JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 18, NO 1, June 2026 P-ISSN 2086 – 4132 E-ISSN 2615 – 1367

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS) contains scientific papers on research findings and theoretical studies of statistics and computational statistics applied in fields, that is published twice a year in June and December. This journal is published by Politeknik Statistika STIS.

Editorial Address:
Politeknik Statistika STIS
Jl. Otto Iskandardinata 64C
Jakarta Timur 13330
Telp. 021-8191437

The editorial accepts scientific papers or research articles on theoretical studies of statistics and computational statistics in fields. The editorial has the right to edit writings without changing the substance of the writing. The contents of the Jurnal Aplikasi Statistika & Komputasi Statistik are cited by referring to the source material.

Editorial Foreword

Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS) Volume 18 Number 1 June 2026 Edition has undergone transformations such as writing articles in English, establishing a journal logo, establishing the Politeknik Statistika STIS publisher logo, changing paper template designs, updating Author Guidelines, etc. The aim of this transformation is to improve the journal's performance and expand the reach of JASKS readers.

This issue consists of 8 articles contributed by 24 authors affiliated with various universities and government institutions from Indonesia and abroad. The contributing institutions include BPS-Statistics Indonesia, BPS-Statistics Bekasi Regency, the Training and Education Center of Statistics Indonesia, Institut Agama Islam Negeri Sorong, UIN Sulthan Thaha Saifuddin Jambi, Universitas Andalas, Institut Teknologi Sepuluh Nopember, Kupang State Polytechnic of Agriculture, Mulawarman University, Universitas Udayana, Universitas Indonesia, Universitas Lambung Mangkurat, and the University of Leeds (United Kingdom). The articles in this issue demonstrate diverse applications of statistical and computational methods to address economic, environmental, agricultural, and social issues in Indonesia. The studies cover structural equation modeling, dimension reduction, spatial regression, geographically weighted modeling, machine learning, ensemble learning, class balancing, topic modeling, sentiment analysis, and forecasting methods. Collectively, these contributions highlight the growing role of statistical and computational approaches in transforming official statistics and socio-economic data into evidence that supports informed decision-making, regional development planning, and public policy formulation.

The editorial board extends its sincere appreciation to all authors and reviewers for their valuable contributions and rigorous evaluations that have ensured the quality of this issue. We also thank the editorial team for their dedication in managing the publication process. It is hoped that this issue will enhance readers' understanding of applied statistical and computational methods, and the editorial team welcomes further scholarly contributions to support the dissemination of statistical knowledge within the community.

Jakarta, June 2026

Editor-in-Chief,

Fitri Kartiasih

JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 18, NO 1, JUNE 2026

Contents

DOI: https://doi.org/10.34123/jurnalasks.v18i1.784

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Financial Literacy's Impact on Interest in Sharia Investment: An Examination Using SEM-PLS

Sella Nofriska Sudrimo[1]*, Urwawuska Ladini[2], Dahlia Misrika[3]

[1]Institut Agama Islam Negeri Sorong, Sorong, Indonesia, [2]UIN Sulthan Thaha Saifuddin Jambi, Jambi, Indonesia, [3]Universitas Andalas, Padang, Indonesia
**Corresponding Author: E-mail address:* sellans@iainsorong.ac.id

ARTICLE INFO

Article history:
Received 17 July, 2024
Revised 19 Dec, 2025
Accepted 2 June, 2026
Published 30 June, 2026

Keywords:
Financial Inclusion; Financial Literacy; Sharia Investment; SEM-PLS; Southwest Papua

Abstract

Introduction/Main Objectives: This research explores the impact of financial literacy, including knowledge, confidence, and ability, on interest in Sharia investment. **Background Problems:** the relationship between financial literacy and interest in sharia investment. **Novelty:** Considering that each province has different community characteristics, especially in the Southwest Papua Province area which is included in the 3T areas (underdeveloped, frontier, and outermost), an analysis was carried out and using SEM-PLS for analysis of Ddta. **Research Methods:** Using the Structural Equation Model Partial Least Square (SEM-PLS) method with the response variable sharia investment interest and the financial literacy variable with the dimensions of knowledge, confidence and ability. **Finding/Results:** The results show that financial literacy has a positive and statistically significant effect on interest in Sharia investment. Although its explanatory power is limited ($R^2 = 0.094$), the findings indicate that financial literacy functions as an enabling factor rather than a sole determinant of Sharia investment interest. These results suggest that improving financial literacy alone is insufficient to substantially increase interest in Sharia investment. Therefore, policies aimed at promoting Sharia investment should integrate financial education with institutional trust-building, product accessibility, and socio-religious engagement, particularly in underdeveloped and frontier regions such as Southwest Papua.

1. Introduction

An intriguing field of study lies within finance, given its significant role in our lives. The state of our finances frequently shapes our overall well-being, with those who are financially secure generally enjoying a higher quality of life compared to those facing financial difficulties. It is undeniable that having financial knowledge significantly impacts one's ability to achieve financial stability. A study conducted by the OECD/INFE International Survey of Adult Financial Literacy (2020) found that individuals with high financial literacy tend to have more savings, more manageable debt, and are better able to plan for retirement. Another finding was obtained from research conducted by The Economic Importance of Financial Literacy: Theory and Evidence [1], that adults who received financial education

(for example, in high school) had higher credit scores, less credit card debt, and were more likely to have an emergency fund than those who did not.

As per the Financial Services Authority Regulation Number 76/POJK, which aims to improve financial literacy for the public and consumers in the financial services sector, it is crucial to comprehend that financial literacy comprises knowledge, skills, and beliefs that influence attitudes and behavior, ultimately resulting in better financial management and decision-making for increased prosperity. Financial literacy is crucial because it enables individuals to successfully plan for the future and make wise financial decisions in everyday life. A thorough understanding of financial literacy helps individuals make long-term financial decisions such as investing, as well as short-term decisions about saving and consumption. Low financial literacy also contributes to the high number of illegal online lending cases and individual bankruptcies. According to data compiled on the Financial Services Authority (OJK) website (2023), the Illegal Financial Activities Eradication Task Force (Satgas PASTI) (formerly the Investment Alert Task Force) recorded that from 2017 to October 31, 2023, the Task Force had shut down 7,502 illegal financial entities, consisting of 1,196 illegal investment entities, 6,055 illegal online lending/PINPR entities, and 251 illegal pawnshop entities. Similarly, the OJK's National Financial Literacy and Inclusion Survey (2022) showed that the financial literacy rate in Indonesia was only 49.68%. This indicates low financial literacy among Indonesians, leading many to fall prey to illegal investments. Investing is the process of carefully allocating money to projects with the potential to generate future profits. People who lack financial literacy can experience problems such as excessive debt, poor investments, or financial instability following unexpected events. Therefore, improving financial literacy among Indonesians is crucial because it directly benefits the financial well-being of individuals and society as a whole.

Financial literacy is essential for many people, especially students who are at a transitional time of life when the responsibility for managing personal financial resources starts to expand. In addition to being susceptible to consumer debt, especially the widespread illicit internet loans in Indonesia, students with low levels of financial literacy run the danger of having trouble saving and managing their expenditures. A person's capacity to make prudent long-term investment decisions may also be hampered by a lack of financial literacy. Students' interest in Sharia-compliant investments is rising along with investment awareness, mostly due to their view of the possible profits. The combination of financial knowledge and profit expectations are factors that encourage individuals to consider investments, particularly Sharia-compliant investments, which are considered to align with religious and ethical values.

Previous studies have documented the important role of financial literacy in shaping individual financial behavior and investment decisions. Several studies have found a positive relationship between financial literacy and investment decisions, indicating that individuals with higher levels of financial knowledge and competence tend to make more rational and informed investment decisions [2]- [3]. This finding confirms that financial literacy serves as a key determinant in financial decision-making.

However, empirical evidence regarding the influence of financial literacy on investment interest still shows mixed results. Some studies find that financial literacy has a significant effect on investment interest, while others report a weak or insignificant effect, particularly among university students [4]. These differing findings indicate that the relationship between financial literacy and investment interest is likely influenced by contextual factors, such as demographic characteristics, the institutional environment, and an individual's level of exposure to financial activities.

Furthermore, most previous studies have focused on Islamic financial literacy as a determinant of Islamic investment interest [5]-[6], while general financial literacy tends to receive less attention. However, general financial literacy—which encompasses basic knowledge, confidence, and skills in financial management—can serve as an important foundation for understanding Sharia-compliant investment principles. Furthermore, empirical studies examining this relationship in remote and underdeveloped regions, such as Southwest Papua, are still very limited.

Based on this gap, this study aims to analyze the effect of general financial literacy on Islamic investment interest among IAIN Sorong students using the Structural Equation Modeling–Partial Least Squares (SEM-PLS) approach. This approach was chosen for its ability to predictively analyze complex relationships between latent variables, even with a relatively limited sample size. Thus, this study is expected to clarify inconsistent findings in previous studies and provide empirical contributions from regional contexts that are still rarely studied in Islamic finance literature.

2. Material and Methods

2.1. *Type of Research*

The impact of financial literacy on interest in Sharia investing is investigated in this study using a quantitative explanatory research approach. While the explanatory approach is utilized to evaluate causal links among variables, quantitative approaches are used since the research variables are measurable and appropriate for statistical analysis. Interest in Sharia investing is considered an endogenous latent variable, while financial literacy is considered an exogenous latent variable. To ascertain the direction and importance of the relationship between these constructs, the study focuses on hypothesis testing. The study uses a cross-sectional approach, gathering data all at once to determine students' present levels of financial literacy and interest in investing. When studying relationships without taking time changes into account, this design is suitable. Data analysis is conducted using the Structural Equation Modeling–Partial Least Squares (SEM-PLS) approach, which is suitable for explanatory research with predictive objectives, relatively small sample sizes, and minimal distributional assumptions.

2.2. *Data and Strategies*

This study was carried out in the State Islamic Institute (IAIN) Sorong, Southwest Papua, over the course of two months, from July to August 2023. The location was selected because IAIN Sorong is a religious university with a sizable Muslim student body, making it a pertinent setting for analyzing interest in sharia-compliant investment.

Based on official statistics from the Bureau of Academic and Student Affairs of IAIN Sorong, 2023, the study's population consisted of all enrolled students at IAIN Sorong during the 2022–2023 academic year, or roughly 2,850. In order to guarantee that every member of the population had an equal chance of being chosen as a respondent, this study used a straightforward random sampling technique due to time and resource restrictions. The sample size is determined using the Slovin formula by [7]:

$$n=\frac{N}{1+Ne^2} \tag{1}$$

with a minimum sample size of 99 respondents and a 10% margin of error. Given the exploratory nature of the study, the restricted access to respondents during academic holidays, and [8]'s suggestion that SEM-PLS research can accept smaller samples as long as they satisfy the minimal statistical requirements, the 10% margin of error was used. The sample of this research was 115 students of IAIN Sorong.

Instrument validity was tested using loading factors in a PLS-based Confirmatory Factor Analysis (CFA), where indicators with loading factors ≥ 0.70 were considered valid [9]. Meanwhile, reliability was measured using two indicators: Cronbach's Alpha and Composite Reliability (CR), with a minimum value of 0.70 for both as the threshold for acceptance [10]. The researchers ensured adherence to the principles of informed consent, confidentiality, and voluntary participation. Each respondent was given an initial explanation of the study's purpose, benefits, and their right to withdraw at any time without consequence. Respondent identities were not included in the questionnaire, and all data were anonymized and used solely for academic purposes [11]-[12].

2.3. *Research Procedures*

This study utilizes quantitative research methods, with financial literacy as the independent variable (X) and interest in sharia investment as the dependent variable (Y). Financial literacy is measured across three dimensions: knowledge, confidence, and skills. The data analysis methodology utilized in this research is the Structural Equation Model-Partial Least Squares (SEM-PLS). The stages of data analysis using SEM-PLS include:

(1) Analysis of the measurement model (outer model)

At this point, convergent and discriminant validity are assessed as part of construct validity testing. Additionally, the composite reliability (CR) value is evaluated as part of construct reliability testing. The loading factor value can be used to calculate the correlation between the reflective indicator score and the latent variable score. PLS is used to quantify the correlation between item/component scores to evaluate the measurement model with reflected indicators. If the loading factor value correlates with the measured construct more than 0.7 ($\lambda = 0{,}5$), It is considered high. However, a loading value measuring scale of 0.5 to 0.6 is thought to be enough for the first stage [13]. According to [8] and [14], a factor loading of 0.50 or more ($\geq 0{,}50$) is deemed to have adequate validity to explain the latent concept. According to several publications, ensuring good validation of the underlying construct requires a loading factor of at least 0.50.

Discriminant validity is a component of construct validity, which aims to ensure that each construct in the model is empirically distinct from other constructs. This means that the indicators (items) used to measure a construct are not excessively correlated with other constructs, ensuring that the construct has unique meaning [9]. Cross-loading correlation is one way to test discriminant validity. When compared to the loading on other constructions, the indicator must have the highest loading on the construct it measures.

The consistency with which indicators within a construct measure the same idea is evaluated using reliability tests. Cronbach's alpha and composite reliability are two techniques used in reliability testing.Assuming that all indicators have the same weight (tau-equivalency) and measuring reliability using the correlation between items [14].

$$\alpha = \frac{k}{k\text{-}1}\left(\frac{\sum \text{Var(i)}}{\text{var(total)}}\right) \tag{2}$$

Table 1. Criteria for Cronbach's Alpha

Cronbach's Alpha	Interpretation
$value \geq 0.70$	Good reliability
$0.60 \leq value < 0.70$	Still acceptable for exploratory research
$value < 0.60$	Not reliable

Reliability is measured by considering the indicator's actual weight (outer loading) [14].

$$CR = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum\left(1\text{-}\lambda_i^2\right)} \tag{3}$$

(2)Analysis of the structural model (inner model)

The inner model evaluates the relationships between latent constructs in the theoretical model: whether the independent (exogenous) variables influence the dependent (endogenous) variables, how strong the influence is, whether the influence is significant, and whether the model has predictive ability (in-sample and out-of-sample).

Tests to analyze the inner model include [9]:

a) Calculate R^2 (Coefficient of Determination) for each endogenous construct

$$R^2 = \frac{\text{var x}}{\text{varian total}} \tag{4}$$

The higher the R^2, the greater the predictive ability of the exogenous construct against the endogenous one.

b) Calculate f^2 (Effect Size)

Determine the degree to which exogenous constructs contribute to specific endogenous constructs [9] and [15].

$$f^2 = \frac{R^2_{\text{included}} - R^2_{\text{excluded}}}{1 - R^2_{\text{included}}} \tag{5}$$

Table 2. Criteria for f^2 (Effect Size)

f^2 values	Category	Interpretation
0.02	Small	The variable makes a small contribution to the endogenous R^2
0.15	Moderat	Variables provide moderate contributions
0.35	Big	Variables make a big contribution

c) Testing Q^2 (Predictive Relevance)

Evaluate the model's prediction power in relation to the endogenous constructs' manifest indicators.

$$Q^2=1-(1-R_1^2)(1-R_2^2)(1-R_3^2)\ldots(1-R_n^2) \quad (6)$$

(3) Conversion of path diagrams to equations

Figure 1 presents the structural framework of the research model and the conversion of the path diagram into an analytical form. The model illustrates the relationships among latent variables examined in this study. Financial Literacy is positioned as a higher-order latent construct, which is reflected by three first-order latent variables, namely Knowledge, Confidence, and Skills. These three dimensions represent the core components of general financial literacy and are hypothesized to positively contribute to its formation.

