program. The list of indicators used can be seen in **Table 1**.

Table 1. KOTAKU Slums Indicators and Its Variables

No.	Indicator	Variable (Unit)
1.	Building Irregularities	The building percentages have no regularity (%)
2.	Building Density	The percentage of the area has a density that does not comply with the provisions (%)
3.	Non-compliance with Building Technical Requirements	The percentage of buildings that do not meet the technical requirements (%)
4.	Environmental Road Service Coverage	Percentage of non-existing road length (%)
5.	Environmental Road Surface Quality	Percentage of road length with damaged surface (%)
6.	Availability of Safe Access to Drinking Water	Percentage of households without access to safe drinking water (%)
7.	Unfulfilled Needs of Drinking Water	Percentage of households not meeting their minimum drinking water needs (%)
8.	Inability to Drain Runoff	Percentage of area affected by inundation (%)
9.	Drainage Unavailability	Percentage of non-existing drainage channel length (%)
10.	Disconnected with the City Drainage System	Percentage of channel length not accessible to the city system (%)
11.	Not Maintained Drainage	Percentage of drainage channel length not maintained (%)
12.	Drainage Construction Quality	Percentage of broken drainage channel length (%)
13.	Wastewater Management System Does Not Meet Technical Standards	Percentage of households without access to the wastewater system according to technical standards (%)
14.	Infrastructure and Facilities for Wastewater Management Does Not Meet the Technical Requirements	Percentage of households with wastewater infrastructure facilities that do not comply with technical requirements (%)
15.	Garbage Infrastructure and Facilities Not By Technical Requirements	Percentage of households with waste processing facilities that do not comply with the technical requirements (%)
16.	Waste Management System that Does Not Meet Technical Standards	Percentage of households with waste processing systems that do not comply with technical standards (%)
17.	Waste Management Facilities and Infrastructure are Not Maintained	Percentage of households with unmaintained waste processing facilities (%)
18.	Unavailability of Fire Protection Infrastructure	Percentage of buildings not served by fire protection infrastructure (%)
19.	Unavailability of Fire Protection Facilities	Percentage of buildings not served by fire protection facilities (%)

The variables in **Table 1** are defined as X1 to X19 respectively. The additional indicators used are also sourced from the KOTAKU database. A list of additional indicators can be seen in the following table.

Table 2. Additional Slums Indicators and Its Variables

No.	Indicator	Variable (Unit)
1.	Population Density	Village population density (Person/Ha)
2.	Building Permit Ownership	Percentage of residential buildings having IMB (%)
3.	Ownership of Freehold Title (SHM)/Building Rights (HGB)/Government Recognized Letters from Building Land	The percentage of residential building land has SHM/HGB/Letters recognized by the government (%)

The variables in **Table 2** are defined as X20 to X22 respectively. Variable X20 measures the second characteristic of slums, namely Adequate Living Space, while variables X21 and X22 measure the fifth characteristic of slums, namely Security of Tenure. These indicators are mapped to see the completeness of the indicators in meeting the characteristics of slum settlements set by UN-Habitat (2006) **[12]**. The mapping can be seen in the following figure.

