e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v8i1.14159

Integrating Geospatial Big Data and Machine Learning for Village Level Rural Urban Classification: Evidence From Toba Regency

Meilani Thereza Br. Saragih[1*], Nurlatifah Hartojo[2]

[1]BPS-Statistics Toba Regency,
Toba, Indonesia 22315,
meilanithereza@bps.go.id

[2]Education and Training Centre, Statistics Indonesia
Jakarta, Indonesia 12620
ifah@bps.go.id

*Correspondence: meilanithereza@bps.go.id

Abstract – *This study aims to develop a data-driven framework for classifying rural and urban areas at the village level in Toba Regency by integrating official statistical data, geospatial big data, and machine learning techniques. The current regional classification still relies on the 2020 baseline and may not adequately reflect recent socio-spatial transformations occurring at finer administrative levels. To address this limitation, this study integrates Village Potential Statistics (PODES) data with spatial indicators derived from big data sources, including population density from WorldPop and the Built-Up Index (BUI) extracted from satellite imagery. The integration of these datasets enables a more comprehensive representation of settlement patterns, spatial development intensity, and demographic distribution across villages. Three supervised machine learning algorithms were implemented to this study: Support Vector Machine (SVM), Naïve Bayes, and Random Forest, with model evaluation using accuracy, precision, recall, and F1-score. The analysis results show that the Random Forest algorithm provides the best performance. Based on the best model, of the 244 villages analyzed, 156 areas were classified as rural and 88 areas as urban. These results indicate a change in status in 47 villages compared to the previous classification. These findings indicate that integrating official statistical data with big data and machine learning methods can capture the dynamics of regional development more adaptively, potentially serving as a complementary approach for compiling regional classifications and formulating more targeted development policies.*

Keywords: *Village Status; Machine Learning; Random Forest; Big Data*

I. INTRODUCTION

Indonesia is known for its abundant natural resources. However, these natural resources have not been optimally utilized to improve the welfare of the community as a whole. However, with abundant natural resources, development processes aimed at enhancing public welfare should be effectively implemented and distributed. Currently, development inequality between regions is one of the strategic issues Indonesia is still facing. Development from the smallest regional level is crucial for achieving equitable development. In response, the government formulated a development plan, outlined in Asta Cita (Presidential Regulation No. 12 of 2025), one of its main points emphasizes the importance of development starting from the villages at the lowest level of society.

To realize development from the smallest regional level, measured and comprehensive development planning is required. One form of this development planning involves the continuity of development in rural and urban areas. Integrated development planning between urban and rural areas allows for a mutually supportive flow of resources, information, and economic opportunities. To achieve this, utilizing the potential of each smallest region is crucial, thus requiring accurate and up-to-date urban and rural classifications in line with current socio-economic dynamics.

An up-to-date classification of rural and urban status plays a crucial role in capturing village development and guiding the accuracy of interventions in development policies. Urban status is defined as the status of an administrative area at the village level that meets the criteria for urban village classification. The current urban and rural classification still refers to Regulation of the Head of the BPS-Statistics Indonesia Number 120 of 2020 concerning the Classification of Urban and Rural Villages in Indonesia, which classifies urban and rural areas into villages based on urban criteria scores, namely population density, percentage of farming families, and access to various facilities that indicate urban areas. However, the Ministry of Village (2025) noted that the number of villages categorized as advanced and independent villages increases annually, indicating that changes in rural and urban status should occur in line with changes in social and economic characteristics, environmental conditions, and access within a region.

The availability of accurate and up-to-date data is a fundamental requirement in development policy planning, including in the classification of regional status. Currently, socioeconomic data is generally only available at the regency level, thus under-supporting decision-making based on smaller territorial units, such as villages. To address this limitation, the use of geospatial big data (GBD) offers significant potential in providing more detailed, small-scale information. GBD is a type of big data that specifically contains spatial or location information (Li, 2020). Geospatial data, such as satellite imagery, points-of-interest (POI), and mobility data, are widely used to identify and classify urban land characteristics, including land use functions and human activities, because they can represent the spatial and temporal dimensions of regional dynamics (Yin et al., 2021). One example of this use is the Built-up Index generated from Sentinel-2 satellite imagery to illustrate the distribution of buildings in an area. Furthermore, gridded population data can also be used to present population distribution in greater detail down to the village level. Therefore, exploring the use of geospatial big data is necessary to address the challenges posed by the limitations of conventional socioeconomic data, particularly in supporting more granular, region-based analysis and decision-making.