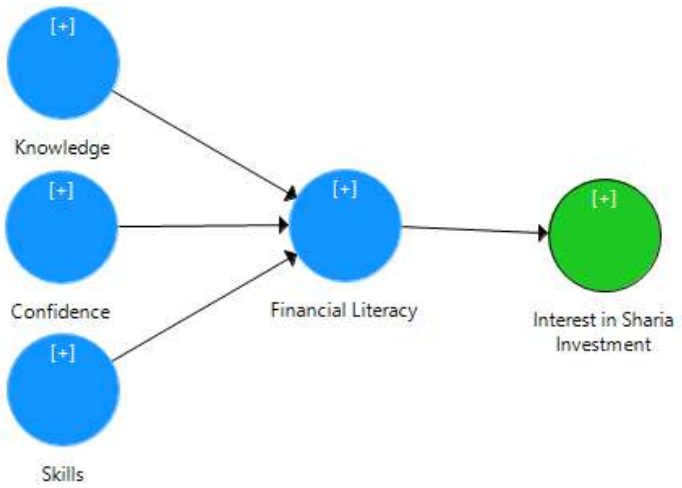


Figure 1. Variable Framework

The model specifies a direct causal relationship between Financial Literacy and Interest in Sharia Investment. This structural path indicates that an increase in an individual's level of financial literacy is expected to enhance their interest in Sharia-compliant investment instruments. The direction of the arrows represents the hypothesized causal influence among latent variables, while the absence of reciprocal arrows indicates a unidirectional relationship in accordance with the theoretical framework. The diagram also reflects the reflective measurement nature of the model, where each latent construct is measured by its respective indicators (not shown in detail in the figure for clarity). The conversion of this path diagram into structural and measurement equations allows for empirical testing using the SEM-PLS approach, enabling the estimation of path coefficients, factor loadings, and error terms associated with each latent variable. Overall, this framework provides a comprehensive representation of the theoretical assumptions underlying the study and serves as the basis for hypothesis testing and subsequent structural model evaluation.:

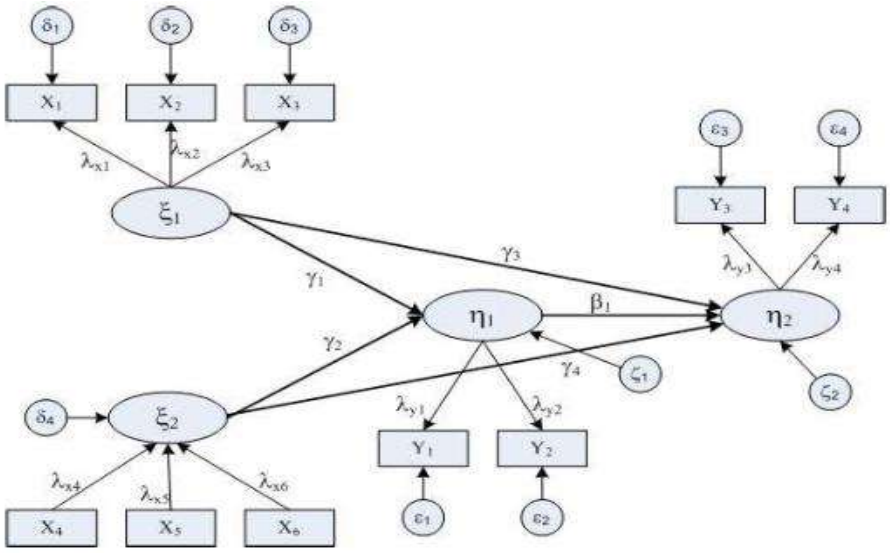


Figure 2. Path Diagram

The Structural Equation Model - Partial Least Squares Model (SEM-PLS) utilized in this study is:

a) Measurement

$$\xi_{ij}=\lambda_{ij}X_i+\delta_{ij} \quad \text{and} \quad \eta_{kj}=\lambda_{kj}Y_k+\varepsilon_{kj} \tag{7}$$

b) Structural (Inner)

$$\eta_m=\sum b_{mp}\eta_p+\sum \gamma_{mq}\xi_q+\zeta_m \tag{8}$$

Which ξ is the exogenous latent variable, η is the endogenous latent variable, λ is an exogenous latent variable factor loading, β is the coefficient of influence of exogenous variables on endogenous variables, ζ is the endogenous latent variable model error, and ε is the exogenous latent variable model error.

(4) *Hypothesis testing*

The hypothesis can be formulated as follows:

H_1 : Knowledge of financial literacy and interest in Sharia-compliant investments are significantly correlated.

H_2 : Financial literacy and knowledge are significantly correlated.

H_3 : Financial literacy and confidence are significantly correlated.

H_4 : Skills and financial literacy are significantly correlated.

3. Results and Discussion

3.1. *Measurement Model Analysis*

The measurement findings, which the measurement model directly observes in the relationship between observable indicators and latent variables, comprise the dimensions of the factor variables. There are two types of measuring models available: formative and reflective. The measurement model is examined using the following two tests: Tests for reliability and validity.

Table 3. Descriptive Statistics

Item	Mean	Median	Min	Max	Standard Deviation
X.X1	3.104	3	2	4	0.651
X.X2	3.261	3	2	4	0.576
X.X3	3.174	3	2	4	0.594
X.X4	3.122	3	2	4	0.621
X.X5	3.217	3	2	4	0.587
X.X6	3.209	3	2	4	0.679
X.X7	3.217	3	2	4	0.587
X.X8	3.165	3	2	4	0.543
X.X9	3.200	3	2	4	0.662
Y.Y1	3.070	3	1	4	0.766
Y.Y2	3.009	3	1	4	0.692
Y.Y3	2.965	3	1	4	0.697
Y.Y4	2.983	3	2	4	0.646
Y.Y5	3.096	3	2	4	0.632
Y.Y6	3.000	3	1	4	0.746
Y.Y7	3.157	3	2	4	0.641
Y.Y8	3.026	3	1	4	0.665
Y.Y9	3.052	3	2	4	0.683

Based on the results of the descriptive analysis in Table 3, it was found that the average value of variable X was in the quite high category (Mean = 3.19), while variable Y was in the sufficient category (Mean = 3.04). This indicates that respondents gave positive responses to the measured aspects, but there is still room for improvement. The standard deviation value ranging from 0.54–0.77 indicates that respondents' answers were relatively consistent and there were no extreme differences between individuals. Thus, in general, respondents assessed that variable X had been implemented well and had a moderate impact on variable Y.

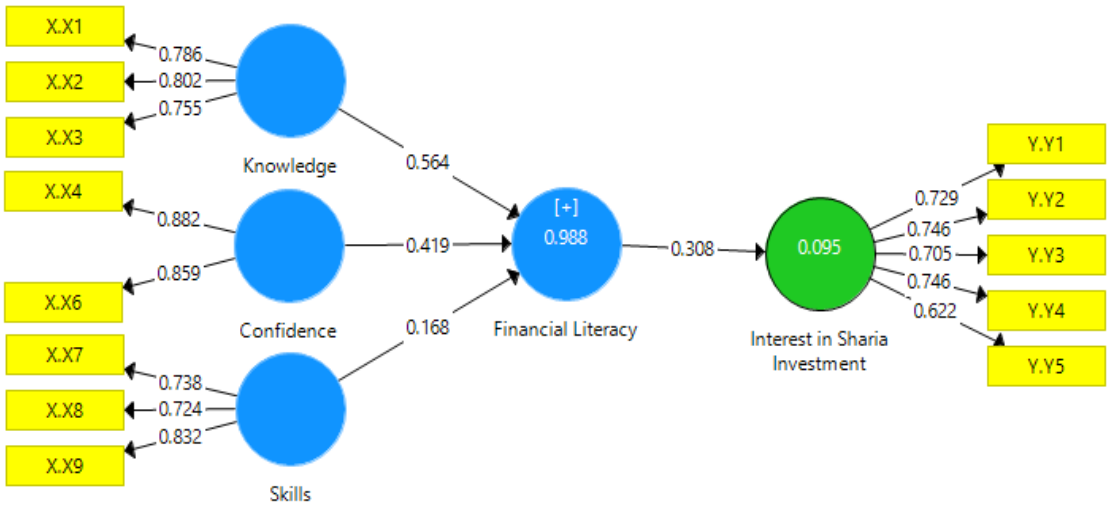


Figure 3. Path Chart and Loading Factor Value After Elimination of Invalid Indicators

According to [16], a validity test is used to determine if the indicators in a research instrument are valid and can accurately measure the latent variables being studied. Study by [17] further explain that the validation test includes assessments for both convergent and discriminant validity.

An indicator is considered to have convergent validity if the loading of its outer value is larger than 0.5. The results, as shown in Figure 3, show that every indication has a loading factor value more than

0.50 ($\lambda \geq 0.5$).This suggests that the underlying construct can be explained with strong validity by all indications of each dimension of the latent variable, including knowledge, confidence, and abilities in the latent variable financial literacy and the latent variable interest in investing in Sharia. Consequently, each indicator satisfies the convergent validity criterion

The test of discriminant validity determines if the observed measures differ from one another. Examining cross-loading values is one method of performing discriminant validity testing for reflecting indicators. When each measure's cross-loading correlation value is higher with its construct than with other constructs, it indicates good discriminant validity. Table 4 contains the precise cross-loading correlation values.

Table 4. Crossloading Correlation Values

	Skill	Confidence	Knowledge	Financial Literacy	Interest in Sharia Investment
X.X1	0.379	0.594	0.786	0.761	0.237
X.X2	0.422	0.432	0.803	0.700	0.168
X.X3	0.472	0.389	0.755	0.665	0.151
X.X4	0.527	0.882	0.548	0.774	0.254
X.X6	0.406	0.859	0.513	0.711	0.233
X.X7	0.738	0.513	0.358	0.502	0.224
X.X8	0.724	0.334	0.426	0.464	0.179
X.X9	0.832	0.396	0.457	0.623	0.261
Y.Y1	0.204	0.199	0.205	0.246	0.756
Y.Y2	0.141	0.244	0.209	0.244	0.764
Y.Y3	0.303	0.159	0.142	0.200	0.710
Y.Y4	0.246	0.227	0.151	0.221	0.756

As shown in Table 4, all indicators exhibit cross-loading correlation values above 0.70, as well as stronger correlations with their respective constructs than with other constructs. When assessing a model's reliability, don't forget to take Cronbach's Alpha and Composite Reliability values into account. If a model's composite reliability score is higher than 0.7 and its Cronbach's alpha value is more than 0.6, it is deemed reliable. Table 5 contains comprehensive reliability test findings.

Table 5. Reliability Test

	Cronbach's Alpha	Composite Reliability
Skills	0.649	0.809
Confidence	0.682	0.863
Financial Literacy	0.799	0.857
Interest in Sharia Investment	0.736	0.834
Knowledge	0.681	0.824

Table 5 shows that the latent variables of financial literacy and interest in Sharia investment, as well as the dimensions of skills, confidence, and knowledge, all have Cronbach's Alpha values larger than 0.6. In the same way, all dimensions and latent variables have Composite Reliability values greater than 0.7. Consequently, we can conclude that every dimension and variable is trustworthy.

3.2. Structural Model Analysis (Inner Model)

The values of R-square, F-square, and Q-square can be used to examine structural models. Take note of the following table.

Table 6. R-square Values

Variable	Adjusted R-square
Financial Literacy	0.988
Interest in Sharia Investment	0.086

Table 6 clearly shows that the R-squared corrected value is 0.086. This suggests that 91.4% of Sharia Investment Interest may be explained by factors not included in the model, with Financial Literacy accounting for 8.6% of the total.

Table 7. F-square Values

Variable	F-square
Skills	1.533
Confidence	8.433
Knowledge	15.125
Financial Literacy	0.104

Table 7 shows that the knowledge, skills, and confidence components of financial literacy have an F-square value greater than 0.35. The derived F-square value falls into the big group according to the F-square assessment criteria shown in Table 7, suggesting that these dimensions have a significant impact on financial literacy.

Conversely, the interest in Sharia investing is positively correlated with the latent variable of financial literacy, with an F-squared value of 0.104. With an f-squared value below 0.15, this suggests that interest in Sharia investing is somewhat influenced by financial literacy.

The Q-square, which expresses how closely the observed data match the model created with the estimated parameters, is next measured. The model is considered predictively relevant if the Q-square value is higher than 0. On the other hand, the model is not predictively relevant if the Q-square value is less than 0 [18]-[20]. The following is the equation used to determine the Q-square:

$$Q^2=1-\left(1-R_1^2\right)\left(1-R_2^2\right)\left(1-R_3^2\right)\ldots\left(1-R_n^2\right)$$
$$Q^2=1-(1-0{,}988)(1-0{,}086)$$
$$Q^2=0{,}989$$

Based on the computations, 0.989 is the Q-square value. When the model's Q-square value is higher than zero, it is considered to have good predictive relevance.

Once the bootstrapping process is completed, the estimated route coefficient value can be used to test the inner model and determine how much the latent component influences the outcome. The sub-discussion on hypothesis testing has a thorough explanation.

3.3. Convert the Path Diagram to an Equation

Conversion of path diagrams into equation form, namely:

Outer model for endogenous latent variables

1. Knowledge = 0,786 X.X1 + 0,802 X.X2 + 0,755 X.X3 + δ_1

   The model above shows that the most dominant indicator is X2 (0.802), meaning that the indicator for the question "I have an adequate understanding of the risks and benefits associated with investing" most strongly reflects the Knowledge construct. The higher the perception of the question indicators on the knowledge variable (X1 to X3), the higher the individual's level of knowledge regarding financial literacy.
2. Confidence = 0,882 X.X4 + 0,859 X.X6 + δ_2

   The above model shows that indicator X4 is the strongest indicator reflecting respondents' confidence in financial management. The confidence construct is very strong as measured by both indicators; respondents tend to have a good level of financial confidence.
3. Skills = 0,738 X.X7 + 0,724 X.X8 + 0,832 X.X9 + δ_3

   The model above shows that Indicator X9 has the highest loading (0.832), indicating that this skill component is the most important in forming Skills. This indicates that the Respondent has

good basic skills in managing or using financial skills (e.g. recording, calculations, or simple analysis).

4. Financial Literacy = 0,564 Knowledge + 0,419 Confidence +0,168 Skills + ε

   This model shows that Knowledge (0.564) has the greatest influence on Financial Literacy, followed by Confidence (0.419), and Skills (0.168). Based on the model, we can see that a person's level of financial literacy is most influenced by financial knowledge and confidence, while practical financial skills contribute less. This indicates that understanding and confidence are more dominant in shaping respondents' financial literacy than technical skills.

Inner model

Interest in Sharia Investment = 0.308 Financial Literacy + ζ

The path coefficient value is 0.308, indicating a positive and moderate relationship between financial literacy and Sharia-compliant investments. This means that the higher a person's financial literacy, the greater their interest in Sharia-compliant investments. Financial literacy increases understanding of the concept of Sharia-compliant investments, reduces uncertainty, and fosters confidence that investments in accordance with Islamic principles can be both financially and spiritually beneficial.

3.4. Hypothesis Testing

The next stage is hypothesis testing, which comes after the measurement model and structural model have been examined and the path diagram has been transformed into a mathematical equation. The values that bootstrapping yields are:

Table 8. Bootstrapping Result

	Original Sample (O)	Sample Mean (M)	Standard Deviation (STDEV)	T Statistics (|O/STDEV|)	P-Values
Skills -> Financial Literacy	0.169	0.171	0.022	7.671	0.000
Confidence -> Financial Literacy	0.420	0.422	0.033	12.759	0.000
Knowledge -> Financial Literacy	0.563	0.559	0.035	15.921	0.000
Financial Literacy -> Interest in Sharia Investment	0.307	0.335	0.093	3.313	0.001

The test results show that the Skills variable has a positive and significant effect on Financial Literacy with an original sample value of 0.169, a T-statistic of 7.671, and a p-value of 0.000. Since the T-statistic is greater than 1.96 and the p-value is less than 0.05, the relationship is declared significant. This indicates that the higher an individual's financial management skills, the better their financial literacy level. In other words, practical skills in managing personal finances, creating budgets, and making appropriate financial decisions significantly contribute to improving a person's financial literacy. The Confidence variable also showed a positive and significant influence on Financial Literacy with an original sample value of 0.420, a T-statistic of 12.759, and a p-value of 0.000. These findings show that a person's degree of financial literacy is significantly influenced by their level of self-confidence. A person's capacity to make wise financial decisions increases with their level of confidence in their ability to comprehend and manage financial concerns. Self-assurance enables people to make more daring financial decisions based on information and logical analysis as opposed to merely relying on gut feeling or outside influences. Additionally, with an original sample value of 0.563, a T-statistic of 15.921, and a p-value of 0.000, the Knowledge variable significantly and favorably affects Financial Literacy. Knowledge is the primary determinant in enhancing financial literacy, as indicated by the greatest coefficient value among these three factors. These findings support the idea that a person's capacity for prudent money management increases with their level of understanding of financial ideas, principles, and products. The key to developing good financial awareness and abilities is knowledge.