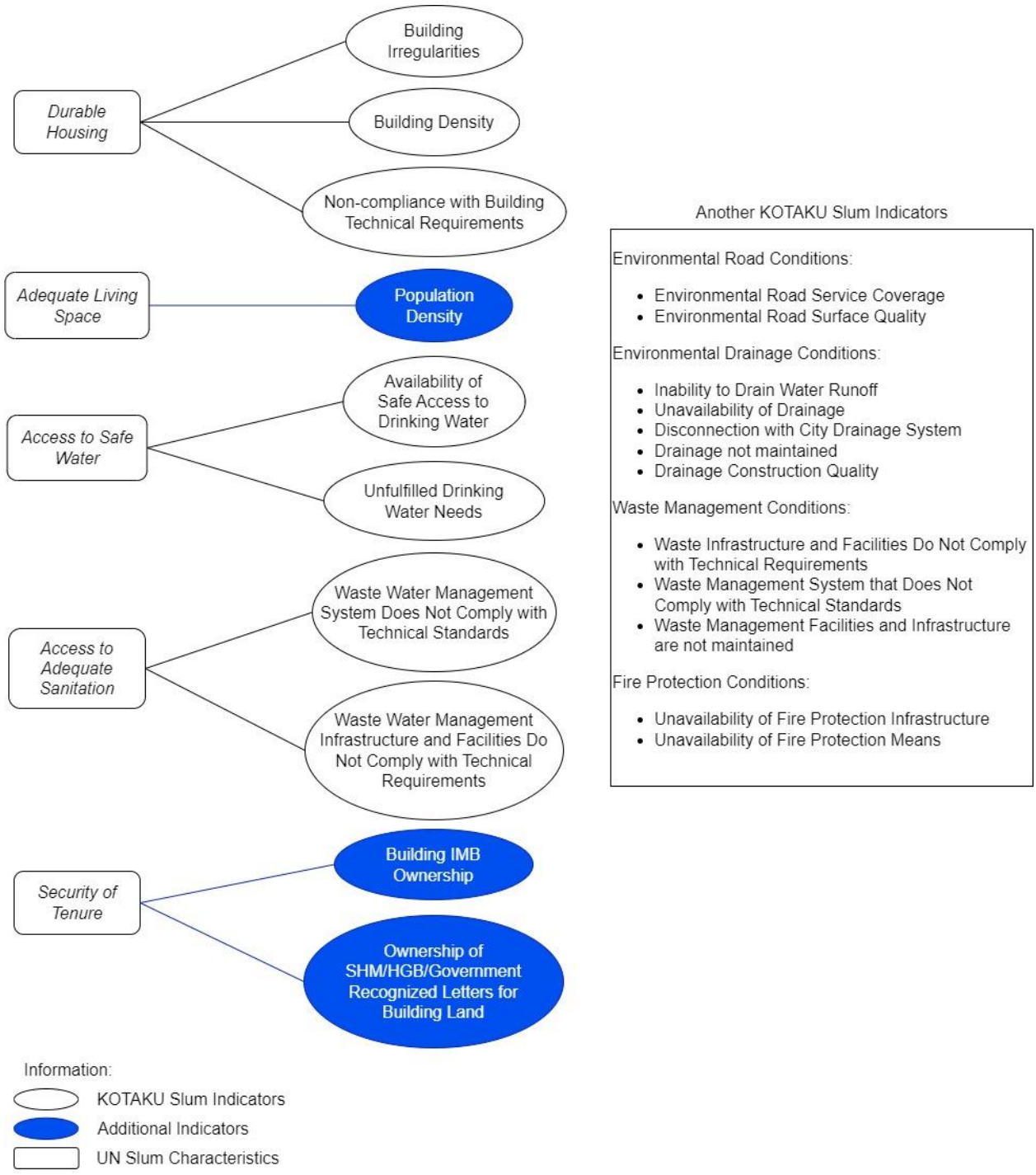

Figure 1. Mapping of KOTAKU Slum Indicators against UN Slum Characteristics

2.1 Data Collection

The data collection process begins by visiting the kotaku.pu.go.id site. This site has a Management Information System (SIM) which contains various information and data related to the KOTAKU program, one of which is the slum profile of 19 parameters/indicators for each village in Indonesia. From this slum profile, early-late slum data from each village can be downloaded as a file in Excel format (.xlsx format). Initial and final slum data are data for 2016 and 2019 respectively, but the data used in this study is only slum data at the end of 2019.

There are 115 data (records) taken from the KOTAKU website. Data is classified into four groups, namely light slums, medium slums, heavy slums, and not slums. This classification follows the categorization carried out by the PUPR ministry in detecting slum settlements in Indonesia [13].

2.2 Data Pre-processing

The data that has been collected is then processed by carrying out several stages. The data pre-processing stages in this study consist of Data Integration and Data Transformation. Data Integration is the process of combining data contained in different data sources, both from internal and external sources, into one integrated data set [20]. Data transformation is the process of changing or transforming data from one form to another that is more suitable for a particular analysis or use [21].

First, data from each urban village in DKI Jakarta are integrated into one data. The data for each village is combined into a new Excel file to simplify data processing. The data that has been integrated is then transformed/converted into a format that is more suitable for analysis. In this case, data from each variable is converted into percentage form, except for the variable X20 (village population density). This process does not change (add/subtract) data that has been previously collected.

2.3 Slum Indicator Modeling

Modeling is done using one of the algorithms in the Ensemble Method, namely Random Forest. Random Forest is an algorithm that is used to classify large amounts of data. Random Forest classification is done by merging trees by conducting training on the data samples they have [22]. Random Forest can be used on several types of data, such as discrete, continuous data, multivariate combinations, and survival data. Random Forest can detect interactions between dependent and independent variables and can explore data with its flexibility [23]. In carrying out analysis using Random Forest, there are no certain assumptions that must be met [24]. The main advantages of Random Forest are its ability to overcome overfitting, have tolerance for imbalanced data, and have good performance on complex data. In addition, Random Forest can provide estimates of the importance of each feature in predictions [25]. The Random Forest algorithm is divided into two parts: first, forming 'k' trees to create a random forest, and second, making predictions with the random forest that has been formed. The steps in implementing Random Forest include [23]:

1. Create sample data by random sampling with a return from the dataset.
2. Use the sample data to build the i-th tree (i = 1, 2, 3, …, k).
3. Repeat steps 1 and 2 until k times.

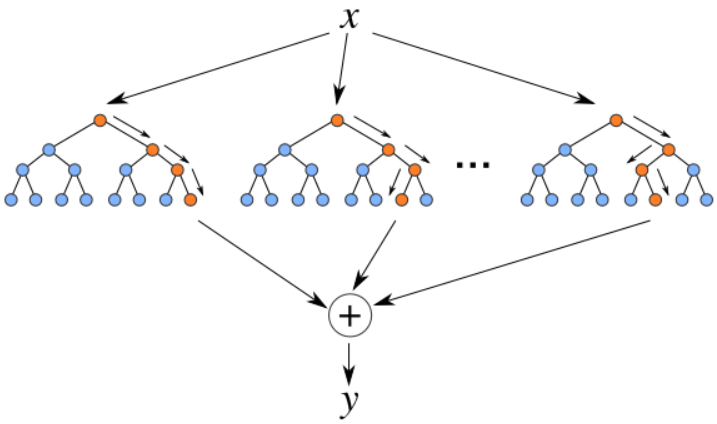

Figure 2. Random forest

In this study, the data was divided into two parts, namely 80% training data and 20% testing data. The training data is used to create the model, while the testing data is used to evaluate the model. Model creation and evaluation using R Package (randomForest) with RStudio software.