Toba Regency in North Sumatra Province is one of the regions that has strategic potential in supporting village development through the utilization of natural resources and the development of the tourism sector. As part of the Lake Toba area, which is designated as one of the national super-priority tourism destinations (Presidential Decree Number 89 of 2024 concerning the Master Plan for the National Tourism Destination of Lake Toba 2024-2044), Toba Regency has a great opportunity to encourage economic growth starting from the smallest administrative level, namely villages. However, there is still uneven development in the Toba Regency. This is reflected in the 2024 Gini ratio, which was recorded at 0.3480 or an increase of 0.053 points compared to 2023 (BPS, 2025). This condition also shows that Toba Regency has the highest Gini ratio among the 8 regencies in the Lake Toba area, namely Toba Regency, Samosir Regency, Simalungun Regency, Karo Regency, North Tapanuli Regency, Humbang Hasundutan Regency, Dairi Regency, and Pakpak Bharat Regency. Based on these conditions, researchers are interested in classifying villages into urban or rural areas in Toba Regency by leveraging big data and machine learning methods as a more advanced classification alternative.

Big data refers to large, diverse, and rapidly generated datasets that require advanced

analytical techniques for processing. These data are commonly characterized by the dimensions of volume, velocity, and variety, and have increasingly been utilized to support data-driven decision-making in various sectors (Al-Sai et al., 2022; Dicuonzo et al., 2022; Tosi et al., 2024).

Geospatial big data integrates large-scale datasets with spatial or location information, enabling the description of regional characteristics, land use, and human activities from sources such as satellite imagery and location-based data (Shi et al., 2020; Yao et al., 2017).

The Built-Up Index (BUI), calculated from the difference between the Normalized Difference Built-up Index (NDBI) and the Normalized Difference Vegetation Index (NDVI), is commonly used to represent the intensity of built-up areas and urban development (Alkire et al., 2021; Zhou et al., 2020). The calculation of the Built-up Index (BUI) can be formulated as follows:

$$\mathbf{BUI} = \mathbf{NDBI} - \mathbf{NDVI} \quad (1)$$

In this study, BUI is utilized as a proxy indicator to approximate the proportion of farming households at the village level.

Adapting Mariyah & Wobcke (2025) study on Proxy Mean Test modeling in Indonesia, this study uses a built-up index as a proxy to estimate the percentage of farming families in each village. This index can reflect residential density and regional economic transition, which are relevant to rural-urban classification and the composition of agricultural households.

Classification analysis (supervised learning) aims to group data into labeled classes by building a model based on training data and predicting classes from new data. Several classification analysis methods exist, such as k-Nearest Neighbor (kNN), decision trees, random forests, support vector machines (SVM), naïve Bayes, and others (Pramana et al., 2023).

Studies discussing the use of machine learning methods in classifying rural and urban areas have been conducted, including Apriliansyah et al. (2021), who classified 438 villages in the Special Region of Yogyakarta Province using the decision tree method. This study used variables such as the number and distance to shopping centers, permanent markets, junior high schools, senior high schools, hospitals, and the percentage of households using electricity in 2020. The results showed that the decision tree method was able to classify villages in the Special Region of Yogyakarta Province with an accuracy level above 89.5%, with 192 villages classified as urban and 246 villages as rural. Furthermore, this study also demonstrated that the decision tree method could be a recommended alternative to the scoring method used by BPS-Statistics Indonesia. Additionally, Almasah & Wijayanto (2023) also compared several data mining methods to classify village status in Purwakarta Regency and West Bandung Regency using variables such as access to kindergartens, junior high schools, senior high schools, village markets, shops, hospitals, hotels, pubs, and salons, as well as the percentage of families using landline telephones and PLN electricity, sourced from the 2021 Village Potential Data Collection (PODES). The results showed that the random forest method had the best classification performance, correctly classifying the data 90% of the time.