The test results also show that Financial Literacy has a positive and significant effect on Interest in Sharia Investment, with an original sample value of 0.307, a T-statistic of 3.313, and a p-value of 0.001. Since the T-statistic value is > 1.96 and the p-value < 0.05, this hypothesis is accepted. This means that the higher a person's level of financial literacy, the greater their interest in investing in Sharia-

based financial instruments. Individuals who understand financial concepts and Sharia principles well will be more interested in choosing investments that align with Islamic values and provide long-term benefits. Based on the results of the bootstrapping test, all hypotheses in this study were accepted as they showed positive and significant effects. In order, the Knowledge variable had the strongest influence on Financial Literacy, followed by Confidence and Skills. Furthermore, Financial Literacy proved to be a significant factor in increasing interest in Sharia Investment. These findings indicate that efforts to improve financial literacy through education, training, and outreach that emphasize knowledge, skills, and confidence will contribute significantly to increasing public interest in Sharia-based investments.

4. Conclusion

The data analysis suggests that interest in modern sharia investing is somewhat influenced by financial literacy, however this effect is considered to be modest. Based on the R^2 analysis, the Financial Literacy construct has an R^2 value of 0.988, meaning 98.8% of the variance in financial literacy is explained by its constituent indicators, namely Knowledge, Confidence, and Skills. This value is considered very high, indicating that the financial literacy measurement model has a very good level of clarity. Conversely, the Interest in Sharia Investment construct has an R^2 value of 0.094, meaning only 9.4% of the variance in Sharia investment interest is explained by financial literacy. Although the relationship between financial literacy and Sharia investment interest is positive and significant, its quantitative contribution is relatively small.

The extreme difference between these two R^2 values indicates that the financial literacy measurement model is very strong, but the structural model between financial literacy and Sharia investment interest is still weak. Therefore, although financial literacy can increase interest in Sharia investment, this influence is not dominant because Sharia investment decisions are also influenced by other factors such as religious values, trust in Sharia institutions, and social influences. Therefore, further model development needs to consider additional constructs as mediating or moderating variables to strengthen the relationship between financial literacy and Sharia investment interest.

This study demonstrates that financial literacy plays an important role in shaping students' interest in Sharia investment. Students with better financial knowledge, higher confidence in financial decision-making, and stronger financial skills tend to show greater openness toward Sharia-compliant investment instruments. This finding confirms that financial literacy remains a relevant factor in encouraging investment interest, even within the context of underdeveloped and frontier regions. However, the findings also suggest that financial literacy alone does not fully determine students' interest in Sharia investment. Investment interest appears to be influenced by a broader set of factors beyond individual financial competence, including institutional trust, accessibility of Sharia financial products, socio-religious considerations, and local contextual conditions. This indicates that interest in Sharia investment is a multifaceted phenomenon that cannot be explained solely through individual-level financial attributes.

Based on these findings, efforts to increase Sharia investment participation should not rely exclusively on financial literacy programs. Instead, financial education initiatives need to be integrated with policies that strengthen Sharia financial institutions, improve product outreach, and enhance public trust, particularly in regions with limited financial infrastructure. Future research is encouraged to incorporate additional contextual and institutional variables to develop a more comprehensive understanding of the determinants of Sharia investment interest.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses primary data obtained from questionnaires distributed to all IAIN Sorong students from July 2023 to August 2023.

Credit Authorship

Selly Nofriska Sudrimo: Conceptualization, Methodology, Data Collection, Data Analysis, Writing–Original Draft. **Urwawuska Ladini**: Manuscript Review, Editing, Writing–Review. **Dahlia Misrika**: Editing, Writing–Review.

References

[1] A. Lusardi and O. S. Mitchell, "The economic importance of financial literacy: Theory and evidence," *Journal of Economic Literature*, vol. 52, no. 1, pp. 5–44, 2014, doi: 10.1257/jel.52.1.5.

[2] A. Yulianto, "Pengaruh literasi keuangan syariah terhadap keputusan penggunaan produk atau layanan lembaga keuangan syariah," *Jurnal Ilmiah Ekonomi Syariah Pascasarjana (JIESP)*, vol. 2, no. 1, pp. 45–56, 2018. [Online]. Available: https://ejournal.kopertais4.or.id/susi/index.php/JIESP/article/download/3634/2470

[3] E. N. Fitrianingsih, "Pengaruh literasi keuangan terhadap keputusan investasi mahasiswa Fakultas Ekonomi dan Bisnis Universitas Muhammadiyah Purwokerto," *Skripsi*, IAIN Purwokerto, 2019. [Online]. Available: https://repository.uinsaizu.ac.id/6289/

[4] M. Widiantika, "Pengaruh persepsi return, motivasi, literasi keuangan, dan pendapatan terhadap minat investasi pada mahasiswa Fakultas Ekonomi Universitas Negeri Jakarta," *Skripsi*, Universitas Negeri Jakarta, 2022. [Online]. Available: https://repository.fe.unj.ac.id/10340/

[5] L. Lutfi, A. Z. Arifin, and A. Syafii, "Sharia financial literacy and investment decisions among students," *Journal of Islamic Economics and Business*, vol. 11, no. 2, pp. 103–118, 2021.

[6] S. Wahyudi and H. Prasetyo, "Sharia-compliant financial literacy and its impact on investment interest," *Journal of Islamic Economics and Finance Studies*, vol. 5, no. 1, pp. 25–36, 2023.

[7] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif*, dan R&D. Bandung: Alfabeta, 2022.

[8] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed. Thousand Oaks, CA: Sage Publications, 2017.

[9] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 3rd ed. Thousand Oaks, CA: Sage Publications, 2022.

[10] J. Henseler, G. Hubona, and P. A. Ray, "Using PLS path modeling in new technology research: Updated guidelines", *Industrial Management & Data Systems*, vol. 116, no. 1, pp. 2–20, 2015, doi: 10.1108/IMDS-09-2015-0382.

[11] E. Babbie, *The Practice of Social Research*, 15th ed. Boston, MA: Cengage Learning, 2020.

[12] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed. Harlow, UK: Pearson Education, 2019.

[13] I. Ghozali, *Konsep dan Aplikasi Structural Equation Modeling berbasis Variance dengan Program SmartPLS 3.0*. Semarang: Badan Penerbit Universitas Diponegoro, 2014.

[14] M. Sarstedt, C. M. Ringle, and J. F. Hair, "Partial least squares structural equation modeling: Guidelines and advances," *European Business Review*, vol. 34, no. 1, pp. 1–25, 2022, doi: 10.1108/EBR-10-2019-0245.

[15] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[16] T. P. Utomo, A. Nugroho, and S. Cahyani, "Analisis validitas dan reliabilitas pada model SEM-PLS," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 21, no. 1, pp. 56–68, 2017, doi: 10.21831/pep.v21i1.15625.

[17] N. Sari, A. Wibowo, and E. Rahmawati, "Pengujian validitas dan reliabilitas model SEM-PLS," *Jurnal Ilmu Sosial dan Humaniora*, vol. 12, no. 2, pp. 210–220, 2023.

[18] I. Ghozali, *Structural Equation Modeling: Metode Alternatif dengan Partial Least Square (PLS)*. Semarang: Badan Penerbit Universitas Diponegoro, 2015.

[19] N. Lajuni, I. Bujang, and Y. Yacob, "Religiosity, financial literacy, and intention to invest: Evidence from Malaysia," *International Journal of Islamic and Middle Eastern Finance and Management*, vol. 13, no. 3, pp. 471–483, 2020, doi: 10.1108/IMEFM-06-2019-0219.

[20] M. Sarstedt, C. M. Ringle, and J. F. Hair, "Partial least squares structural equation modeling: A useful tool for management research," *European Management Journal*, vol. 41, no. 1, pp. 98–110, 2023, doi: 10.1016/j.emj.2022.05.004.

[21] OECD/INFE, *International Survey of Adult Financial Literacy*. Paris: OECD Publishing, 2020.

[22] Otoritas Jasa Keuangan, *Survei Nasional Literasi dan Inklusi Keuangan*. Jakarta: Otoritas Jasa Keuangan, 2022.

[23] Otoritas Jasa Keuangan, *Laporan Satgas PASTI: Penutupan Entitas Keuangan Ilegal*, 2023. [Online]. Available: https://www.ojk.go.id

DOI: https://doi.org/10.34123/jurnalasks.v18i1.909

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Analysis of Factors Influencing Waste Generation in East Java

Syefa Ilmi Beandita Putri[1*], Sri Pingit Wulandari[2]

[1,2]*Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*
**Corresponding Author: E-mail address:* beandita.putri@gmail.com

ARTICLE INFO

Article history:
Received 22 Dec, 2025
Revised 12 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
East Java; Geographically Weighted Regression; SDGs Point 12; Spatial Heterogeneity; Waste Generation

Abstract

Introduction/Main Objectives: Waste accumulation poses a serious threat to environmental sustainability and hinders the achievement of Sustainable Development Goal (SDG) 12 regarding responsible consumption and production patterns. **Background Problems:** East Java consistently ranks second highest in waste generation among Indonesian provinces; this paper investigates the demographic, economic, and environmental determinants of waste generation, specifically addressing the research question of how these factors vary across regencies. **Novelty:** This study extends previous waste generation studies by applying Geographically Weighted Regression (GWR) to the East Java context, initially considering demographic, economic, and environmental variables, and identifying spatial variations in the significant determinants of waste generation. **Research Methods:** Secondary data from 35 regencies/cities in 2023 were analyzed using GWR with a Bisquare Fixed kernel, which was selected as the optimal weighting function compared to Fixed kernels and OLS. **Finding/Results:** Surabaya City recorded the highest waste generation, while the GWR model achieved a goodness-of-fit of 92.72%, higher than the multiple linear regression model. The results confirm that the influence of waste generation determinants is not uniform across regions, indicating significant spatial heterogeneity in East Java.

1. Introduction

Waste-related issues are critical environmental problems that require serious attention from both governments and the public. Along with changing consumption patterns, waste management has become a national issue, as higher consumption levels tend to increase waste volume and diversify its composition. This issue is also closely aligned with Sustainable Development Goal (SDG) 12, which aims to ensure sustainable consumption and production patterns. In 2023, only about 59.74% of waste in Indonesia was managed correctly, equivalent to approximately 20,441,184.59 tons out of a total waste generation of 43,375,225.12 tons [1]. East Java Province, which has the second-largest population in Indonesia at approximately 41.53 million people, is also the province with the second-highest waste generation, amounting to 5,947,865.29 tons in 2023.

These conditions indicate that waste generation in East Java is not only an environmental issue but also a regional planning issue that requires evidence-based, spatially specific policy responses. Global waste management studies emphasize that reliable waste data and local-level analysis are essential for designing effective strategies for waste prevention, collection, treatment, and reduction [2], [3]. Therefore, the benefit of this study is to provide empirical information on how demographic, economic, and environmental factors influence waste generation differently across regencies and cities in East Java. The results can help local governments identify which factors should be prioritized in each area. For

example, regions where population-related factors have a stronger influence may require the strengthening of household waste reduction programs, waste sorting at source, collection services, and community-based 3R facilities. Meanwhile, regions where micro and small industrial activities have a stronger influence may require targeted assistance, technical guidance, and monitoring of business waste management. Thus, the findings of this study are expected to support more targeted waste management policies rather than applying the same policy approach to all regions.

Several previous studies have examined the determinants of municipal solid waste generation using different analytical approaches, study areas, and explanatory variables. Popli et al. [4], analyzed solid waste generation in an urban region of Laos using sociodemographic and economic parameters and showed that these factors are relevant for predicting waste generation rates. Lu et al. [5] also emphasized the importance of data-driven modeling for municipal solid waste generation across different cities, noting that waste generation patterns may vary with urban characteristics and local conditions. More recently, Fontaine et al. [6] developed a framework for predicting solid waste generation by incorporating socioeconomic and demographic factors with municipal waste collection data, confirming that population-related and socioeconomic variables remain important in explaining waste generation.

In the Indonesian context, Putri [7] applied Geographically Weighted Regression (GWR) to model waste generation in Central Java and found that GWR outperformed multiple linear regression by capturing spatial variability across regions. The significant variables in Putri's study were total population, expected years of schooling, and the open unemployment rate. However, Putri's study was limited to Central Java and focused primarily on sociodemographic and socioeconomic variables. In addition, previous studies have generally focused on different location, used global regression or prediction approaches, or failed to examine how the determinants of waste generation vary locally across regencies/cities in East Java. Therefore, the present study differs from Putri's research in terms of study locus, data period, and variable coverage. This study focuses on regencies and cities in East Java in 2023 and initially integrates demographic, economic, and environmental variables, including access to improved sanitation and food management sites. Thus, the novelty of this study lies not only in the use of GWR but also in its application to the East Java context, with broader initial variable coverage, and in identifying local variations in the determinants of waste generation across regencies and cities.

Geographically Weighted Regression (GWR) is an extension of linear regression that produces location-specific parameter estimates for each observation, allowing the model to reflect spatially varying relationships between waste generation and its influencing factors at the regency and city levels [8]. In spatial analysis, relationships can be represented through area-based or point-based approaches. In this study, spatial relationships are area-based, as the data structure and interpretation are based on the regency/city administrative area as the unit of analysis. Waste generation, demographic, economic, and environmental variables are aggregated for each regency/city, so the findings are interpreted as regional characteristics rather than individual-level behavior. However, the GWR estimation process requires spatial coordinates to measure geographical proximity between observations. Therefore, each regency/city is represented by its latitude and longitude as a representative point of the administrative area. These coordinates are used to calculate Euclidean distances and construct the spatial weighting matrix in the GWR model [9].

The use of point-based coordinates is considered appropriate in this study because the main objective is to examine spatial heterogeneity, namely, whether the influence of demographic, economic, and environmental factors on waste generation differs across locations. A point-based distance-weighting approach assigns larger weights to nearby regencies/cities than to more distant regions, consistent with the local-estimation principle of GWR [8]. In contrast, an area-based contiguity approach is better suited to models that explicitly model spatial dependence between neighboring polygon areas. Since this study aims to estimate location-specific regression parameters rather than model spatial autocorrelation using an adjacency matrix, representative point coordinates serve as the basis for the GWR weighting scheme. Nevertheless, the area-based nature of the data remains central to interpretation, as all results are reported and discussed at the regency/city level.

2. Material and Methods

The target population of this study comprises all 38 regencies/cities in East Java Province, a region identified as one of the largest contributors to waste in Indonesia. The research context focuses on environmental sustainability, specifically on modeling waste-generation patterns to support Sustainable Development Goal (SDG) 12 on responsible consumption and production. The unit of analysis is the administrative area at the regency/city level. However, the final sample used in this study consists of 35 regencies/cities because waste generation data for three regions, namely Bondowoso Regency,

Probolinggo Regency, and Pasuruan Regency, were not available in the SIPSN database for 2023. Therefore, these three regions were excluded from the analysis due to data unavailability, as they did not follow the number of regencies/cities used in previous research on Central Java. Thus, the use of 35 regencies/cities in this study was determined solely by the completeness of the 2023 waste generation data.