2.4 Model Evaluation

Modeling was carried out twice, namely modeling the slum indicators without additions and with additions. Each model will produce variable importance values from each input variable and go through a parameter tuning process to adjust the parameter values in the model and obtain optimal results. This process uses the Random Search CV (Cross-Validation) method. Random Search CV is a method for tuning

parameters in machine learning models by searching for random combinations of hyperparameter values from a predetermined range of values. This method makes it possible to find the hyperparameter combination that gives the best performance for the model without having to try all the combinations sequentially. Random Search CV is more efficient than Grid Search or other methods because the method does not try all combinations of values sequentially, but chooses them randomly. According to Bergstra and Bengio, **Figure 3** illustrates the training flow of Random Search CV **[26]**.

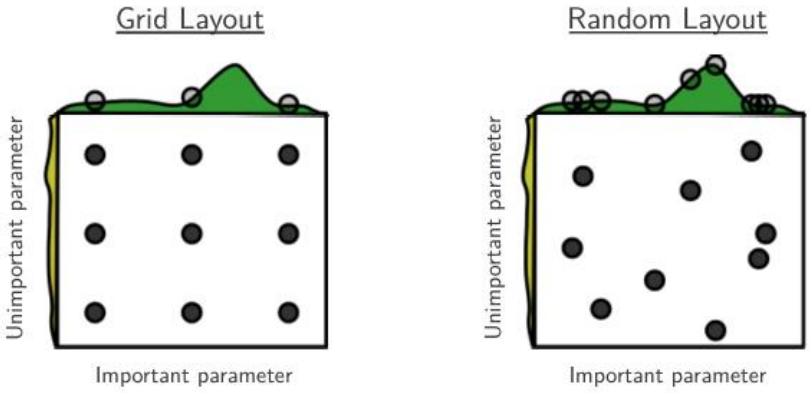

Figure 3. Illustration of the Application of Hyperparameter Optimization

Grid and random search of nine trials for optimizing a function $f(x,y) = g(x) + h(y) \approx g(x)$ with low effective dimensionality. Above each square $g(x)$ is shown in green, and on the left of each square $h(y)$ is shown in yellow. With grid search, nine trials only test $g(x)$ in three distinct places. With random search, all nine trials explore distinct values of g. This failure of grid search is the rule rather than the exception in high dimensional hyper-parameter optimization **[26]**.

The models that have been made are then evaluated using the Confusion Matrix. The confusion matrix is a tabulation of calculations based on evaluating the performance of a classification model based on the number of research objects that are correctly and incorrectly predicted **[3]**. Several evaluation metrics such as accuracy, precision, recall, specificity, and F1-score are calculated to evaluate the performance of the model as a whole or by class. Accuracy, precision, recall, specificity, and F1-score are calculated using the formula **[27]**.

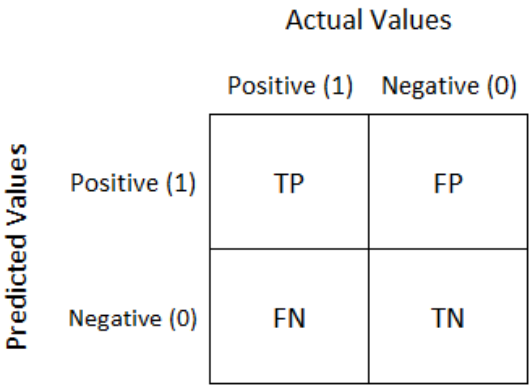

Figure 4. Confusion Matrix

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

$$Precision = \frac{(TP)}{(TP + FP)} \quad (2)$$

$$Recall = \frac{(TP)}{(TP + FN)} \quad (3)$$

$$Specificity = \frac{(TN)}{(TN + FP)} \quad (4)$$

$$F1\ Score = \frac{2 \times (Recall\ \times Precision)}{(Recall\ +\ Precision)} \quad (5)$$

where:

TP: True Positive, when the predicted results are positive and it's true.

TN: True Negative, when the predicted results are negative and it's true.

FP: False Positive, when the predicted results are positive and it's false.

FN: False Negative, when the predicted results are negative and it's false.