However, these studies focused solely on access to urban facilities. However, two other variables, population density and the percentage of farming families, also form the basis for the rural and urban classification by the BPS-Statistics Indonesia. Therefore, this study is expected to complement previous research and provide more in-depth findings, particularly regarding the use of big data to address the limitations of village level data for classifying rural and urban status.

II. METHODS

This study analyzed 231 villages and 13 sub-districts in Toba Regency. The research variables used were based on the criteria for urban villages as stipulated in Regulation of the Head of the BPS-Statistics Indonesia Number 120 of 2020. These variables included: area status, presence of kindergartens (KD), junior

high schools (JHS), senior high schools (SHS), market, shop clusters, hospitals, entertainment venues (hotels, billiards, pubs, discotheques, karaoke bars, salons), percentage of families using landline telephones, and percentage of families using PLN electricity, as sourced from the 2024 Village Potential Data Collection (PODES) by the BPS-Statistics Indonesia. Population density data were obtained from the WorldPop population grid, and the percentage of agricultural families was analyzed using a proxy, the Built-Up Index (BUI), to represent the intensity of regional physical development.

The detailed research variables are as follows.

Table 1. Research Variables

Variable	Variable Description	Measurement Scale
Status	1: Urban 0: Rural	nominal
Density	Population Density (people/km^2)	ratio
BUI	Built Up Index	ratio
KD	1: exist or ≤2.5 km 0: >2.5 km	nominal
JHS	1: exist or ≤2.5 km 0: >2.5 km	nominal
SHS	1: exist or ≤2.5 km 0: >2.5 km	nominal
Market	1: exist or ≤2 km 0: >2 km	nominal
Shop	1: exist or ≤2 km 0: >2 km	nominal
Hospital	1: exist or ≤5 km 0: >5 km	nominal
Entertaiment	1: exist 0: not exist	nominal
Telephone	Percentage of families using landline telephones (%)	ratio
Electricity	Percentage of families using PLN electricity (%)	ratio

This study used classification analysis, which follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) business cycle (Schröer et al., 2021).

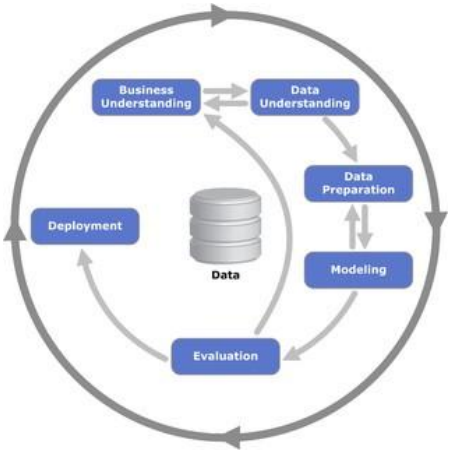

Figure 1. CRISP-DM Flow

a. Business Understanding

Business understanding aims to understand the primary business or policy objectives of a data mining project. Understanding the context is crucial for formulating appropriate and relevant analysis objectives. This process involves identifying the primary problem to be solved and how data mining can provide strategic solutions for decision-makers.

b. Data Understanding

Data Understanding encompasses initial data collection, exploration, and data quality verification. A thorough understanding of the data structure, quality, and available attributes is crucial before moving on to the next stage.

c. Data Preparation

This process encompasses all the activities required to construct the final dataset for modeling. These stages include attribute selection, data transformation, cleaning, coding categorical data, merging data from multiple sources if necessary, and adjusting for imbalanced data. The quality of this stage will significantly impact the subsequent performance of the model.

d. Modelling

At this stage, various machine learning algorithms are applied to build classification or prediction models. The entire modeling and evaluation process was carried out using R software with the caret package.