Data collection processes were conducted correctly using secondary data sources. The data regarding waste generation was retrieved from the National Waste Management Information System (SIPSN) under the Ministry of Environment and Forestry. Meanwhile, demographic and economic data were obtained from Statistics Indonesia (BPS) official publications for East Java Province, and environmental health data were sourced from the East Java Provincial Health Office. The temporal scope of the data is the year 2023, selected to ensure the most recent and comprehensive representation of the province.

2.1 Preparations

The preparations for this research included collecting secondary data from 35 regencies/cities in East Java Province in 2023. The data were obtained from the official website of Statistics Indonesia (BPS) and the National Waste Management Information System (SIPSN). In this study, p represents the number of independent variables, n indicates the number of observations (35 regencies/cities), Y_i is the value of the dependent variable (waste generation) on i-th observation, β_0 is the constant, β_j represents the regression coefficient of independent variable X_j, X_{ij} indicates the value of j-th independent variable on i-th observation. The independent variables used are Total Population (X_{i1}) measured in persons, Expected Years of Schooling (X_{i2}) measured in years, Open Unemployment Rate (X_{i3}) measured in percent, Number of Micro and Small Industrial Enterprises (X_{i4}) measured in units/enterprises, Percentage of Households with Access to Proper Sanitation (X_{i5}) measured in percent, and Percentage of Food Management Sites (X_{i6}) measured in percent.

2.2 Performing Multicollinearity Checking

Multicollinearity is a condition in which independent variables are highly correlated. The multicollinearity test identifies multicollinearity using the Variance Inflation Factor (VIF) criterion and correlation coefficient value. If the VIF value is greater than 10, then multicollinearity occurs [10]. The VIF value can be obtained using Formula (1):

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1}$$

$$R_j^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where VIF_j denotes the Variance Inflation Factor for the j-th independent variable, and R_j^2 denotes the coefficient of determination obtained from regressing the j-th independent variable on all other independent variables in the model. In the calculation of R_j^2, SSR represents the regression sum of squares, SST represents the total sum of squares, $\hat{y}$ is the predicted value for the i-th observation, y_i is the observed value for the i-th observation, $\bar{y}$ is the mean value of the observed data, and n denotes the number of observations. A higher R_j^2 value indicates that the j-th independent variable can be strongly explained by the other independent variables, which may lead to a higher VIF_j value and indicate potential multicollinearity. Furthermore, the Pearson correlation coefficient measures the linear relationship between two variables; the higher the absolute value of the coefficient, the stronger the relationship [11]. The calculation of the correlation coefficient value is explained in Formula (2).

$$r_{x_a,x_b} = \frac{n\sum_{i=1}^{n} x_{ai}x_{bi} - \left(\sum_{i=1}^{n} x_{ai}\right)\left(\sum_{i=1}^{n} x_{bi}\right)}{\sqrt{n\sum_{i=1}^{n} x_{ai}^2 - \left(\sum_{i=1}^{n} x_{ai}\right)^2}\sqrt{n\sum_{i=1}^{n} x_{bi}^2 - \left(\sum_{i=1}^{n} x_{bi}\right)^2}} \tag{2}$$

Where r_{x_a,x_b} denotes the Pearson correlation coefficient between the a-th independent variable x_a and the b-th independent variable x_b. The notation x_{ai} represents the observed value of variable x_a in the i-th observation, while x_{bi} represents the observed value of variable x_b in the i-th observation. The symbol n denotes the total number of observations. The value of the correlation coefficient ranges from

-1 to 1. A value close to 1 indicates a strong positive linear relationship, a value close to -1 indicates a strong negative linear relationship, and a value close to 0 indicates a weak linear relationship [12].

2.3 Multiple Linear Regression Modeling

The initial stage involves modeling the relationship between waste generation and the predictor variables using Multiple Linear Regression (Ordinary Least Squares - OLS) to obtain a global model [10]. The mathematical equation for the OLS model is expressed as follows:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i \quad (3)$$

Where y_i denotes the value of the dependent variable for the *i*-th observation, namely waste generation; β_0 denotes the intercept or constant term β_j denotes the regression coefficient of the *j*-th independent variable; X_{ij} denotes the value of the *j*-th independent variable for the *i*-th observation; ε_i denotes the error term for the *i*-th observation; (i = 1, 2, ..., n) denotes the observation index; and (j = 1, 2, ..., k) denotes the independent variable index. Furthermore, n denotes the number of observations, and k denotes the number of independent variables in the model.

The method used to estimate the parameters is Ordinary Least Squares (OLS). Parameter estimates are stated in Formula 4.

$$\hat{\beta} = (X^T X)^{-1}(X^T y) \quad (4)$$

The significance of the global parameters is evaluated using the F-test for simultaneous influence and the t-test for partial influence. The IIDN residual test is a requirement in multiple linear regression analysis. The IIDN residual test is conducted to determine whether the residuals of the regression model meet the requirements of being identical, independent, and normally distributed [13].

2.4 Performing Spatial Heterogeneity Test

The spatial heterogeneity test was conducted using the Breusch–Pagan test. This test examines whether the variance of the residuals from the global regression model is constant across observations [14]. Spatial heterogeneity means that the relationship between waste generation and its influencing factors is not uniform across regencies/cities [15]. Therefore, if spatial heterogeneity is detected, using Geographically Weighted Regression (GWR) is appropriate, as it allows regression parameters to vary locally at each observation location. The Breusch-Pagan test formula is shown in Formula (5).

$$BP = \frac{1}{2}\boldsymbol{f}^T \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f} \quad (5)$$

Where BP denotes the Breusch Pagan test statistic, $\boldsymbol{f}$ denotes an n × 1 vector whose *i*-th element is f_i, $\boldsymbol{Z}$ denotes the n × p matrix of standardized independent variables, $\boldsymbol{Z}^T$ denotes the transpose of matrix $\boldsymbol{Z}$ and $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$ denotes the inverse matrix of $\boldsymbol{Z}^T\boldsymbol{Z}$. Furthermore, f_i is shown in Formula (6).

$$f_i = \left(\frac{\varepsilon_i^2}{\sigma^2} - 1\right) \quad (6)$$

Where ε_i denotes the residual value of the i-th observation, σ^2 denotes the residual variance, n denotes the number of observations, and p denotes the number of parameters in the model. The null hypothesis of the Breusch-Pagan test states that there is no spatial heterogeneity, while the alternative hypothesis states that there is spatial heterogeneity across observation locations.

2.5 Geographically Weighted Regression (GWR) Modeling

GWR enhances the global model by allowing regression parameters to vary across locations, capturing local variations [8]. The GWR model Equation is expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{k} \beta_{ij}(u_i, v_i) X_{ij} + \varepsilon_i \quad (7)$$

Where y_i denotes the value of the dependent variable for the *i*-th observation, namely waste generation; $\beta_0(u_i, v_i)$ denotes the intercept parameter at the location of the *i*-th observation; $\beta_{ij}(u_i, v_i)$ denotes the local regression coefficient of the *j*-th independent variable at the location of the *i*-th observation; X_{ij} denotes the value of the *j*-th independent variable for the *i*-th observation; ε_i denotes the error term for the *i*-th observation; u_i denotes the latitude coordinate of the *i*-th observation; v_i denotes the longitude coordinate of the *i*-th observation; i = 1, 2, …, n denotes the observation index; and j = 1, 2, …, k denotes the independent variable index. Furthermore, n represents the number of observations, while k represents the number of independent variables used in the model.

The spatial weighting function in GWR plays an important role because it represents the spatial proximity among observation locations. The weighting function is determined by first calculating the Euclidean distance between locations. The Euclidean distance d_{il} denotes the distance between the *i*-th and *l*-th observation locations, which is determined using latitude u_i and longitude v_i coordinates [16]. The Euclidean distance is calculated using Formula 8.

$$d_{il} = \sqrt{(u_i - u_l)^2 + (v_i - v_l)^2} \quad (8)$$

Where d_{il} denotes the Euclidean distance between the *i*-th observation location and the *l*-th observation location. The symbol u_i denotes the latitude coordinate of the *i*-th observation, while u_l denotes the latitude coordinate of the *l*-th observation. Furthermore, v_i denotes the longitude coordinate of the *i*-th observation, while v_l denotes the longitude coordinate of the *l*-th observation. The indices i and l represent different regency/city observation locations. A smaller d_{il} value indicates that two regions are geographically closer, while a larger d_{il} value indicates that two regions are farther apart [17].

The GWR spatial weighting function can be calculated using kernel functions. Kernel functions are commonly used in GWR to assign different weights to observations based on their geographical proximity to the target location. These kernel functions can be specified using either fixed or adaptive bandwidths [8], [18]. Fixed kernel functions use the same bandwidth for all observation locations, while adaptive kernel functions allow the bandwidth to vary according to the spatial distribution of the observations. The kernel functions considered in this study include Fixed Gaussian, Fixed Bisquare, Fixed Tricube, Adaptive Gaussian, Adaptive Bisquare, and Adaptive Tricube. The weighting function chosen for this analysis is the Fixed Bisquare Kernel, which was selected based on its lowest AIC and CV values compared to other kernels. The h value is a non-negative parameter known as bandwidth. Bandwidth determines the extent of spatial influence around each target location and is used to assign weights to other observation locations in the GWR model [8], [17]. The method for determining the bandwidth value is using Cross Validation, as described in Formula 9 [9].

$$CV(h) = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(h)]^2 \quad (9)$$

To estimate the parameters, a spatial weighting matrix is required. This study utilizes the Euclidean distance metric to measure the proximity between locations *i* and *j*.

The GWR parameter model estimation used is the Weighted Least Square (WLS) method by providing different weightings for each observation location [8]. The GWR parameter estimation is explained by the Formula in 10.

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i) \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}(u_i, v_i) \boldsymbol{y} \quad (10)$$

Finally, the model's performance is evaluated and compared against the global OLS model using the Coefficient of Determination (R^2) and Akaike Information Criterion (AIC). A higher R^2 and a lower AIC indicate a better-fitting model [19]. To provide a clearer overview of the analytical procedure used in this study, the steps of the Geographically Weighted Regression analysis are summarized in the flow chart shown in Figure 1.

3. Results and Discussion

The characteristics of waste generation data and its influencing factors based on descriptive statistics are presented in Table 1.

Table 1. Descriptive statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
Waste Generation (Y_i)	169,939.01	123,696.13	27,988.20	657,016.64
Total Population (X_{i1})	1,081,209	699,795	135,414	2,893,698
Expected Years of Schooling (X_{i2})	13.57	0.90	11.97	15.77
Open Unemployment Rate (X_{i3})	4.70	1.46	1.71	8.05
Number of Micro and Small Industrial Enterprises (X_{i4})	22,251	15,155	2,009	67,609
Percentage of Households with Access to Improved Sanitation (X_{i5})	86.02	11.23	50.30	98.18

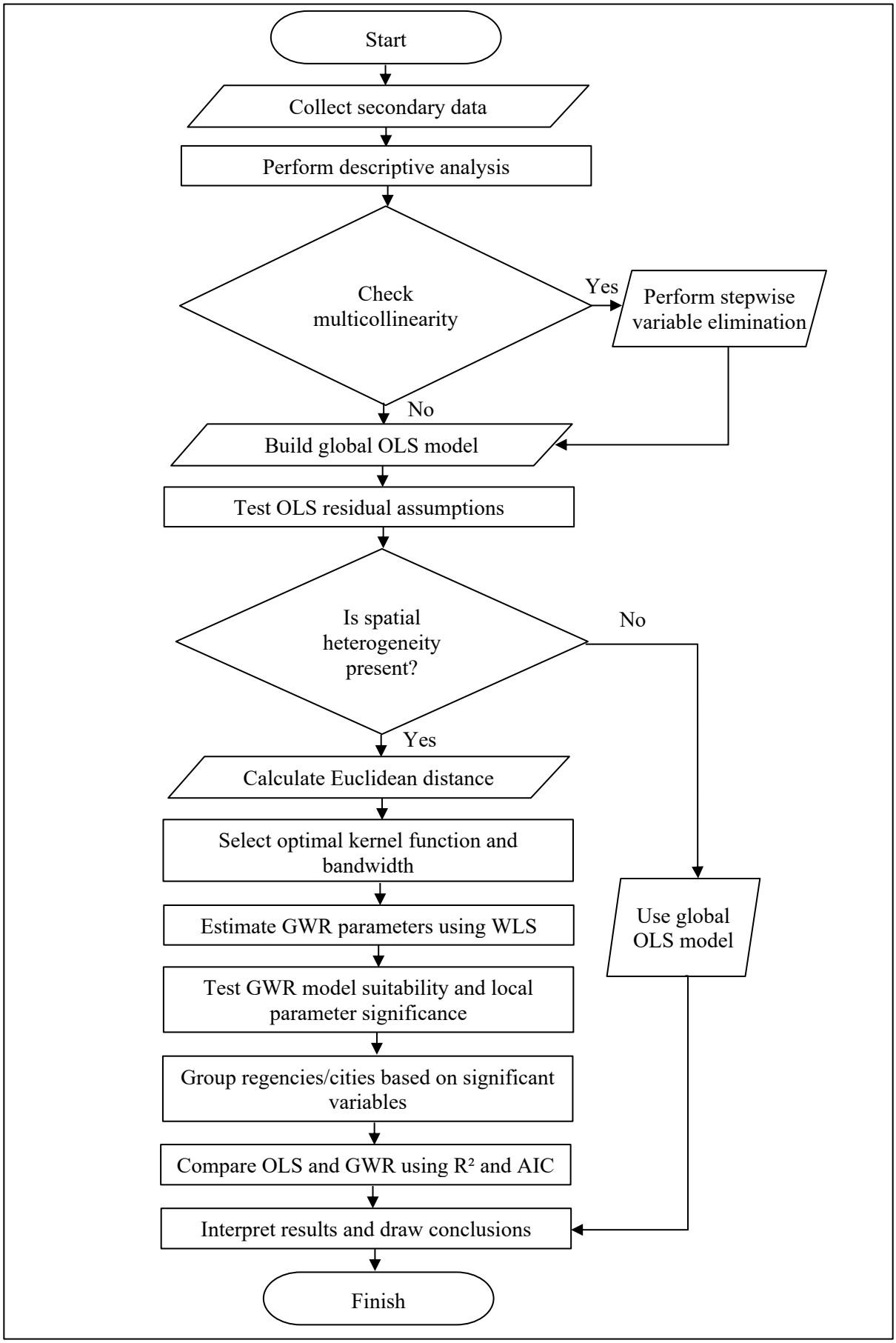


Figure 1. Flow chart

Table 1 shows that annual waste generation in East Java in 2023 averaged 169,939.01 tons per year, with a standard deviation of 123,696.13 tons per year, indicating substantial variation across regencies and cities. Blitar Regency records the lowest waste generation, while Surabaya City has the highest. The population variable also exhibits considerable regional disparity, with an average of 1,081,209 people and a standard deviation of 699,795, primarily driven by Surabaya City, the most populous area.

Other socioeconomic factors, including expected years of schooling, the open unemployment rate, the number of micro and small industrial enterprises, access to improved sanitation, and the percentage of food management facilities, display notable variation among regencies and cities. This variation reflects differences in educational attainment, labor market conditions, economic activity, and basic infrastructure across East Java. The spatial characteristics of waste generation are further illustrated through map visualizations presented in Figure 1.