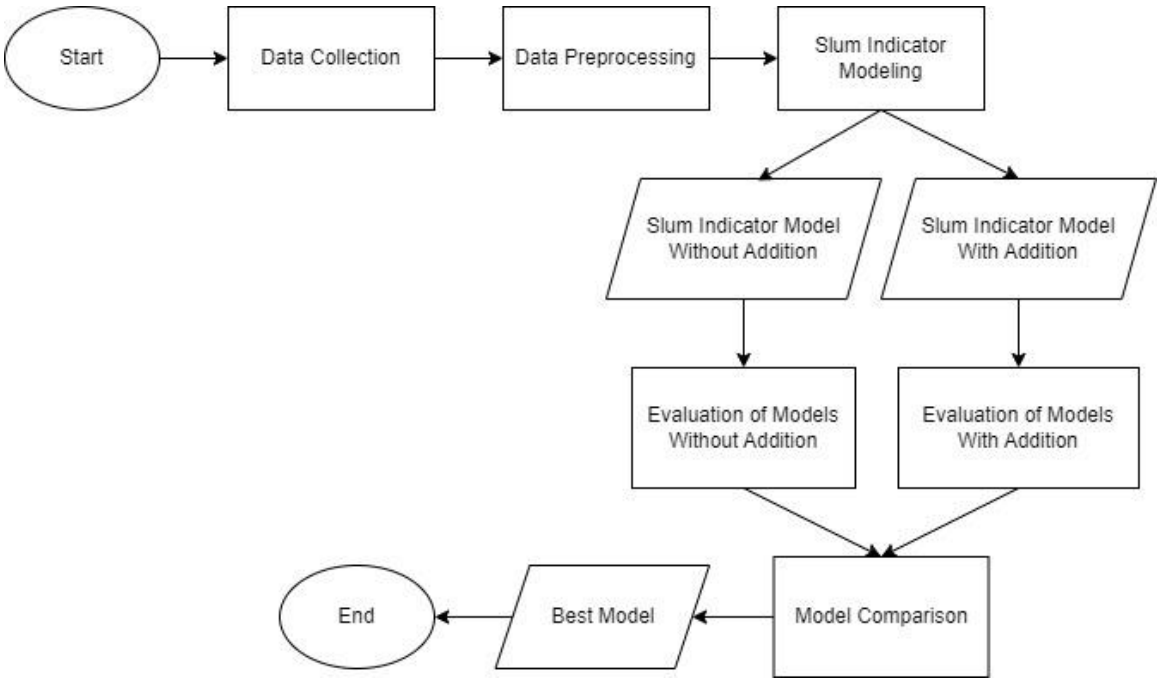

Figure 5. Research flow

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis of DKI Jakarta Slums

There are 267 urban villages in DKI Jakarta. Based on KOTAKU data, there are 118 urban villages in DKI Jakarta have been recorded. Of the 118 urban villages recorded, 3 of them are urban villages in Seribu Islands Regency. In this research, the Seribu Islands Regency was not included in the research locus.

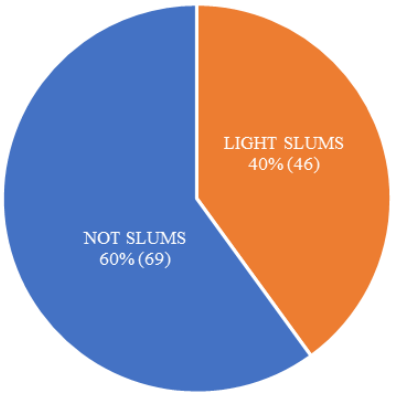

Figure 6. Percentage of the Number of Light Slums and not Slums in DKI Jakarta Urban Areas in 2019

The 115 urban villages in DKI Jakarta consist of 30 urban villages in the city of West Jakarta, 18 urban villages in the city of Central Jakarta, 20 urban villages in the city of South Jakarta, 24 urban villages in the city of East Jakarta, and 23 urban villages in the city of North Jakarta. Based on **Figure 6**, in 2019, the number of urban villages identified as light slums was 46 (40% of 115 urban villages), while the rest were not slums.

Table 3. Five-Number Summary and Mean of Input Variables

Variable	Minimum	1st Quartile	Median	3rd Quartile	Maximum	Mean
X1	0	0	0	0.07175	0.914	0.07849
X2	0	0	0	0.1520	1	0.1172
X3	0	0.03565	0.0812	0.1752	1	0.1345
X4	0	0	0	0	0.0676	0.0005974
X5	0	0	0.0157	0.1828	0.8333	0.1137
X6	0	0.0563	0.2083	0.4057	1	0.2985
X7	0	0.05225	0.1667	0.3889	0.99	0.25189
X8	0	0.3524	0.8407	1	1	**0.6659**
X9	0	0.0008	0.1139	0.2233	0.6	0.1386
X10	0	0	0	0	0.0382	0.0003322
X11	0	0.7052	0.8573	0.9629	1	**0.7792**
X12	0	0.0635	0.1806	0.3957	0.9966	0.2411
X13	0	0.0018	0.0759	0.243	0.99	0.1587
X14	0	0.00085	0.0706	0.23935	0.99	0.15199
X15	0	0	0	0.4683	0.9735	0.2229
X16	0	0	0	0	0.5261	0.02411
X17	0	0.9976	1	1	1	**0.8337**
X18	**0**	**0**	**0**	**0**	**0**	**0**
X19	**0**	**0**	**0**	**0**	**0**	**0**
X20	6.95	156.77	275.71	650.39	9087.67	625.32
X21	0	0.0298	0.2059	0.3918	1	0.2428
X22	0	0.2227	0.4205	0.5351	1	0.4241

Based on **Table 3**, several input variables have a fairly high mean value. These variables are X8 (percentage of area affected by inundation), X11 (percentage of drainage channel length not maintained), and X17 (percentage of households with waste processing facilities not maintained) with successive averages of 66.59%, 77 .92%, and 83.37%. This means that urban villages in DKI Jakarta in 2019 have three main obstacles in overcoming slum settlements, including problems related to inundation, drainage channels that are not maintained, and waste management infrastructure that is not maintained.

Variables X18 (percentage of buildings not served by fire protection facilities) and X19 (percentage of buildings not served by fire protection facilities) have a zero value for the five-number summary or their average. This indicates that urban villages in DKI Jakarta in 2019 had no problems related to fire protection conditions. In other words, fire protection facilities and infrastructure in each sub-district have been well served.

3.2 Slum Indicator Modeling (Without Addition)

The modeling begins by using 19 variables which are indicators of the formation of slum settlements according to the KOTAKU PUPR program. The Random Forest model is then created using training data. This model uses the default parameter with the number of trees to grow (*ntree* = 500) and the number of variables randomly sampled as candidates at each split (*mtry* = 4) for classification. The results of this modeling can be seen in the following table and figure.