- Random Forest
  Random forest is a classification method that uses decision trees as a base classifier, which are formed and combined (Kulkarni & Sinha, 2014). Put, a random forest can be likened to a collection of decision trees. This method can accommodate large databases with thousands of input variables and address class imbalances in population data (unbalanced datasets). The first step in using this method is to form a decision tree using bootstrap sampling. Then, each decision tree predicts a value. The results of each prediction are combined by the random forest using a majority vote to determine the final value (Primajaya & Sari, 2018).

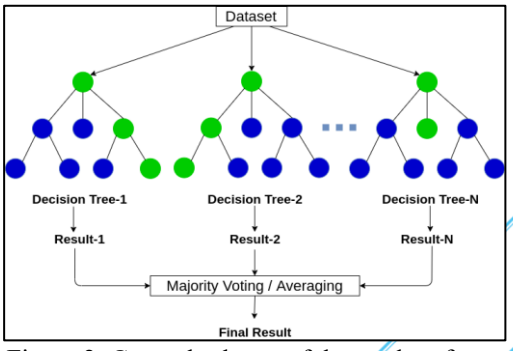

Figure 2. General scheme of the random forest method

- Support Vector Machine (SVM)
  This method is a classification method that utilizes data projection into a high-dimensional space to separate and model the data as a linear function. This classification method minimizes errors that arise during the data training process and has advantages in managing small-scale data and addressing binary classification problems (Cervantes et al., 2020).

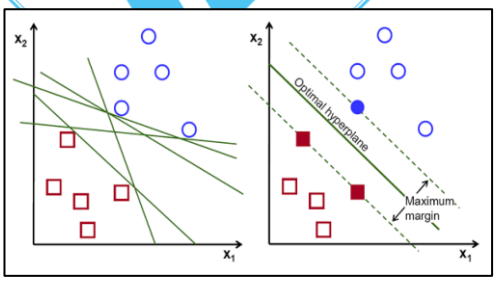

Figure 3. General scheme of the SVM method

- Naïve Bayes
  Naïve Bayes is a classification method that uses probability and statistical methods to determine the final decision. This method is quite popular due to its simplicity and the involvement of all variables in determining the final classification decision (Wibawa et al., 2019). This method is relatively easy to use, making it widely used for classification in various fields.

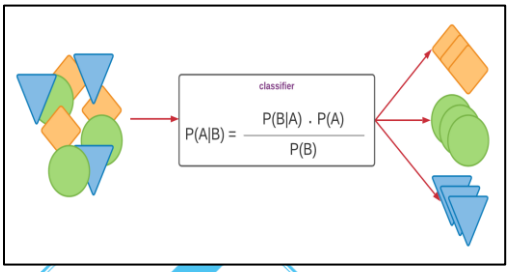

Figure 4. General schematic of the Naive Bayes method

These three algorithms were chosen because they have different characteristics in classification. Support Vector Machine is effective at separating classes in high-dimensional spaces, Naïve Bayes is a simple, probabilistic classification model, and Random Forest is known for its ability to capture complex relationships between variables and is relatively robust against overfitting.

To determine the best model, an evaluation will be conducted based on accuracy, precision, recall, and F1-score values. These four metrics were chosen because they can describe the overall model performance, both in balanced and imbalanced data contexts (Obi, 2023).

The modeling stage also includes validation using k-fold cross-validation to measure overall model performance and avoid overfitting.