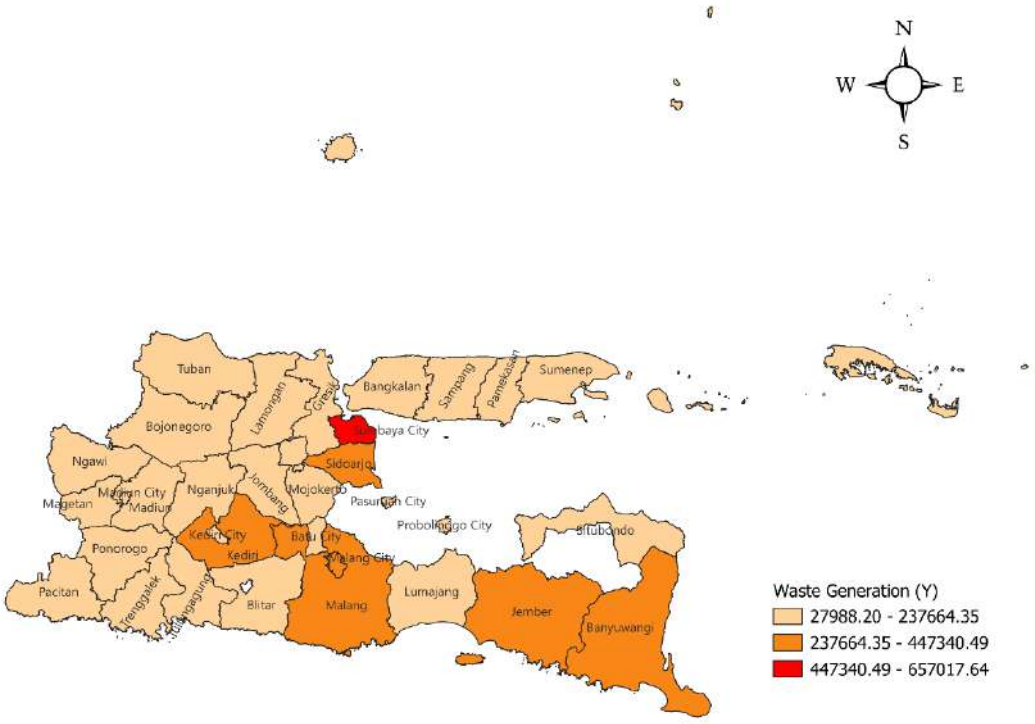


Figure 2. Mapping of waste generation in East Java Province

3.1 Multicollinearity Checking

Multicollinearity was examined using correlation coefficients and variance inflation factors (VIFs). The assessment of multicollinearity based on correlation coefficients is presented in Table 2.

Table 2. Correlation matrix

	Y	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}
X_{i1}	0,908					
	0,000					
X_{i2}	0,128	-0,066				
	0,462	0,708				
X_{i3}	0,375	0,297	**0,418**			
	0,026	0,083	**0,012**			
X_{i4}	0,237	**0,440**	**-0,378**	**-0,433**		
	0,170	**0,008**	**0,025**	**0,009**		
X_{i5}	-0,021	-0,133	**0,518**	**0,406**	**-0,521**	
	0,905	0,445	**0,001**	**0,015**	**0,001**	
X_{i6}	0,029	-0,041	**0,311**	**0,325**	-0,219	**0,412**
	0,867	0,813	**0,069**	**0,057**	0,207	**0,014**

Table 2 indicates that several pairs of predictor variables exhibit statistically significant correlations (p-values < 0.10). Furthermore, the results of the Variance Inflation Factor (VIF) analysis show that all

predictor variables have VIF values below 10. Although the VIF values do not exceed the commonly used threshold, the presence of significant correlations among predictor variables suggests potential multicollinearity. Therefore, variable elimination was performed using the stepwise method, resulting in the final set of predictor variables consisting of X_{i1}, X_{i2}, and X_{i4}.

3.2 Multiple Linear Regression Modeling

The parameter estimates of the multiple linear regression model for factors influencing waste generation are expressed as follows:

$$\hat{Y} = -251554 + 0.17X_1 + 19157X_2 - 1.16X_4$$

The estimated multiple linear regression model indicates that, holding other variables constant, a one-person increase in total population (X_1) in East Java Province is associated with an increase in waste generation of 0.17 tons/year. Furthermore, a one-year increase in expected years of schooling (X_2) is associated with an increase in waste generation of 19,157 tons/year, assuming other variables remain constant. Meanwhile, a one-unit increase in the number of micro and small industrial enterprises (X_4) is associated with a 1.16-ton/year decrease in waste generation, ceteris paribus.

The negative coefficient of the number of micro and small industrial enterprises (X_4) is not fully consistent with the theoretical expectation that industrial activity tends to increase waste generation. This result may be related to differences in the characteristics of micro and small enterprises across regions, including the implementation of waste minimization, reuse of production materials, recycling, or practices oriented towards a circular economy. Circular economy approaches emphasize reducing waste, keeping materials in use, and reusing or recycling resources, so an increase in the number of business units does not always directly lead to higher residual waste generation [20]. Therefore, the negative sign of (X_4) should be interpreted cautiously as an indication that the relationship between micro and small industrial activity and waste generation may vary across regencies/cities. This condition supports the need for spatially local analysis using GWR.

Subsequently, a simultaneous significance test was conducted. Based on the analysis, the F-test statistic value of 71.73 exceeds the critical value $F_{(0.1;3;31)}$ is 2.27 and is supported by a p-value of 0.00, which is smaller than the significance level α is 0.1. Therefore, the null hypothesis is rejected, indicating that at least one predictor variable has a statistically significant effect on waste generation. Furthermore, the results of the partial significance tests are presented in Table 3. The t-test results indicate that variables X_{i1}, X_{i2}, and X_{i4} have statistically significant effects on waste generation.

Table 3. Partial significance test statistics

Variable	t-statistic	$t_{(0,1/2;28)}$	P-Value
X_{i1}	13.70		0.00
X_{i2}	2.01	1.70	0.05
X_{i4}	1.84		0.08

Subsequently, the results of the identical residuals test show that the test statistic value of 22.74 exceeds the critical value $F_{(0.1;1;33)}$ is 2.86 and is supported by a p-value of 0.00, which is smaller than the significance level α is 0.1. Therefore, the null hypothesis is rejected, indicating that the residuals of the regression model are not identical. This result suggests the presence of spatial heterogeneity due to differences in characteristics across regencies and cities.

The independence of residuals test yields a Durbin–Watson statistic of 2.17, which is smaller than 4 − dU is 2.35 and greater than dU is 1.65. Thus, the null hypothesis is accepted, indicating that the residuals of the regression model are independent.

The normality test of residuals produces a test statistic value of 0.13, which is smaller than the critical value $KS_{(0.1;35)}$ is 0.20 and is supported by a p-value of 0.56, which is greater than the significance level α is 0.1. Therefore, the null hypothesis is accepted, indicating that the residuals of the regression model are normally distributed.

3.3 Performing Spatial Heterogeneity Test

The spatial heterogeneity test using the Breusch–Pagan test yields a test statistic value of 16.79, which exceeds the critical value of 6.25 and is supported by a p-value of 0.00, which is smaller than the significance level α is 0.1. Therefore, the null hypothesis is rejected, indicating the presence of spatial

heterogeneity. Consequently, the analysis can be extended using the Geographically Weighted Regression (GWR) approach.

3.4 Geographically Weighted Regression (GWR) Modeling

In the GWR analysis, the weighting values were determined using kernel functions, as presented in Table 4. Table 4 indicates that the Fixed Bisquare kernel function was selected for the GWR analysis, as it provides the best overall model performance based on the evaluation criteria.

Table 4. Comparison of optimal Kernel Functions

Kernel Function	CV	R^2	AIC
Fixed Gaussian	110489182025	0,88	850,28
Adaptive Gaussian	110964903827	0,88	849,67
Fixed Bisquare	120870672140	0,93	836,76
Adaptive Bisquare	121002609225	0,90	845,95
Fixed Tricube	119746559578	0,92	839,05
Adaptive Tricube	133953245074	0,90	844,33

After selecting the Fixed Bisquare kernel function, the factors influencing waste generation were modeled for the first observation location (u_1, v_1), namely Pacitan Regency, which is used as an example for interpreting the modeling results. The waste generation model for Pacitan Regency is expressed as follows:

$$\hat{Y} = -548164.14 + 0.21X_1 + 36448.47X_2 + 0.74X_4$$

This model indicates that, holding other variables constant, a one-person increase in population results in an increase of waste generation by 0.21 tons per year. Furthermore, a one-year increase in expected years of schooling, ceteris paribus, leads to an increase in waste generation by 36,448.47 tons per year. In addition, an increase of one percent in the number of micro and small industrial enterprises, while other variables remain constant, increases waste generation by 0.74 tons per year. An increase in expected years of schooling, accompanied by higher waste generation, indicates that improvements in educational attainment in Pacitan Regency have not yet been fully aligned with increased awareness and behavioral changes in waste management [21], [22]. In line with this, population growth reflects rising household consumption, which directly increases waste generation [4], [23]. In addition to demographic factors, the growth in the number of micro and small industrial enterprises suggests that business activities also contribute to increased waste generation when not supported by responsible waste management practices [24]. These conditions indicate that both demographic and economic factors are important determinants of waste generation in Pacitan Regency. Therefore, efforts in education and outreach are needed for communities and micro- and small-business actors on waste reduction, waste segregation, and sustainable waste management, particularly among groups with increasing consumption and production activities, alongside local educational and economic development.

The results of the GWR model goodness-of-fit test show that the test statistic value of 1.73 exceeds the critical value $F_{(0.1;31;23,34)}$ is 1.68 and is supported by a p-value of 0.08, which is smaller than the significance level α is 0.1. Therefore, the null hypothesis is rejected, indicating a statistically significant difference between the Ordinary Least Squares (OLS) and the GWR models. This result confirms that the GWR approach provides a better representation of the data's spatial variation.

Following the confirmation of model suitability, partial significance tests of the GWR model parameters were conducted using a significance level of α is 0.1 and degrees of freedom of 23.34, with a critical value of $t_{(0.1;23.34)}$ is 1.71 The results indicate that the total population variable (X_{i1}), is statistically significant in 35 regencies and cities, expected years of schooling (X_{i2}), and is significant in 26 regencies and cities, and the number of micro and small industrial enterprises (X_{i4}) is significant in 11 regencies and cities. These findings highlight the spatially varying influence of explanatory variables on waste generation across East Java, as presented in Table 5.

Table 5. Partial significance test of GWR model parameters

No.	Regency/City	$\lvert t_{X_{i1}} \rvert$	$\lvert t_{X_{i2}} \rvert$	$\lvert t_{X_{i4}} \rvert$	$t_{(\frac{0,1}{2};23.34)}$
1.	Pacitan Regency	**12.87**	**2.09**	0.60	1.71
2.	Ponorogo Regency	**13.91**	**2.39**	0.58	1.71
3.	Trenggalek Regency	**13.74**	**2.65**	0.59	1.71
4.	Tulunganggung Regency	**13.78**	**2.68**	0.98	1.71
5.	Blitar Regency	**13.03**	**2.56**	1.58	1.71
6.	Kediri Regency	**14.04**	**2.44**	**1.72**	1.71
7.	Malang Regency	**12.24**	**2.47**	1.71	1.71
8.	Lumajang Regency	**10.15**	**1.93**	1.10	1.71
9.	Jember Regency	**9.24**	1.29	1.53	1.71
10.	Banyuwangi Regency	**6.19**	0.78	0.99	1.71
11.	Situbondo Regency	**7.26**	0.73	1.02	1.71
12.	Sidoarjo Regency	**12.27**	**2.16**	**1.95**	1.71
13.	Mojokerto Regency	**13.72**	**2.28**	**2.36**	1.71
14.	Jombang Regency	**14.28**	**2.30**	**2.31**	1.71
15.	Nganjuk Regency	**14.30**	**2.24**	**1.82**	1.71
16.	Madiun Regency	**14.19**	**2.28**	1.06	1.71
17.	Magetan Regency	**13.88**	**2.15**	0.73	1.71
18.	Ngawi Regency	**14.21**	**1.96**	1.43	1.71
19.	Bojonegoro Regency	**14.16**	1.61	**2.31**	1.71
20.	Tuban Regency	**13.66**	1.22	**2.51**	1.71
21.	Lamongan Regency	**13.35**	**1.84**	**2.38**	1.71
22.	Gresik Regency	**12.53**	**1.94**	**2.13**	1.71
23.	Bangkalan Regency	**10.43**	1.46	1.69	1.71
24.	Sampang Regency	**9.33**	1.19	1.19	1.71
25.	Pamekasan Regency	**9.31**	0.75	1.15	1.71
26.	Sumenep Regency	**8.20**	0.59	0.83	1.71
27.	Kediri City	**14.05**	**2.46**	1.63	1.71
28.	Blitar City	**13.31**	**2.62**	1.33	1.71
29.	Malang City	**12.24**	**2.47**	1.71	1.71
30.	Probolinggo City	**10.06**	**1.79**	1.17	1.71
31.	Pasuruan City	**10.79**	**2.04**	1.42	1.71
32.	Mojokerto City	**13.71**	**2.28**	**2.36**	1.71
33.	Madiun City	**14.24**	**2.23**	1.16	1.71
34.	Surabaya City	**14.23**	**2.23**	1.15	1.71
35.	Batu City	**13.11**	**2.50**	**2.00**	1.71

Note: Values shown in bold indicate statistically significant parameters.

Based on the results of the analysis, a summary of the significant variables for each regency and city in East Java is presented in Table 6.

Table 6. Significant variables for each regency/city

No.	Regency/ City	Variables	Significant Factors	No.	Regency/ City	Variables	Significant Factors
1	Pacitan Regency	X_1, X_2	Demographic	19	Bojonegoro Regency	X_1, X_4	Demographic, Economic
2	Ponorogo Regency	X_1, X_2	Demographic	20	Tuban Regency	X_1, X_4	Demographic, Economic
3	Trenggalek Regency	X_1, X_2	Demographic	21	Lamongan Regency	X_1, X_2, X_4	Demographic, Economic
4	Tulunganggung Regency	X_1, X_2	Demographic	22	Gresik Regency	X_1, X_2, X_4	Demographic, Economic
5	Blitar Regency	X_1, X_2	Demographic	23	Bangkalan Regency	X_1	Demographic
6	Kediri Regency	X_1, X_2, X_4	Demographic, Economic	24	Sampang Regency	X_1	Demographic
7	Malang Regency	X_1, X_2	Demographic	25	Pamekasan Regency	X_1	Demographic
8	Lumajang Regency	X_1, X_2	Demographic	26	Sumenep Regency	X_1	Demographic
9	Jember Regency	X_1	Demographic	27	Kediri City	X_1, X_2	Demographic
10	Banyuwangi Regency	X_1	Demographic	28	Blitar City	X_1, X_2	Demographic
11	Situbondo Regency	X_1	Demographic	29	Malang City	X_1, X_2	Demographic
12	Sidoarjo Regency	X_1, X_2, X_4	Demographic, Economic	30	Probolinggo City	X_1, X_2	Demographic
13	Mojokerto Regency	X_1, X_2, X_4	Demographic, Economic	31	Pasuruan City	X_1, X_2	Demographic
14	Jombang Regency	X_1, X_2, X_4	Demographic, Economic	32	Mojokerto City	X_1, X_2, X_4	Demographic, Economic
15	Nganjuk Regency	X_1, X_2, X_4	Demographic, Economic	33	Madiun City	X_1, X_2	Demographic
16	Madiun Regency	X_1, X_2	Demographic	34	Surabaya City	X_1, X_2	Demographic
17	Magetan Regency	X_1, X_2	Demographic	35	Batu City	X_1, X_2, X_4	Demographic, Economic
18	Ngawi Regency	X_1, X_2	Demographic				

Table 6 shows that the variables significantly influencing waste generation vary across regencies and cities. Regencies and cities were further grouped based on similarities in the significant variables affecting waste generation, as presented in Table 7.