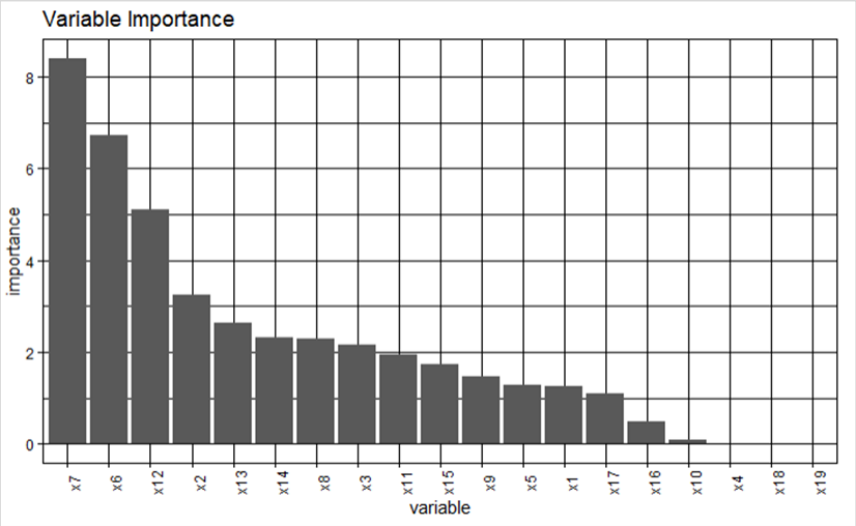

Figure 7. Graph of Variable Importance Value of 19 Variables (default parameters)

Variable importance in **Figure 7** is a term used in machine learning to measure the extent to which each feature (variable) contributes to model performance or the predictions made by the model. This can help in understanding what features are the most important or have a big impact on the prediction process, as well as which features are less relevant or have a small impact **[28]**. If we look at the value of the variable importance in **Figure 7**, variable X7 has the highest contribution in constructing the model, followed by variables X6, X12, and so on. This means that variable X7 (percentage of households whose minimum drinking water needs are not met) is the variable that contributes the most to the status of urban slum settlements in DKI Jakarta province.

Table 4. Confusion Matrix Data Training 19 Variables (Default Parameter)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	21 (22.83%)	13 (14.13%)	0.38235294
Not Slums	4 (4.35%)	54 (58.69%)	0.06896552
Accuracy		81.52%	
Precision		61.76%	
Recall		84.00%	
Specificity		80.59%	
F1-score		71.18%	

Based on **Table 4**, the model that has been made produces an accuracy of 81.52%. After modeling, parameter tuning is performed to find the optimal *mtry* parameter value for the model. The method used in this process is Random Search CV. The results are presented in the following table.

Table 5. Resampling Results from Model Tuning Parameters Without Additional Indicators

Mtry	Accuracy	Kappa
2	82.56%	59.74%
4	84.32%	64.13%
6	**85.39%**	**66.96%**
7	84.32%	64.45%
12	84.32%	66.04%
14	83.57%	65.29%
16	81.80%	61.21%
19	81.80%	61.21%

From **Table 5**, it is found that the *mtry* parameter value that can make the model more optimal is *mtry* = 6. This value is determined based on the results of the accuracy of each *mtry* parameter candidate with *mtry* = 6 having the highest accuracy of 85.39%. After the optimal *mtry* is determined, modeling is again carried out using the new parameter values. The results of this modeling can be seen in the following table and figure.

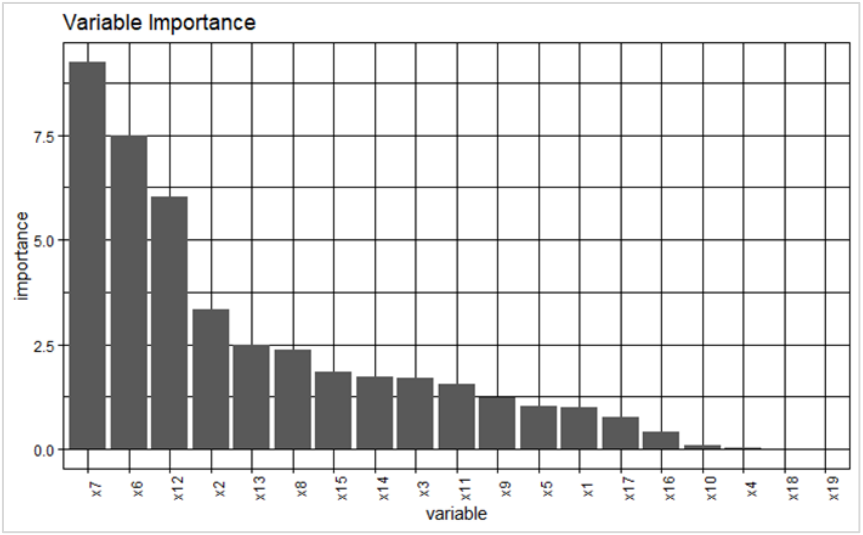

Figure 8. Graph of Variable Importance Value of 19 Variables (parameter *mtry* = 6)

If we look at the value of the variable importance in **Figure 8**, variable X7 (percentage of households whose minimum drinking water needs are not met) still has the highest contribution in constructing the model, followed by variables X6, X12, and so on. This means that changes in the parameters in the model do not have a significant effect on the contribution of each model-forming variable.

Table 6. Confusion Matrix Data Training 19 Variables (Parameter *mtry* = 6)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	24 (26.09%)	10 (10.87%)	0.2941176
Not Slums	5 (5.43%)	53 (57.61%)	0.0862069
Accuracy	83.69%		
Precision	70.59%		
Recall	82.76%		
Specificity	84.13%		
F1-score	76.19%		

Based on **Table 6**, the model that has been created using the parameter *mtry* = 6 produces an accuracy of 83.69%, better than the model using the default parameters. Next, the model with *mtry* = 6 parameters is evaluated using data testing. The evaluation results are presented in the following table.