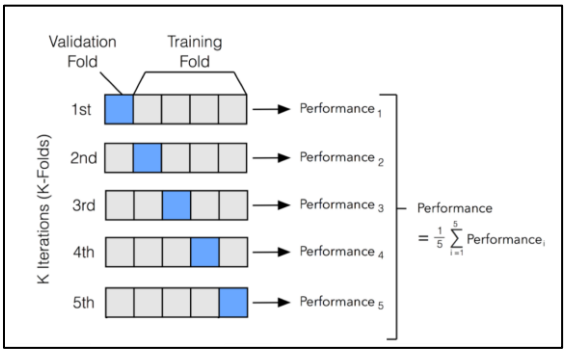

Figure 5. Cross Validation

This method is efficient and commonly used in machine learning to estimate generalization error, and offers a balance between bias and variance compared to other validation approaches (Bates et al., 2022).

e. Evaluation

After the model is built, this stage is used to assess its performance and determine whether the results meet the initial objectives of the business understanding stage. The model is evaluated based on performance metrics (accuracy, precision, recall, and F1-score).

- Accuracy
  Accuracy indicates how often the model produces correct predictions on all test data.
- Precision
  Precision measures the accuracy of the model in classifying data into the positive class.
- Recall
  Recall, or sensitivity, assesses the model's ability to recognize all true positive cases.
- F1-score
  The F1-score is the harmonic mean of precision and recall, and is used to balance the two. A high F1-score indicates a balance between the ability to detect and avoid misclassification errors.

f. Deployment

The final stage is presenting the results in a format that is easy to understand and can be used in the decision-making process. The output can be a visualization of the results, a report, or an integrated information system.

III. RESULTS AND DISCUSSION

The processed data consisted of two categories (classes): urban and rural. In this study, no missing values were found, so imputation was not performed. However, differences in the units of numeric variables, namely density, BUI, telephone, and electricity, were found. Therefore, data standardization was necessary. The purpose of data standardization is to standardize the units of all research variables. The standardization method used in this study was the min-max scaling method. Min-max standardization is performed by subtracting the minimum value for each variable from the data value and dividing it by the range of that variable.

Next, a data balance check was performed. In the Toba Regency case study, data imbalance was found, or a significant difference between the number of rural and urban areas. 197 villages were recorded as rural and 47 as urban.

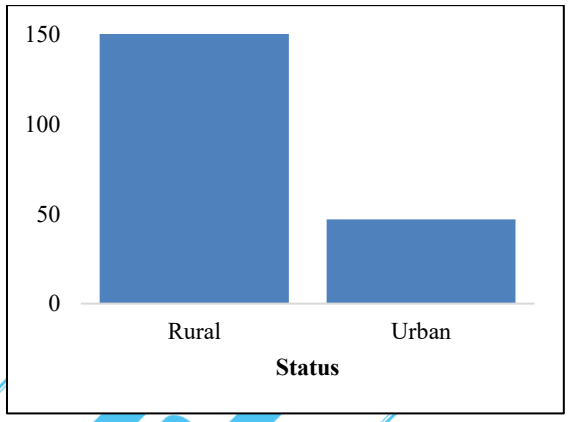

Figure 6. Distribution of Village Classification

Figure 6 shows a data imbalance between the rural and urban categories, with an imbalance ratio of 0.239. Therefore, adjustments are needed using a resampling method. To balance the distribution of rural and urban data, the researchers used the Random Over-Sampling Examples (ROSE) algorithm and obtained an imbalance ratio of 0.937 (as seen in Figure 7).

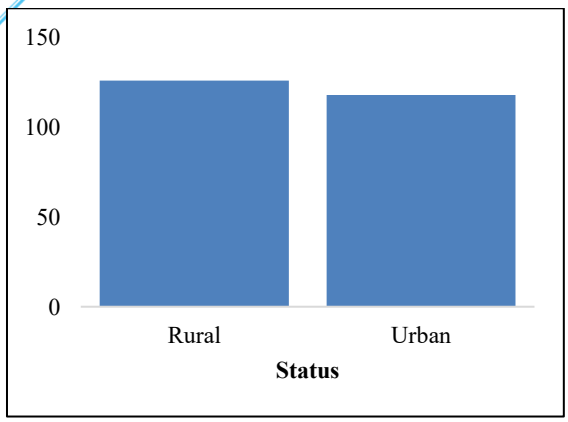

Figure 7. ROSE Resampling Results

To evaluate the classification model for rural and urban areas in Toba Regency, researchers used the 10-fold cross-validation method. Generally, this method randomly divides the data into 10 equally sized subsets.