Table 7 summarizes the grouping of regencies and cities based on the variables that significantly influence waste generation. Grouping is important because it shows that the determinants of waste generation are not uniform across East Java. Each group reflects a different local mechanism in which demographic and economic factors influence waste generation. Therefore, the same waste management policy may not be equally effective for all regencies/cities.

Group 1 consists of Jember Regency, Banyuwangi Regency, Situbondo Regency, Bangkalan Regency, Sampang Regency, Pamekasan Regency, and Sumenep Regency, in which only the total population (X_1) is significant. This indicates that waste generation in these regions is primarily driven by population size. In this group, the effects of expected years of schooling (X_2) and the number of micro and small industrial enterprises (X_4) are not statistically significant, suggesting that household and population-related waste generation is more dominant than education-related consumption patterns or small industrial activities. Therefore, waste management strategies in this group should prioritize

household waste reduction, source-based waste sorting, improvements in waste collection services, and community-based waste management programs.

Table 7. Grouping of regencies/cities based on significant variables

Group	Regency/City	Variables	Number of Regencies/Cities	Significant Factors
1	Jember Regency, Banyuwangi Regency, Situbondo Regency, Bangkalan Regency, Sampang Regency, Pamekasan Regency, Sumenep Regency	X_1	7	Demographic
2	Pacitan Regency, Ponorogo Regency, Trenggalek Regency, Tulunganggung Regency, Blitar Regency, Malang Regency, Lumajang Regency, Madiun Regency, Magetan Regency, Ngawi Regency, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Madiun City, Surabaya City	X_1, X_2	17	Demographic
3	Bojonegoro Regency, Tuban Regency	X_1, X_4	2	Demographic, Economic
4	Kediri Regency, Sidoarjo Regency, Mojokerto Regency, Jombang Regency, Nganjuk Regency, Lamongan Regency, Gresik Regency, Mojokerto City, Batu City	X_1, X_2, X_4	9	Demographic, Economic

Group 2 consists of regions where total population (X_1) and expected years of schooling (X_2) are significant. This group includes several regencies and cities such as Pacitan, Ponorogo, Trenggalek, Tulungagung, Blitar, Malang, Lumajang, Madiun, Magetan, Ngawi, Kediri City, Blitar City, Malang City, Probolinggo City, Pasuruan City, Madiun City, and Surabaya City. The significance of X_2 indicates that waste generation in these regions is not only related to the number of residents but also to social characteristics associated with education. Higher expected years of schooling may reflect changes in lifestyle, consumption behavior, access to goods and services, and public activities, which can increase the volume and diversity of waste generated. Therefore, this group requires not only population-based waste management but also environmental education programs, school-based waste literacy, and campaigns to strengthen responsible consumption behavior.

Group 3 consists of Bojonegoro Regency and Tuban Regency, where the total population (X_1) and the number of micro and small industrial enterprises (X_4) are significant. This pattern indicates that both household activities and local economic activities influence waste generation in these regions. The significance of X_4 suggests that micro and small enterprises contribute to waste generation through production, packaging, distribution, and trade activities. Therefore, waste management policies in this group should combine household waste management with specific assistance for micro and small enterprises, such as training on waste reduction, recyclable packaging, separation of production waste, and monitoring of business-related waste disposal.

Group 4 consists of Kediri Regency, Sidoarjo Regency, Mojokerto Regency, Jombang Regency, Nganjuk Regency, Lamongan Regency, Gresik Regency, Mojokerto City, and Batu City, where the total population (X_1), expected years of schooling (X_2), and the number of micro and small industrial enterprises (X_4) are all significant. This group indicates a more complex pattern of waste generation because demographic, social, and economic factors simultaneously influence it. The regions in this group require more integrated waste management policies, including household waste services, education-based behavioral change programs, and targeted waste management assistance for micro and small enterprises. The significance of the three variables in this group indicates that waste generation is not only a population issue but is also closely related to human development and local economic activity.

The spatial distribution of regencies and cities in East Java according to these groupings is illustrated in Figure 3.

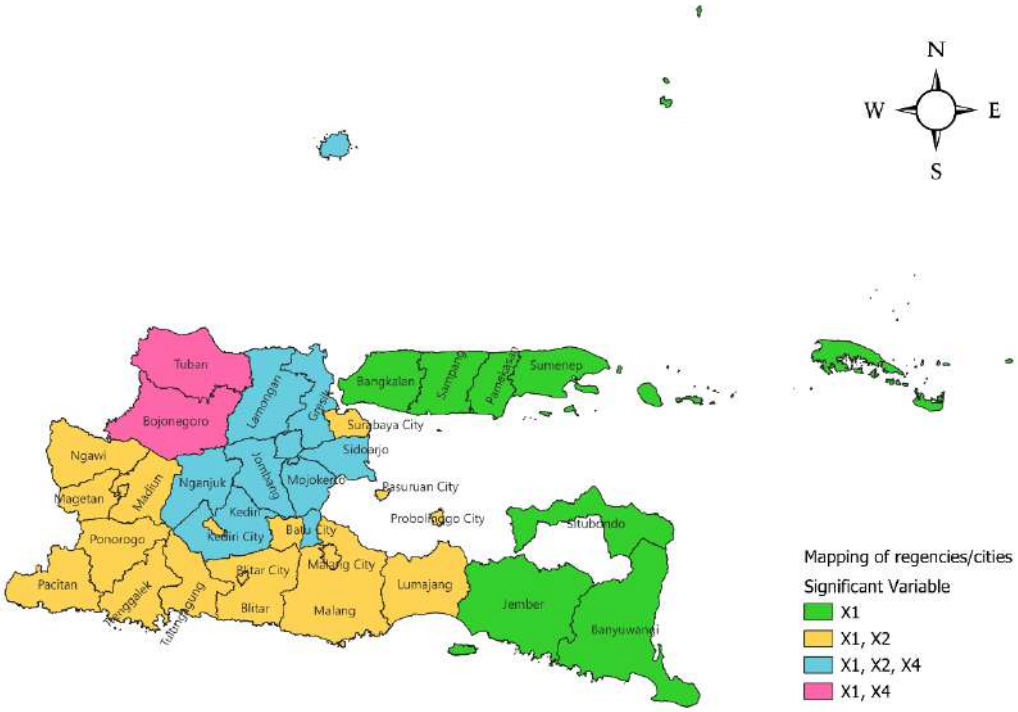


Figure 3. Mapping of regencies/cities based on significant variables

The selection of the best model was conducted by comparing multiple linear regression and Geographically Weighted Regression (GWR) using the coefficient of determination (R^2) and the Akaike Information Criterion (AIC), as presented in Table 8.

Table 8. Best model selection

Method	R2	AIC
Multiple linear regression	86.19%	856.57
GWR	92.72%	836.76

Table 8 shows that the GWR model has a higher R^2 and a lower AIC than the multiple linear regression model, indicating superior model performance. This result suggests that GWR better captures spatial heterogeneity across regions, whereas multiple linear regression applies a single global Equation to all regions.

4. Conclusion

Based on the results of the Geographically Weighted Regression model with a fixed bisquare kernel, the influence of demographic and economic factors on waste generation varies across East Java's regencies and cities, indicating spatial heterogeneity. The GWR model captures this spatial variation and achieves a goodness-of-fit of 92.72%, making it more representative than the multiple linear regression model. The results show that total population (X_1), expected years of schooling (X_2), and the number of micro and small industrial enterprises (X_4) have different levels of significance across regions. Meanwhile, environmental variables initially considered in the early stage of the analysis were not included in the final model because they were eliminated during variable selection due to multicollinearity. For local governments, these findings indicate that waste management policies should be tailored to regional characteristics. Regions where population is the dominant factor should prioritize household waste reduction, waste sorting at the source, and improvement of waste collection services, while regions where micro and small industrial enterprises also have a significant effect should strengthen outreach, technical assistance, and monitoring of business-related waste management.

This study has several limitations. First, the analysis covers only 35 regencies/cities in East Java because waste generation data for Bondowoso Regency, Probolinggo Regency, and Pasuruan Regency were unavailable in the SIPSN database for 2023. Therefore, the results do not fully represent all

administrative regions in East Java. Second, this study uses cross-sectional data from a single year, so it cannot capture changes in waste generation patterns over time. Third, the analysis is limited to variables available from secondary data sources, and several potentially relevant factors, such as income level, population density, tourism activity, consumption patterns, land use, waste management facilities, and local waste management policies, were not included in the model. In addition, the use of representative coordinates at the regency/city level may not fully capture spatial variation within each administrative area. Therefore, future research is recommended to use more complete data covering all regencies and cities in East Java, apply multi-year or panel data, include additional socioeconomic, infrastructure, and waste management variables, and compare GWR with other spatial approaches such as spatial lag models, spatial error models, or spatial panel models.

Ethics approval

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Data obtained via the Statistics Indonesia of East Java Province website https://jatim.bps.go.id/, the National Waste Management Information System (SIPSN) website https://sipsn.menlhk.go.id/, and the East Java Provincial Health Office publication (East Java Province Health Profile).

Credit Authorship

Syefa Ilmi Beandita Putri: Conceptualization, Methodology, Data Collection, Data Analysis, Software Development, Visualization, Writing–Original Draft. **Sri Pingit Wulandari:** Conceptualization, Methodology, Manuscript Review, Research Advisor, Writing–Review.

References

[1] SIPSN, "Timbulan Sampah." Accessed: May 18, 2025. [Online]. Available: https://sipsn.kemenlh.go.id/sipsn/public/data/timbulan

[2] F. Kaza, S.; Yao, L.; Bhada-Tata, P.; Van Woerden, *What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050. World Bank.* Washington DC: The World Bank, 2018.

[3] U. N. E. Programme, "Global Waste Management Outlook 2024: Beyond an Age of Waste," Nairobi, 2024. [Online]. Available: https://www.unep.org/resources/global-waste-management-outlook-2024

[4] K. Popli, C. Park, S. Han, and S. Kim, "Prediction of Solid Waste Generation Rates in Urban Region of Laos Using Socio-Demographic and Economic Parameters with a Multi Linear Regression Approach," *MDPI*, vol. 13, no. 6, 2021, doi: https://doi.org/10.3390/su13063038.

[5] W. Lu, W. Huo, H. Gulina, and C. Pan, "Development of machine learning multi-city model for municipal solid waste generation prediction," vol. 16, no. 9, 2022.

[6] L. Fontaine, R. Legros, and J. Frayret, "Solid waste generation prediction model framework using socioeconomic and demographic factors with real-time MSW collection data," 2025, doi: 10.1177/0734242X241231414.

[7] F. E. Putri, "Pemodelan Timbulan Sampah Berdasarkan Faktor Sosioekonomi dan Sosiodemografi di Jawa Tengah Menggunakan Geographically Weighted Regression," Institut

Teknologi Sepuluh Nopember Surabaya, 2024.
[8] A. S. F. C. B. M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: John Wiley & Sons, 2002.
[9] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," vol. 28, no. 4, 1996.
[10] D. C. . E. A. P. G. G. V. Montgomery, *Introduction to Linear Regression Analysis (5th ed.)*. Wiley, 2013.
[11] N. N. Kachouie, "Association Factor for Identifying Linear and Nonlinear Correlations in Noisy Conditions," pp. 1–19, 2020, doi: 10.3390/e22040440.
[12] X. Wang and M. Braun, "Explainable machine learning-based fatigue assessment of 316L stainless steel fabricated by laser-powder bed fusion," *Int. J. Fatigue*, vol. 190, no. August 2024, p. 108588, 2025, doi: 10.1016/j.ijfatigue.2024.108588.
[13] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models Fifth Edition*. New York: McGraw-Hill/Irwin, 2005.
[14] S. W. Tyas and L. A. Puspitasari, "Geographically weighted generalized poisson regression model with the best kernel function in the case of the number of postpartum maternal mortality in east java," *MethodsX*, vol. 10, no. January, p. 102002, 2023, doi: 10.1016/j.mex.2023.102002.
[15] R. Cano-guervos, J. Chica-olmo, and J. Chica-garcia, "Journal of Retailing and Consumer Services The effect of disruptive change on the spatial variation of commercial rental prices : The case of the COVID-19 pandemic," *J. Retail. Consum. Serv.*, vol. 82, no. September 2024, p. 104111, 2025, doi: 10.1016/j.jretconser.2024.104111.
[16] S. W. Tyas, Gunardi, and L. A. Puspitasari, "Geographically weighted generalized poisson regression model with the best kernel function in the case of the number of postpartum maternal mortality in east java," *MethodsX*, vol. 10, no. January, p. 102002, 2023, doi: 10.1016/j.mex.2023.102002.
[17] M. N. Lessani and Z. Li, "SGWR : similarity and geographically weighted regression," *Int. J. Geogr. Inf. Sci.*, vol. 38, no. 7, pp. 1232–1255, 2024, doi: 10.1080/13658816.2024.2342319.
[18] I. Gollini, B. Lu, M. Charlton, C. Brunsdon, and P. Harris, "GWmodel : An R Package for Exploring Spatial," vol. 63, no. 17, 2015.
[19] C. Sutherland, D. Hare, P. J. Johnson, D. W. Linden, R. A. Montgomery, and E. Droge, "Practical advice on variable selection and reporting using Akaike information criterion," no. Dic, 2023.
[20] N. Nadja *et al.*, "Conceptualizing the Circular Economy (Revisited) An Analysis of 221 Definitions," *Resour. Conserv. Recycl.*, vol. 194, p. 107001, 2026, doi: 10.1016/j.resconrec.2023.107001.
[21] J. K. D. D. G. V. M. A. P. Dinis, "Raising Awareness on Solid Waste Management through Formal Education for Sustainability : A Developing Countries," vol. 6, no. 1, 2021, doi: https://doi.org/10.3390/recycling6010006.
[22] W. Fadhullah, N. Iffah, N. Imran, S. Norkhadijah, S. Ismail, and M. H. Jaafar, "Household solid waste management practices and perceptions among residents in the East Coast of Malaysia," *BMC Public Health*, vol. 22, no. 1, pp. 1–20, 2022, doi: 10.1186/s12889-021-12274-7.
[23] Z. Zhang *et al.*, "Municipal solid waste management challenges in developing regions : A comprehensive review and future perspectives for Asia and Africa," *Sci. Total Environ.*, vol. 930, no. April, p. 172794, 2024, doi: 10.1016/j.scitotenv.2024.172794.
[24] L. K. Ncube, A. U. Ude, E. N. Ogunmuyiwa, R. Zulkifli, and I. N. Beas, "An Overview of Plastic Waste Generation and Management in Food Packaging Industries," *MDPI*, vol. 6, no. 1, 2021, doi: https://doi.org/10.3390/recycling6010012.

DOI: https://doi.org/10.34123/jurnalasks.v18i1.954

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Dimension Reduction of Socioeconomic Factors in Deforestation Analysis in Indonesia Using Sparse PCA

Mitha Rabiyatul Nufus[1*], Jenike Gracelya Noke[2] , Eusabius Paul Pega[3]

[1]Forest Management Study Program, Department of Forestry, Kupang State Polytechnic of Agriculture, Kupang,, Indonesia; [2]Fisheries Agribusiness Study Program, Department of Fisheries and Marine Affairs, Kupang State Polytechnic of Agriculture, Kupang, Indonesia; [3]Horticultural Industrial Technology Study Program, Department of Food Crops and Horticulture, Kupang State Polytechnic of Agriculture, Kupang, Indonesia
**Corresponding Author: E-mail address: mitha.nufus@staff.politanikoe.ac.id*

ARTICLE INFO

Article history:
Received 25 March, 2026
Revised 12 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Deforestation; Dimensionality Reduction; Indonesia; Socioeconomic Factors; Sparce Principal Component Analysis; Spatial Analysis

Abstract

Introduction/Main Objectives: Deforestation remains a major environmental challenge in Indonesia under diverse socio-economic conditions. This study applies Sparse Principal Component Analysis (SPCA) to identify the key socio-economic variables associated with deforestation patterns. **Background Problems:** Analyses of deforestation drivers often involve numerous correlated variables, leading to multicollinearity and making interpretation difficult. Therefore, an approach is needed to reduce data dimensionality while retaining the most relevant information. **Novelty:** This study employs SPCA to simultaneously perform dimensionality reduction and variable selection, producing a more interpretable framework for identifying socio-economic factors related to deforestation at the provincial level in Indonesia. **Research Methods:** Provincial-level socio-economic data from Statistics Indonesia were analyzed using SPCA to address multicollinearity and derive interpretable components. Spatial autocorrelation was assessed using Moran's I. **Finding/Results:** SPCA reduced the variables into two interpretable components and identified six key contributing variables while excluding three with limited influence. Moran's I values for the first (0.402) and second (0.258) sparse principal components indicated significant positive spatial clustering of provinces with similar deforestation-related characteristics. **Research Limitations:** The analysis is limited to provincial-level secondary data and may not fully capture local-scale variations or all determinants of deforestation.