Table 7. Confusion Matrix Data Testing 19 Variables (Parameter *mtry* = 6)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	8 (34.78%)	1 (4.35%)	0.11111111
Not Slums	4 (17.39%)	10 (43.48%)	0.28571429
Accuracy	78.26%		
Precision	88.89%		
Recall	66.67%		
Specificity	90.91%		
F1-score	76.19%		

In **Table 7**, it can be seen that the model produces 18 correct predictions and 5 wrong predictions. This means that 18 urban villages in DKI Jakarta have been correctly detected by the model with details of 8 urban villages which are slum predicted by the model and 10 urban villages which are not slum predicted by the model. Model accuracy reaches 78.26% in predicting testing data. The precision, recall, and specificity values

generated by the model are 88.89%, 66.67%, and 90.91%, respectively, so the F1-score value of the model is 76.19%.

3.3 Slum Indicator Modeling (With Addition)

The next model uses 22 slum-forming variables. These variables consist of 19 variables forming slums according to the KOTAKU PUPR version and 3 additional variables based on the characteristics of slum settlements from UN-Habitat (2006). This modeling is still started by using the default parameters for *ntree* = 500 and *mtry* = 4. The modeling results can be seen in the following table and figure.

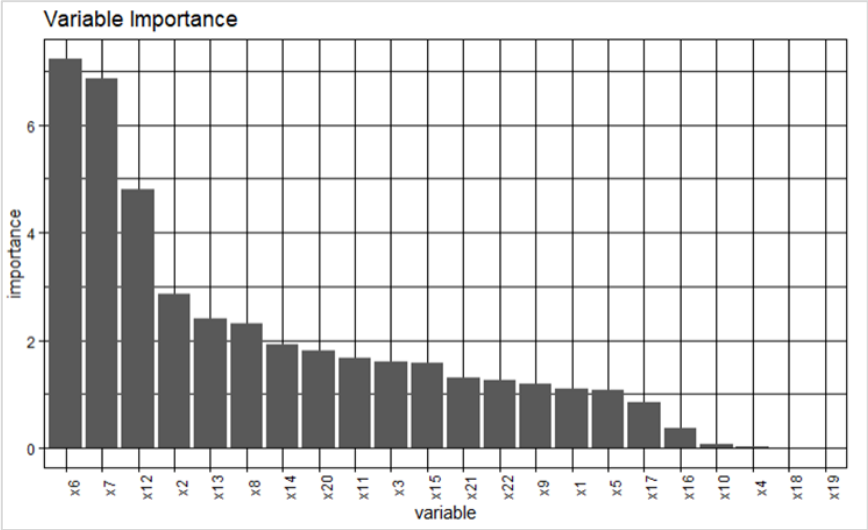

Figure 9. Graph of Variable Importance Value of 22 Variables (default parameter)

If we look at the value of the variable importance in **Figure 9**, variable X6 has the highest contribution in constructing the model, followed by variables X7, X12, and so on. This means that variable X6 (percentage of households without access to safe drinking water) is the variable that contributes the most to the status of urban slum settlements in DKI Jakarta province. Additional variables in this model also provide a high enough contribution to the preparation of the model.

Table 8. Confusion Matrix Data Training 22 Variables (Default Parameter)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	22 (23.91%)	12 (13.04%)	0.35294118
Not Slums	4 (4.35%)	54 (58.69%)	0.06896552
Accuracy	82.61%		
Precision	64.71%		
Recall	84.62%		
Specificity	81.82%		
F1-score	73.34%		

Based on **Table 8**, the model that has been made produces an accuracy of 82.61%. Parameter tuning was carried out again for the model with the addition of the slum indicator. The results are presented in the following table.

Table 9. Resampling Results from Tuning Model Parameters with Additional Indicators

mtry	Accuracy	Kappa
2	81.36%	56.17%
4	82.82%	60.22%
6	83.18%	61.36%
12	**83.78%**	**64.05%**
14	81.96%	60.64%
16	80.93%	58.31%
19	79.51%	55.50%
22	79.14%	55.09%

From **Table 9**, it is found that the *mtry* parameter value that can make the model with more optimal additions is *mtry* = 12. This value is determined based on the results of the accuracy of each *mtry* parameter candidate with *mtry* = 12 having the highest accuracy of 83.78%. After the optimal *mtry* is determined, modeling is again carried out using the new parameter values. The results of this modeling can be seen in the following table and figure.

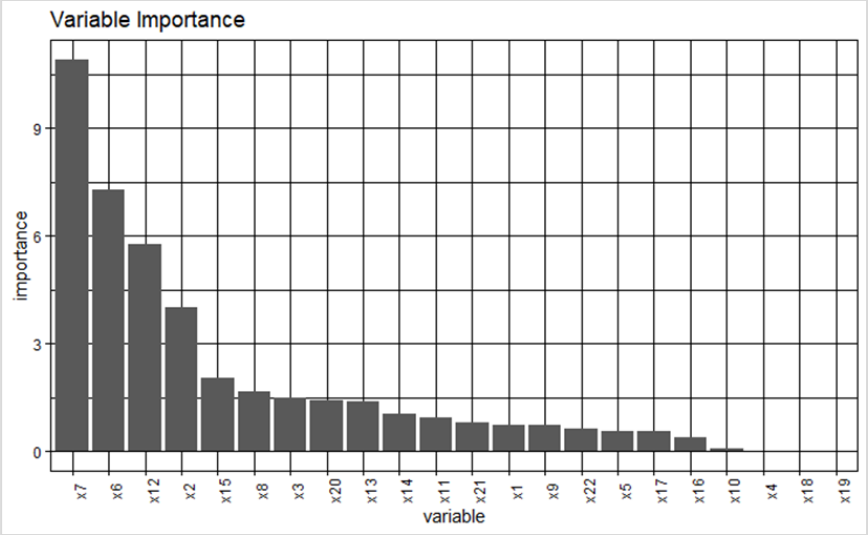

Figure 10. Graph of Variable Importance Value of 22 Variables (parameter *mtry* = 12)

If we look at the value of variable importance in **Figure 10**, there is a change in the order of contribution between variables X7 and X6. This means that changes in the parameters in the model have a significant influence on the contribution of each model-forming variable.