At each iteration, the model is trained using 9 subsets and tested on the remaining 1 subset. This process is repeated 10 times, giving each data set a chance to be tested once. The evaluation results from all iterations, such as accuracy, precision, recall, and F1-score, are then averaged to obtain an estimate of the best model performance.

Support vector machine, naïve Bayes, and random forest models will be evaluated to determine the best model for rural and urban classification in Toba Regency. The results of the classification algorithm measurements are evaluated using accuracy, precision, recall, and F1-score values to determine differences or comparisons in the performance of each classification model (Syahputra & Wibowo, 2023).

Table 2. Model Evaluation Results

Validation Indicator	*Support Vector Machine*	*Naïve Bayes*	*Random Forest*
Accuracy	0.8813	0.9345	0.9383
Precision	0.8896	0.9381	0.9444
Recall	0.8808	0.9436	0.9365
F1-Score	0.8833	0.9363	0.9391

Based on the evaluation results in Table 2, the random forest model demonstrated the best performance across three key indicators: accuracy (93.83%), precision (94.44%), and F1-score (93.91%), reflecting the model's ability to correctly classify data, generate accurate positive predictions, and maintain a balance between false positives and false negatives. Although the naïve Bayes model achieved the highest recall value of 94.36%, indicating its ability to detect a large proportion of true positives, it fell short in terms of precision and F1-score, which are crucial for maintaining overall accuracy in the context of imbalanced data. Therefore, considering all evaluation metrics, random forest was selected as the best model for classifying rural and urban areas in Toba Regency because it provided more stable, reliable, and accurate results across various performance evaluation aspects.

By applying the random forest algorithm to 244 villages in Toba Regency, classification results were obtained that classified 156 areas as rural and 88 areas as urban. These results indicate a change in status in 47 villages compared to the previous classification based on the BPS scoring method, with 3 areas changing from urban to rural, and 44 other areas changing from rural to urban. A visualization of the comparison between the Random Forest classification results and the BPS-Statistics Indonesia scoring method is presented in Figure 8 below.

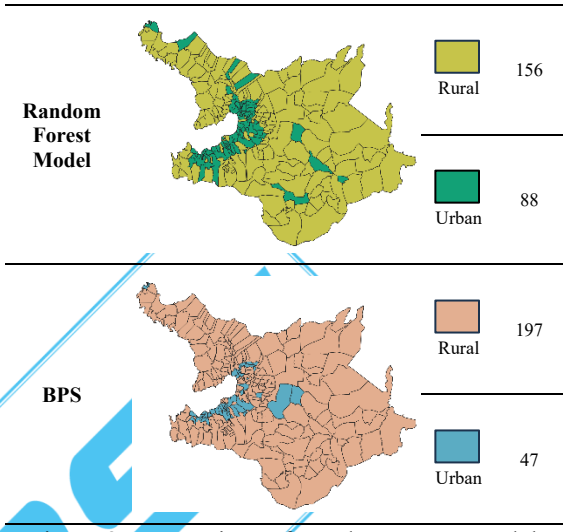

Figure 8. Comparison of Random Forest Model Classification with BPS Classification

These changes reflect regional dynamics likely influenced by infrastructure development, socioeconomic activity, and residential growth. Machine learning algorithms such as random forests, supported by up-to-date data and big data sources, enable the identification of spatial patterns that are more adaptive to actual conditions on the ground. These findings also have practical implications for formulating regional development policies, particularly in supporting village-based development planning and prioritizing more targeted infrastructure and public services at the local level.

This situation is also evident spatially, where several areas previously categorized as rural now exhibit high levels of physical development, a decline in the agricultural sector's dominance, and an increase in non-agrarian economic activity, captured through the Built-Up Index (BUI) variable. The random forest model is able to capture this complexity by simultaneously combining PODES data and big data geospatial indicators. This finding is

also in line with Zhang et al. (2018), which shows that artificial intelligence techniques for pattern recognition, such as neural networks, support vector machines, and random forests, are effective for classifying satellite images and identifying spatial patterns in regions. Therefore, machine learning methods, particularly random forests, have the potential to strengthen regional classification processes, making them more responsive to actual development dynamics.