1. Introduction

Deforestation is widely recognized as a major global environmental concern due to its significant implications for climate change, biodiversity loss, and ecosystem degradation. Indonesia contains one of the largest tropical forest areas in the world; however, significant forest loss has occurred over recent decades due to agricultural expansion, land-use change, and forest fires. This decline in forest cover contributes to carbon emissions, threatens biodiversity, and alters ecosystem functions, thereby affecting both environmental sustainability and climate change mitigation efforts [1]. Beyond its environmental consequences, deforestation also generates substantial social and economic implications, particularly for communities whose livelihoods depend directly on forest resources. Forests serve not only ecological functions but also provide essential goods and services that sustain local economies and support household well-being. Consequently, the promotion of sustainable forest management has become a

critical policy priority aimed at preserving environmental stability while simultaneously fostering sustainable development, especially in many developing countries. In a regional context, countries endowed with extensive tropical forest areas face significant challenges in balancing economic development with environmental conservation. Economic activities such as agricultural expansion, infrastructure development, and increasing investment frequently drive land-use changes that may accelerate deforestation processes. A growing body of literature indicates that socioeconomic factors including population growth, income levels, and the development of the agricultural sector are closely associated with the dynamics of forest cover change [2]. This finding indicates that deforestation is shaped not solely by ecological factors but also by the evolving social and economic dynamics within a region.

Indonesia, recognized as one of the countries with the largest extent of tropical forests in the world, plays a strategic role in maintaining global ecological balance. Indonesia, recognized as one of the countries with the largest tropical forest resources in the world, plays a strategic role in maintaining global ecological balance. According to the Ministry of Environment and Forestry, Indonesia's forest area was estimated at approximately 120.4 million hectares in 2023, representing a substantial proportion of the country's land area and distributed across major regions such as Kalimantan, Sumatra, and Papua. These forest ecosystems perform essential ecological functions, including acting as significant carbon sinks, regulating climate systems, and providing habitats for numerous endemic species with high ecological value [3]. However, pressure on forest areas in Indonesia continues to intensify alongside population growth, economic expansion, and the increasing demand for land to support various development activities.

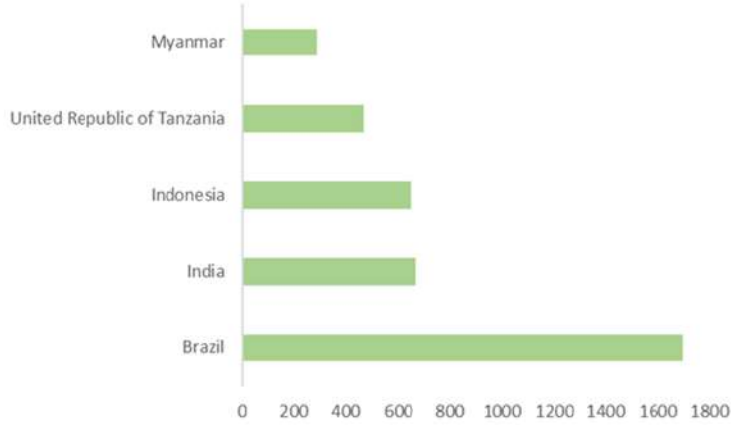


Figure 1. Deforestation rate

Based on the graphical trend, Indonesia experienced an average deforestation rate of approximately 650 thousand hectares per year (650 kha/year) during the 2015–2022 period [4]. This figure positions Indonesia among countries experiencing relatively high levels of forest loss compared with several other nations possessing extensive tropical forest areas. The elevated rate of deforestation in Indonesia is generally associated with multiple interacting drivers, including the expansion of agricultural and plantation land, population growth, and increasing development activities that require the conversion of forest areas into other land uses. These circumstances highlight the critical importance of implementing sustainable forest management practices and strengthening land-use control policies to mitigate the rate of forest cover loss in Indonesia while ensuring the long-term sustainability of forest ecosystems.

The phenomenon of deforestation in Indonesia does not occur in a linear pattern but rather exhibits fluctuations over time. Statistical records indicate that the national rate of deforestation has experienced notable variations in recent years, influenced by a combination of environmental, economic, and social factors. Socioeconomic variables such as population growth, poverty levels, and economic development are frequently associated with increasing pressure on forest resources. Population growth, for instance, tends to intensify the demand for land to support settlements, agricultural expansion, and various economic activities, thereby increasing the likelihood of forest area conversion [3]. A similar observation was reported by Njoya et al. (2024), who indicated that population growth accompanied by the expansion of economic activities tends to intensify the demand for land resources, thereby accelerating land-use change.

In empirical studies, the analysis of the drivers of deforestation frequently involves numerous explanatory variables that exhibit substantial intercorrelation. Socio-economic indicators such as

population size, the Human Development Index, poverty rate, and other development-related measures often display complex relationships with one another. This condition can lead to multicollinearity problems and make it difficult to determine which variables exert the most significant influence on deforestation dynamics. To address this issue, a statistical approach capable of reducing the dimensionality of the data while retaining essential information is required. One commonly applied technique for this purpose is Principal Component Analysis (PCA), which reduces the number of variables by transforming the original variables into a smaller set of uncorrelated components. Through this transformation, PCA captures most of the variability contained in the dataset using a limited number of principal components. Despite its advantages, conventional PCA has limitations in terms of interpretability. Each principal component is typically formed as a linear combination of nearly all original variables, which complicates the interpretation of the resulting components. To overcome this limitation, the Sparse Principal Component Analysis (Sparse PCA) method was developed. Sparse PCA introduces sparsity constraints that allow principal components to be constructed using only a subset of variables, thereby producing components with fewer non-zero loadings. As a result, the derived components become easier to interpret and provide more informative insights for identifying the key socio-economic factors associated with deforestation. Through this approach, a set of principal components is expected to be identified that can represent the underlying structure of relationships among socio-economic variables in a more concise and informative manner. By reducing the dimensionality of the data, the analysis is anticipated to capture the essential patterns of association among variables while minimizing redundancy. Numerous studies have examined the socioeconomic factors associated with deforestation using conventional statistical methods and Principal Component Analysis (PCA). Although PCA is effective in reducing data dimensionality and addressing multicollinearity, the resulting components are often difficult to interpret because most variables contribute to each component simultaneously. As a result, identifying the key socioeconomic factors underlying deforestation remains challenging. Furthermore, the application of Sparse Principal Component Analysis (SPCA), which produces a more parsimonious and interpretable component structure, has received limited attention in studies of deforestation in Indonesia. Therefore, this study applies SPCA to reduce the dimensionality of socioeconomic variables related to deforestation in Indonesia and to identify the most influential factors through a more interpretable component structure. The findings are expected to provide a clearer understanding of the socioeconomic dimensions associated with deforestation and support the development of more targeted forest management policies.

2. Material and Methods

2.1. Data Sources

This study utilizes secondary data at the provincial level covering all provinces in Indonesia. The data were obtained from BPS–Statistics Indonesia. The dataset consists of deforestation-related and socioeconomic indicators, including the Human Development Index, population density, population growth rate, total population, poverty rate, Gross Regional Domestic Product (GRDP) per capita, mean years of schooling, agricultural land area, and the number of forest fires. The analysis includes all provinces in Indonesia. The dataset represents provincial-level indicators in Indonesia that reflect social, economic, and environmental conditions associated with the dynamics of forest cover change. The use of data from a national statistical agency is essential because such datasets are collected through standardized procedures and rigorous verification processes, ensuring a high level of reliability for empirical analysis. Furthermore, regional statistical data are widely employed in environmental research to examine the relationship between socioeconomic factors and land-use change, particularly in studies addressing deforestation and environmental degradation [5]. The variables employed in this study consist of one dependent variable and several independent variables representing socio-economic indicators and pressures on forest resources. These variables were selected because, from a conceptual standpoint, they are frequently utilized in studies examining the relationship between socio-economic dynamics and changes in forest cover. Such indicators are particularly relevant in the context of developing countries, where development activities and socio-economic transformation often exert significant pressure on forested landscapes [5]. The inclusion of multiple socio-economic indicators in the analysis may lead to a high degree of intercorrelation among variables. To address this issue, the present study employs a dimensionality reduction approach using the Sparse Principal Component Analysis (Sparse PCA) method. This technique enables the transformation of a set of correlated variables into a smaller number of more interpretable principal components while enforcing sparsity in the component loadings. Consequently, the method facilitates a clearer identification of the key socio-economic factors associated with deforestation in Indonesia.

Table 1. Research variables

Variable	Description	Unit	Source	Year
Y	Deforestation Rate	Hectares per Year	Statistics Indonesia (BPS)	2023
X1	Human Development Index	Index (0-100)	Statistics Indonesia (BPS)	2023
X2	Population Density	People/km²	Statistics Indonesia (BPS)	2023
X3	Population Growth Rate	Percent per Year	Statistics Indonesia (BPS)	2023
X4	Total Population	Thousand People	Statistics Indonesia (BPS)	2023
X5	Poverty Rate	Percentage	Statistics Indonesia (BPS)	2023
X6	Gross Regional Domestic Product (GRDP) per Capita	Million Rupiah/Person/Year	Statistics Indonesia (BPS)	2023
X7	Mean Years of Schooling	Years	Statistics Indonesia (BPS)	2023
X8	Agricultural Land Area	Thousand Hectares	Statistics Indonesia (BPS)	2023
X9	Number of Forest Fires	Number of incidents	Statistics Indonesia (BPS)	2023

The study considered several socioeconomic indicators as predictor variables (X), including the Human Development Index, population density, poverty rate, GRDP per capita, educational attainment, agricultural land area, and forest fire incidence. These variables were included in the Sparse Principal Component Analysis to identify the underlying socioeconomic dimensions associated with deforestation. The deforestation rate (Y) was included only for descriptive and interpretative purposes. It was not used in the construction of the sparse principal components, as SPCA is an unsupervised technique that operates solely on the predictor variables. Instead, the deforestation rate was presented to provide context and facilitate the interpretation of the extracted socioeconomic patterns.

2.2 Correlation

Correlation is a statistical measure employed to quantify both the strength and direction of a linear association between two variables. Within the context of multivariate analysis, it serves as an initial step for examining the underlying relationships among variables prior to conducting more advanced modelling procedures. Among the available measures, the Pearson correlation coefficient is the most widely utilized for assessing linear dependence. Mathematically, the Pearson correlation coefficient is expressed as follows,

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right)\left(n\sum Y^2 - \left(\sum Y\right)^2\right)}} \quad (1)$$

where r denotes the correlation coefficient, n represents the number of observations, X and Y correspond to the variables under analysis. The correlation coefficient ranges between $-1 \leq r \leq 1$. A positive value ($r > 0$) indicates a direct relationship between variables, whereas a negative value ($r < 0$) reflects an inverse relationship. Furthermore, as the absolute value of the coefficient ($|r|$) approaches 1, the strength of the association becomes increasingly strong. In the context of multivariate analysis, strong correlations among independent variables may serve as an initial indication of multicollinearity. Several studies suggest that absolute correlation coefficients exceeding approximately 0.7–0.8 reflect a substantial linear association, which may imply overlapping information and potential redundancy among variables [6]. Correlation analysis is often complemented by graphical approaches, such as scatterplot matrices, to facilitate a more comprehensive examination of both linear and nonlinear relationships among variables.

2.3 Multicollinearity

Multicollinearity refers to a situation in which two or more independent variables in a model exhibit strong linear associations. This phenomenon is frequently encountered in social, economic, and environmental datasets, where indicators are often inherently interrelated. From a conceptual standpoint, multicollinearity inflates the variance of regression parameter estimators, leading to unstable coefficient estimates and reduced interpretability. As a consequence, the overall reliability of the model may decline, particularly in terms of statistical inference and the assessment of variable significance [7]. One of the most widely applied approaches for diagnosing multicollinearity is the *Variance Inflation Factor* (VIF). This metric quantifies the extent to which the variance of an estimated regression coefficient is amplified due to linear dependencies among the independent variables. In other words, VIF reflects how strongly a predictor is explained by the remaining predictors in the model.

Mathematically, VIF is expressed as:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2}$$

where R_i^2 represents the coefficient of determination obtained by regressing the i-th independent variable on the remaining independent variables in the model. The interpretation of Variance Inflation Factor (VIF) values can be outlined as; a VIF equal to 1 indicates the absence of correlation among explanatory variables. Values ranging between 1 and 5 suggest a moderate degree of association that is generally acceptable within regression models. When the VIF exceeds 5, it signals the presence of potential multicollinearity, while values greater than 10 are commonly regarded as evidence of severe multicollinearity, which may substantially affect the stability and reliability of parameter estimates.

A high VIF value indicates that a given variable is substantially explained by the remaining predictors in the model, leading to an inflation of the variance of its estimated regression coefficient. From a mathematical standpoint, this increase in variance directly affects the standard error, which can be expressed as follows,

$$SE\left(\beta_i\right) = SE_{OLS(\beta_i)}\sqrt{VIF_i} \tag{3}$$

Accordingly, higher VIF values are associated with increased uncertainty in parameter estimation, reflecting greater instability in the estimated coefficients.

2.4 Sparse PCA

Principal Component Analysis (PCA) is a widely employed multivariate statistical technique designed to reduce the dimensionality of high-dimensional datasets. The method operates by transforming a set of potentially correlated variables into a smaller number of new variables, known as principal components, which are mutually orthogonal and capture the largest possible proportion of variance in the original data. Through this transformation, PCA enables the extraction of the most informative structure of the dataset while minimizing redundancy among variables. Nevertheless, a notable limitation of classical PCA lies in the fact that each principal component is typically expressed as a linear combination of all original variables. Consequently, the resulting components may involve contributions from many variables simultaneously, which often complicates the interpretability of the components in empirical applications [8]. To address these limitations, the Sparse Principal Component Analysis method was developed as an extension of Principal Component Analysis. This approach introduces the concept of *sparsity* in the coefficient vectors of the principal components. In practice, Sparse PCA produces principal components that involve only a limited subset of variables with non-zero loadings, while the remaining variables are assigned coefficients equal to zero. Consequently, this technique not only performs dimensionality reduction but also implicitly carries out variable selection. Such a property improves the interpretability of the resulting components and facilitates a clearer understanding of the relationships among variables within the analyzed dataset [9].

Conceptually, Sparse Principal Component Analysis (Sparse PCA) aims to derive principal components that not only explain the variability of the data, as in classical Principal Component Analysis (PCA), but also exhibit a sparse coefficient structure. This approach introduces additional constraints or penalty terms on the component loadings, encouraging many coefficients to shrink toward zero. As a result, the resulting components are constructed from only a limited subset of variables that contribute most strongly to the underlying data structure. This sparsity property enhances interpretability by allowing the principal components to be associated with a smaller [10]. The Sparse Principal Component

Analysis (Sparse PCA) approach has been widely applied in the analysis of high-dimensional datasets, including genomic, economic, and socio-economic data, due to its capability to simplify complex variable structures while preserving the essential information contained in the data. By imposing sparsity constraints on the component loadings, Sparse PCA produces principal components that involve only a limited subset of variables, thereby enhancing interpretability. In addition to functioning as a dimensionality reduction technique, Sparse PCA can also serve as a variable selection method in multivariate analysis, as it effectively identifies the most relevant variables that contribute to the underlying data structure [9].