Table 10. Confusion Matrix Data Training 22 Variables (Parameter *mtry* = 12)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	23 (25%)	11 (11.96%)	0.3235294
Not Slums	7 (7.61%)	51 (55.43%)	0.1206897
Accuracy		80.43%	
Precision		67.65%	
Recall		76.67%	
Specificity		82.26%	
F1-score		71.88%	

Based on **Table 10**, the model that has been created using the parameter *mtry* = 12 produces an accuracy of 80.43%, worse than the model using the default parameter. Furthermore, models with default parameters are still evaluated using testing data. The evaluation results are presented in the following table.

Table 11. Confusion Matrix Data Testing 22 Variables (Default Parameter)

Predicted Values	Actual Values		Class Error
	Light Slums	Not Slums	
Light Slums	8 (34.78%)	1 (4.35%)	0.11111111
Not Slums	4 (17.39%)	10 (43.48%)	0.28571429
Accuracy		78.26%	
Precision		88.89%	
Recall		66.67%	
Specificity		90.91%	
F1-score		76.19%	

In **Table 11**, it can be seen that the model produces 18 correct predictions and 5 wrong predictions. This means that 18 urban villages in DKI Jakarta have been correctly detected by the model with details of 8 urban villages which are slum predicted by the model and 10 urban villages which are not slum predicted by the model. Model accuracy reaches 78.26% in predicting testing data. The precision, recall, and specificity values generated by the model are 88.89%, 66.67%, and 90.91%, respectively, so the F1-score value of the

model is 76.19%. These results are the same as modeling the slum indicator without adding the parameter *mtry* = 6.

3.4 Comparison of Slum Indicator Models

Modeling the slum indicators without additions produces an optimal model if the model is changed from *mtry* = 4 to *mtry* = 6 while modeling the slum indicators with additions produces an optimal model if the model uses the default parameters (*mtry* = 4). When the two models are compared, the results obtained are as follows.

Table 12. Comparison of Model Evaluation Results

Model	Accuracy	Precision	Recall	Specificity	F1-score
Without Additions (*mtry* = 6)	78.26%	88.89%	66.67%	90.91%	76.19%
With Additions (*mtry* = 4)	78.26%	88.89%	66.67%	90.91%	76.19%

Based on Table 12, it is found that the slum indicator model without additions (*mtry* = 6) has the same evaluation results as the slum indicator model with additions (default parameter or *mtry* = 4). This means that the slum indicator model can be optimized in two ways, the first is to do parameter tuning if the indicators in the model are not added and the second is to add indicators if the model does not go through the parameter tuning process (still using the default parameters). Optimizing the model in a second way is supported by research conducted by Mahabir et al in 2020. In that study, Mahabir et al. suggest that the addition of slum indicators can improve the performance of the detection model.

4. CONCLUSIONS

Based on the results and discussion in the previous section, the following can be concluded that villages in urban DKI Jakarta in 2019 had three main obstacles in overcoming slum settlements, including problems related to inundation, drainage channels that were not maintained, and waste management infrastructure that was not maintained. Even so, settlements in DKI Jakarta do not have problems related to fire protection conditions. In other words, fire protection facilities and infrastructure in each village have been well served.

The indicators that contribute most to the modeling of urban slum indicators in DKI Jakarta are the Availability of Safe Access to Drinking Water and Not Fulfilling Needs for Drinking Water. These indicators were obtained from the slum indicator model without additional indicators (parameter *ntree* = 500 and *mtry* = 6), as well as the slum indicator model with additional indicators (default parameter or *ntree* = 500, *mtry* = 4). This shows that slum settlements in DKI Jakarta can be caused by the unavailability of adequate access to water (the failure to fulfill one of the characteristics of slum settlements, namely Access to Safe Water).

The slum indicator model without additions has good performance after going through the parameter tuning process and produces a model with parameters *ntree* = 500 and *mtry* = 6. On the other hand, the slum indicator model with additions has good performance if it does not go through the parameter tuning process or retains its initial parameters (default parameters) namely *ntree* = 500 and *mtry* = 4. The two models have the same evaluation results. Their F1-score value is 76.19%, this means that these models have good precision and recall and can perform maximum detection.

Based on the conclusions of this study, the researchers suggest to the government, especially the Ministry of Public Works and Public Housing to be able to apply various techniques or methods from machine learning to improve the ability to detect slums in all regions of Indonesia. Researchers also provide input for the development of a more accurate and reliable slum settlement detection model. Through modeling slum indicators using the Ensemble Method, it is necessary to select the most appropriate and optimal type of method to increase the accuracy of the detection of slums.

REFERENCES

[1] E. Ranguelova, B. Weel, D. Roy, M. Kuffer, K. Pfeffer, and M. Lees, "Image based classification of slums, built-up and non-built-up areas in Kalyan and Bangalore, India," *Eur J Remote Sens*, vol. 52, no. sup1, pp. 40–61, Mar. 2019, doi: 10.1080/22797254.2018.1535838.