IV. CONCLUSION

This study demonstrates that the random forest algorithm is the most optimal method for classifying rural and urban areas in Toba Regency, when compared to naive Bayesian and support vector machine models. The evaluation results show that the random forest algorithm excels in almost all key evaluation metrics, namely accuracy, precision, recall, and F1-score. Applying this model yielded a new classification, indicating that 47 villages changed status compared to the previous classification. These findings demonstrate that a big data-based machine learning approach can provide a more adaptive regional classification to the dynamics of regional development and has the potential to support the formulation of more targeted development policies.

Beyond simply selecting a model, this study emphasizes the importance of utilizing big data-based machine learning as an alternative approach that can provide more sophisticated, objective, and evidence-based regional classifications. Integrating official statistical data from BPS-Statistics Indonesia with modern spatial data can enhance the quality of analysis, deepen understanding of micro-level regional characteristics, and support the formulation of development policies that are more responsive to changing regional conditions.

As a follow-up, it is recommended that future development of regional classification systems involve closer collaboration between the BPS-Statistics Indonesia, local governments, and academics. Future researchers can also explore more relevant big data-based variables, such as population mobility data, settlement density from high-resolution satellite imagery, the quality of digital infrastructure, and the intensity of location-based economic activity. In-depth interviews with local governments can also be added to complement the classification generated by the machine learning model. This multidimensional approach can not only improve the accuracy of classification results but also broaden the research's contribution to the development of science and the formulation of development policies that are more responsive to the ever-changing realities of the region.

REFERENCES

Al-Sai, Z. A., Husin, M. H., Syed-Mohamad, S. M., Abdin, R. M. S., Damer, N., Abualigah, L., & Gandomi, A. H. (2022). Explore Big Data Analytics Applications and Opportunities: A Review. *Big Data and Cognitive Computing*, *6*(157), 1–23. https://doi.org/10.3390/bdcc6040157

Alkire, S., Kanagaratman, U., & Suppa, N. (2021). The Global Multidimensional Poverty Index (MPI): 2021. *OPHI MPI Methodological Note 51*, 1–39. https://www.ophi.org.uk/wp-content/uploads/OPHI_MPI_MN_51_2021_4_2022.pdf

Almasah, M. Z., & Wijayanto, A. W. (2023). Comparison of Data Mining Methods in Classifying Village Status of Purwakarta and West Bandung Regencies (Podes 2021). *Eigen Mathematics Journal*, *6*(1), 5–10. https://doi.org/10.29303/emj.v6i1.156

Apriliansyah, Pangestika, A., Ramadhanty, A. P., Putra, G. M., Putri, G. S. N. D. S., & Nooraeni, R. (2021). Classification of Village/Sub-district Status in Special Region of Yogyakarta Using the Decision Tree Model (Case Study of Field Work Practice Data of Politeknik Statitika STIS 2020). *Engineering, MAthematics and Computer Science (EMACS) Journal*, *3*(1), 33–41. https://doi.org/10.21512/emacsjournal.v3i1.6787

Bates, S., Hastie, T., & Tibshirani, R. (2022). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association*, 1–43. https://doi.org/10.1080/01621459.2023.2197686

BPS-Statistics Sumatera Utara Province. (2025). *Gini Ratio by Regency/Municipality in Sumatera Utara, 2024*. https://sumut.bps.go.id/en/statistics-table/2/NDY3IzI=/gini-ratio-sumatera-utara-menurut-kabupaten-kota.html

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 1–27. https://doi.org/10.1016/j.neucom.2019.10.118

Dicuonzo, G., Galeone, G., Shini, M., & Massari, A. (2022). Towards the Use of Big Data in Healthcare: A Literature Review. *Healthcare*, *10*(1232), 1–16. https://doi.org/10.3390/healthcare10071232

Kulkarni, V. Y., & Sinha, P. K. (2014). Effective Learning and Classification using Random Forest Algorithm. *International Journal of Engineering and Innovative Technology (IJEIT)*, *3*(11), 267–273.