[2] United Nations, "Slum profile in human settlements," 2009.

[3] R. Mahabir, P. Agouris, A. Stefanidis, A. Croitoru, and A. T. Crooks, "Detecting and mapping slums using open data: a case study in Kenya," *Int J Digit Earth*, vol. 13, no. 6, pp. 683–707, Jun. 2020, doi: 10.1080/17538947.2018.1554010.

[4] United Nations, "The Millennium Development Goals Report 2015," 2015.

[5] A. Ezeh *et al.*, "The history, geography, and sociology of slums and the health problems of people who live in slums," *The Lancet*, vol. 389, no. 10068, pp. 547–558, Feb. 2017, doi: 10.1016/S0140-6736(16)31650-6.

[6] Undang-Undang RI, *Undang-Undang Republik Indonesia Nomor 1 Tahun 2022 Tentang Perumahan dan Kawasan Permukiman*. 2011.

[7] E. E. Surtiani, "Faktor-Faktor yang Mempengaruhi Terciptanya Kawasan Permukiman Kumuh di Kawasan Pusat Kota (Studi Kasus: Kawasan Pancuran, Salatiga)," Doctoral Dissertation, Universitas Diponegoro, Semarang, 2006.

[8] B. Pujiyono, Arfian, and R. Subiyakto, "Pencegahan dan Peningkatan Kualitas Permukiman Kumuh di Kabupaten Bogor," *KRESNA: Jurnal Riset dan Pengabdian Masyarakat*, vol. 1, no. 1, pp. 11–17, 2021, [Online]. Available: https://jurnaldrpm.budiluhur.ac.id/index.php/Kresna/

[9] S. A. Jamna, "5 Kota Terbesar di Indonesia, Nomor 1 Jumlah Penduduk Sangat Padat." [Online]. Available: https://economy.okezone.com/read/2023/06/26/470/2837034/5-kota-terbesar-di-indonesia-nomor-1-jumlah-penduduk-sangat-padat?page=2

[10] R. Setyo Cahyono and J. Adianto, "Dampak Keterbatasan Akses Perumahan terhadap Kondisi Sosial Ekonomi Masyarakat Berpenghasilan Rendah di Permukiman Kumuh di DKI Jakarta," *JIMPS: Jurnal Ilmiah Mahasiswa Pendidikan Sejarah*, vol. 8, no. 3, pp. 1536–1542, 2023, doi: 10.24815/jimps.v8i3.25231.

[11] Badan Pusat Statistik Republik Indonesia, *Indikator Perumahan dan Kesehatan Lingkungan 2022*. Jakarta, 2022.

[12] UN Habitat, *The state of the world's cities 2006/2007*. Earthscan, 2006.

[13] Kementerian Pekerjaan Umum dan Perumahan Rakyat, "Tentang Program Kota Tanpa Kumuh (KOTAKU)." [Online]. Available: https://kotaku.pu.go.id/page/6880/tentang-program-kota-tanpa-kumuh-kotaku

[14] Salmaa, "Rating Scale : Pengertian, Ciri-ciri, Bentuk, Kesalahan-kesalahan, dan Contoh." [Online]. Available: https://penerbitdeepublish.com/rating-scale

[15] R. Rahman, "Machine Learning: Membuat Masa Depan Lebih Cerah." [Online]. Available: https://jayjay.co/machine-learning

[16] E. Lutins, "Ensemble Methods in Machine Learning: What are They and Why Use Them?" [Online]. Available: https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

[17] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, B. H. Springer, Ed., Springer, Berlin, Heidelberg, Jul. 2000, pp. 1–15. doi: https://doi.org/10.1007/3-540-45014-9_1.

[18] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *International Journal of Pharmaceutical Research*, vol. 12, no. 4, pp. 56–66, Oct. 2020, doi: 10.31838/ijpr/2020.12.04.013.

[19] H. Baradaran Rezaei, A. Amjadian, M. V. Sebt, R. Askari, and A. Gharaei, "An ensemble method of the machine learning to prognosticate the gastric cancer," *Ann Oper Res*, vol. 328, no. 1, pp. 151–192, Sep. 2023, doi: 10.1007/s10479-022-04964-1.

[20] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput Surv*, vol. 41, no. 3, pp. 1–52, Jul. 2009, doi: 10.1145/1541880.1541883.

[21] J. Han, M. Kamber, and J. Pei, *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*, 3rd ed. 2011.

[22] Syaidatussalihah and Abdurahim, "Classification of Poverty Status using the Random Forest Algorithm," *EIGEN MATHEMATICS JOURNAL*, vol. 5, no. 1, pp. 37–44, Jun. 2022, doi: 10.29303/emj.v5i1.133.

[23] M. L. Suliztia, "PENERAPAN ANALISIS RANDOM FOREST PADA PROTOTYPE SISTEM PREDIKSI HARGA KAMERA BEKAS MENGGUNAKAN FLASK," Thesis, Universitas Islam Indonesia, Yogyakarta, 2020.

[24] M. J. Wulansari, "ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI SESEORANG TERKENA PENYAKIT DIABETES MELITUS MENGGUNAKAN REGRESI RANDOM FOREST (Studi Kasus : Data Diabetes di Virginia Amerika Serikat)," Thesis, Universitas Islam Indonesia, Yogyakarta, 2018.

[25] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: http://www.stat.berkeley.edu/

[26] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012, [Online]. Available: http://scikit-learn.sourceforge.net.

[27] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning." [Online]. Available: https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

[28] A. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide For Data Scientist*. Sebastopol, California: O'Reilly Media, Inc, 2016.