Li, Z. (2020). *Geospatial Big Data Handling with High Performance Computing: Current Approaches and Future Directions*. 53–76. https://doi.org/10.1007/978-3-030-47998-5_4

Mariyah, S., & Wobcke, W. (2025). Evaluating area-level features for proxy means test models: evidence from rural, semi-urban and urban districts in poverty targeting. *Journal of Computational Social Science*, *8*(3). https://doi.org/10.1007/s42001-025-00405-8

Ministry of Villages and Development of Disadvantaged Regions of the Republic of Indonesia. (2025). *Village Development Index (IDM) Status*. https://idm.kemendesa.go.id/

Obi, J. C. (2023). A Comparative Study of Several Classification Metrics and Their Performances on Data. *World Journal of Advanced Engineering Technology and Sciences*, *08*(01), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054

Pramana, S., Yuniarto, B., Santoso, I., Nooraeni, R., & Suadaa, L. H. (2023). *Data Mining with R Concepts and Implementation* (2 ed.). In Media.

Presidential Regulation (Perpres) No. 12 of 2025 concerning the National Medium-Term Development Plan (RPJMN) 2025-2029., Pub. L. No. 12 of 2025 (2025).

Presidential Regulation (Perpres) No. 89 of 2024 concerning the Master Plan for the National Tourism Destination of Lake Toba for 2024 - 2044, Pub. L. No. 89 of 2024 (2024).

Primajaya, A., & Sari, B. N. (2018). Random Forest Algorithm for Prediction of Precipitation. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)*, *1*(1), 27–31. https://doi.org/10.24014/ijaidm.v1i1.4903

Regulation of The Head of BPS-Statistics Indonesia Number 120 of 2020 Concerning Urban and Rural Village Classification in Indonesia 2020, Pub. L. No. 120 of 2020 (2020).

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Shi, K., Chang, Z., Chen, Z., Wu, J., & Yu, B. (2020). Identifying and evaluating poverty using multisource remote sensing and point of interest (POI) data: A case study of Chongqing, China. *Journal of Cleaner Production*, *255*, 120245. https://doi.org/https://doi.org/10.1016/j.jc

lepro.2020.120245

Syahputra, H., & Wibowo, A. (2023). Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, *9*(1), 165–173. https://doi.org/10.26555/jiteki.v9i1.25861

Tosi, D., Kokaj, R., & Roccetti, M. (2024). 15 years of Big Data: a systematic literature review. *Journal of Big Data*, *11*(73), 1–39. https://doi.org/10.1186/s40537-024-00914-9

Wibawa, A. P., Kurniawan, A. C., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. *International Journal of Recent Contributions from Engineering, Science & IT (iJES)*, *7*(2), 91. https://doi.org/10.3991/ijes.v7i2.10659

Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., & Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, *31*(4), 825–848. https://doi.org/https://doi.org/10.1080/13658816.2016.1244608

Yin, J., Dong, J., Hamm, N. A. S., Li, Z., Wang, J., Xing, H., & Fu, P. (2021). Integrating remote sensing and geospatial big data for urban land use mapping: A review. *International Journal of Applied Earth Observation and Geoinformation*, *103*, 102514. https://doi.org/10.1016/j.jag.2021.102514

Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., & Zhang, S. (2018). Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery. *Sensors*, *18*(11), 1–21. https://doi.org/10.3390/s18113717

Zhou, C., Li, F., Zhang, J., Zhao, J., Zhang, Y., & Wang, J. (2020). Analysis of Spatial and Temporal Variations of Vegetation Index in Liaodong Bay in the last 30 years based on the GEE Platform. *IOP Conference Series: Earth and Environmental Science*, *502*, 1–8. https://doi.org/10.1088/1755-1315/502/1/012037