Consider a data matrix $X \in R^{(}n \times p)$where (n) represents the number of observations and (p) denotes the number of variables. In the classical Principal Component Analysis (PCA) framework, the first principal component is obtained by maximizing the variance of the projected data, which can be formulated as follows,

$$\max_{\omega} \omega^T \sum \omega \tag{4}$$

With the constraint $\|\omega\|_2 = 1$, where ω denotes the loading vector of the principal component and $\sum$ represents the covariance matrix of the observed data. In the framework of Sparse Principal Component Analysis (Sparse PCA), this formulation is modified by incorporating a sparsity constraint on the loading vector. The objective is to produce principal components that not only retain the ability to explain the variability of the data, as in classical PCA, but also contain a limited number of non-zero loadings to enhance interpretability. One commonly adopted formulation is presented in Equation (1). In addition to the constraint $\|\omega\|_2 = 1$, a sparsity restriction is imposed in the form $\|\omega\|_1 \leq c$. Alternatively, the same constraint can be expressed through a penalty function approach.

$$\max_{\omega} \left(\omega^T \sum \omega - \lambda \|\omega\|_1 \right) \tag{5}$$

where,

ω : loading vector of the principal component
$\sum$: covariance matrix of the data
$\|\omega\|_1$: L_1norm controlling the level of sparsity
λ : penalty parameter that determines the degree of sparsity

The incorporation of an L_1 penalty encourages sparsity in the loading vectors by shrinking a number of coefficients exactly to zero. As a consequence, only a subset of variables contributes to the construction of the principal components. Through this mechanism, Sparse Principal Component Analysis (Sparse PCA) serves not only as a dimensionality reduction technique but also as an effective variable selection approach within multivariate analysis [10]. The approach proposed by Hui Zou reformulates principal component analysis (PCA) as a penalized regression problem. Within this framework, the objective of Sparse PCA is not only to capture the maximum variance in the data as in classical Principal Component Analysis but also to impose sparsity on the loading vectors through appropriate penalty terms. Consequently, the objective function of Sparse PCA can be expressed as follows [11].

$$\min_{A,B} \left\| \mathbf{X} - \mathbf{XBA}^T \right\|^2 + \lambda_1 \sum_j |b_j| + \lambda_2 \sum_j b_j^2 \tag{6}$$

With the constraint applied $A^T A = I$.

Where,

X : the standardized data matrix
A : the matrix representing the principal component scores
B : the matrix of principal component loadings
λ_1 : the L_1 regularization parameter (lasso) that promotes sparsity in the model
λ_2 : the L_2 regularization parameter (ridge) that ensures the stability of parameter estimation

The combination of L_1 and L_2 penalties is commonly referred to as the elastic net penalty. The L_1 penalty promotes sparsity by shrinking some coefficients exactly to zero, thereby producing a simpler and more interpretable loading structure. In contrast, the L_2 penalty contributes to stabilizing the

estimation process by mitigating the effects of multicollinearity among variables. Consequently, the elastic net framework enables the extraction of principal components that are both parsimonious and robust in the presence of correlated predictors.

2.5 Step of Analysis

The analytical procedure in this study was carried out through a series of systematic stages to ensure the reliability and interpretability of the results. The analysis began with the collection of secondary data representing socio-economic factors associated with deforestation across Indonesia. The dataset was subsequently cleaned and standardized to improve data quality and ensure comparability among variables. Descriptive statistical analysis and a correlation matrix were then employed to provide an initial overview of the data distribution and to explore the relationships among the socio-economic variables. To identify potential multicollinearity, the Variance Inflation Factor (VIF) was calculated for all explanatory variables. The suitability of the dataset for dimension reduction was further evaluated using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity.

Following these preliminary assessments, all variables were standardized prior to dimensionality reduction. Principal Component Analysis (PCA) was first performed to investigate the latent structure of the socio-economic variables and to summarize the original information into a smaller number of components. The appropriate number of principal components was determined by considering eigenvalues, the cumulative proportion of explained variance, and the scree plot. To improve component interpretability while retaining the most relevant variables, Sparse Principal Component Analysis (SPCA) was subsequently applied. The resulting loading matrix was examined to identify the socio-economic variables that contributed most strongly to each sparse component. Component scores and loading patterns were then visualized to facilitate the interpretation of observation clusters and variable contributions. Finally, the spatial distribution of the SPCA scores was mapped and analyzed to reveal regional differences in socio-economic characteristics related to deforestation across Indonesia. The overall findings were subsequently synthesized to draw conclusions and discuss their implications. All statistical analyses and graphical visualizations were performed using R software (version 4.5.3) with the support of several packages, including stats, car, psych, FactoMineR, factoextra, and elasticnet.

3. Results and Discussion

3.1 Descriptive Statistics

The distribution of provinces with the highest values for each research variable related to socio-economic conditions and pressures on forest resources is presented in Figure 2. In general, the distribution of maximum values across several indicators appears to be concentrated in specific provinces, indicating substantial disparities in development characteristics among regions in Indonesia. The highest values for the Human Development Index (X1), population density (X2), gross regional domestic product (GRDP) per capita (X6), and average years of schooling (X7) are observed in Special Capital Region of Jakarta. This pattern reflects the relatively advanced level of socio-economic development in the country's principal metropolitan area. In contrast, variables related to land-use pressure, such as agricultural land area (X8) and the number of forest fires (X9), reach their maximum values in provinces including Riau and Central Kalimantan, regions widely recognized for intensive land expansion activities and higher susceptibility to forest fire events. Previous research highlights that agricultural expansion and fire events are among the major drivers of forest loss in tropical regions, particularly in Southeast Asia [12]. Furthermore, the largest population size (X4) is recorded in West Java, highlighting the considerable demographic pressure experienced in this province. These patterns collectively suggest that regional development dynamics, demographic intensity, and land-use transformation play a crucial role in shaping environmental pressures and the potential risk of deforestation across Indonesia.

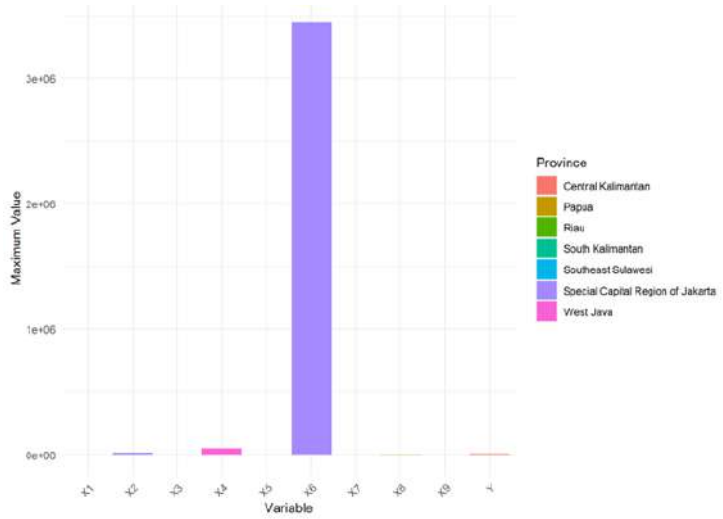


Figure 2. Descriptive statistics of each variable

3.2 Correlation

The correlation matrix and scatterplot visualization reveal varying degrees of association among variables (X1–X9), including both positive and negative relationships. Several variable pairs exhibit relatively strong and statistically significant correlations, indicating the presence of underlying structural relationships within the dataset. Notably, X1 shows a strong positive association with X7 and negative correlations with X3 and X5, while X4 and X6 demonstrate a strong linear relationship. The presence of relatively high correlation coefficients suggests potential multicollinearity among predictors, which may affect the stability and interpretability of regression-based models. Previous studies have highlighted that multicollinearity can distort parameter estimation and reduce model reliability [13]. Therefore, the application of Sparse PCA is justified, as it enables dimensionality reduction while maintaining interpretability through sparse loading structures.

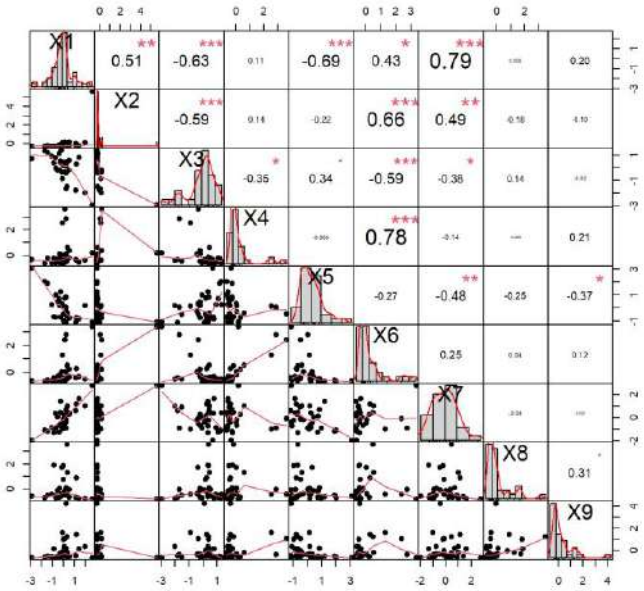


Figure 3. Matrix correlation of each variables

3.3 Multicollinearity

The Variance Inflation Factor (VIF) values for each variable, along with their corresponding visual representation, are presented as follows.

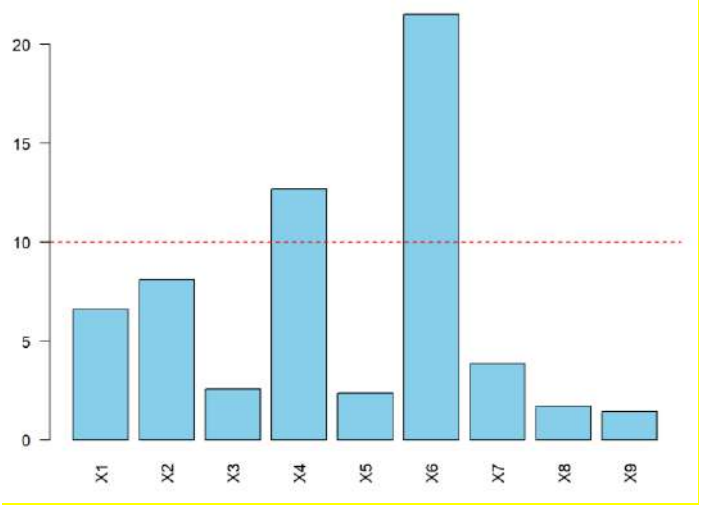


Figure 4. The visualization of VIF

The Variance Inflation Factor (VIF) analysis indicates that most variables exhibit relatively low VIF values, suggesting an acceptable level of multicollinearity. However, variables X4 and especially X6 display substantially high VIF values, exceeding common threshold levels. This suggests that these variables are highly linearly dependent on other predictors in the model. High VIF values indicate inflated variance in regression coefficients, which may lead to unstable parameter estimates and reduced interpretability. Although no universal cutoff exists, values between 5 and 10 are commonly used as indicators of potential multicollinearity [14]. Therefore, the observed values for X4 and X6 confirm the presence of strong multicollinearity. These findings justify the application of dimensionality reduction techniques such as Sparse PCA to address redundancy and improve model robustness.

3.4 Scree Plot of Sparse PCA

The scree plot is a graphical technique commonly employed in Principal Component Analysis to determine the appropriate number of principal components to retain during dimensionality reduction. This plot presents the eigenvalues, or the proportion of variance explained by each component, arranged from the largest to the smallest. Through this visualization, researchers can identify the point at which the contribution of additional components begins to decline substantially, often referred to as the *elbow*. Components appearing before this point are generally considered sufficient to capture the most relevant information contained in the dataset, while subsequent components contribute relatively little additional explanatory power.

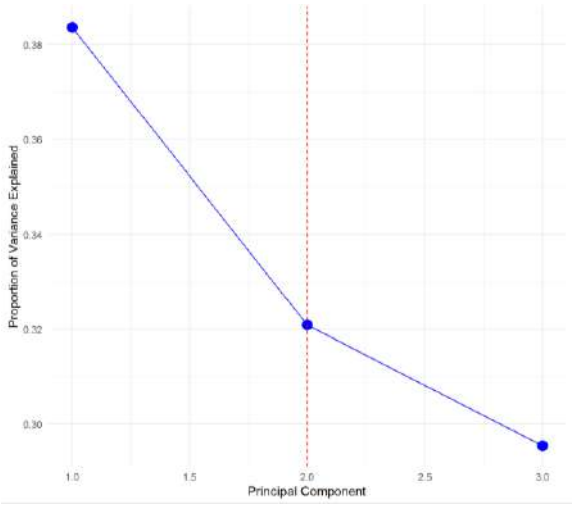


Figure 5. Scree plot sparse PCA

Consequently, the scree plot provides a clear visual basis for selecting an optimal number of components, enabling analysts to preserve essential information while reducing the complexity of multivariate data structures. For this reason, it is widely applied in statistical studies and

multidimensional data analysis as a practical tool for guiding component selection in PCA-based dimensionality reduction.

Based on the scree plot and component selection criteria, two components were retained for further interpretation. The first and second components contributed approximately 38.4% and 32.1% of the retained explained variance, respectively, indicating that these components captured the dominant structure of the socio-economic data.

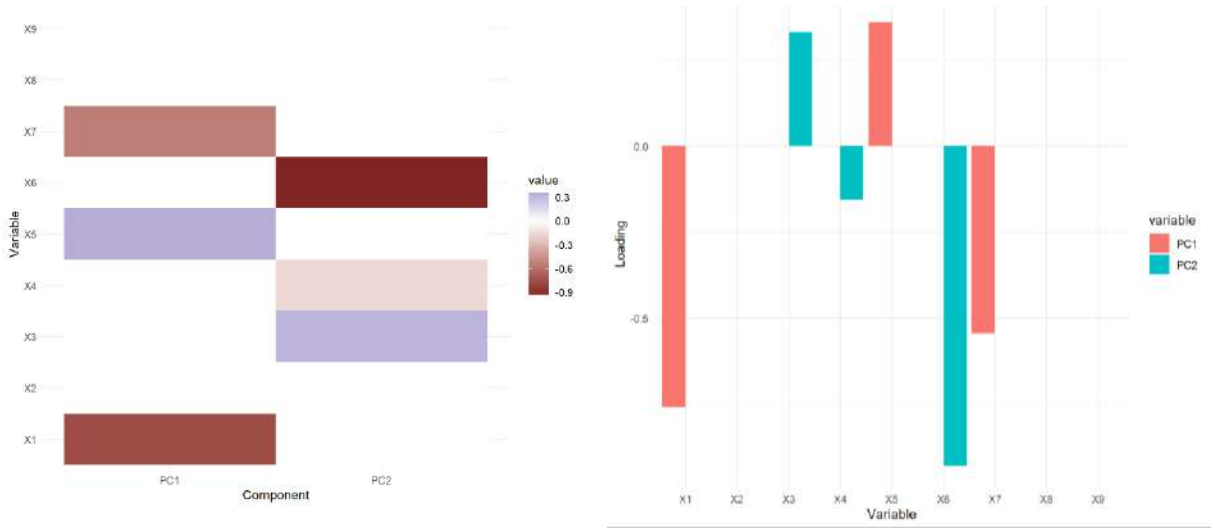
3.5 Sparse Loading

In Sparse Principal Component Analysis (Sparse PCA), the loading coefficients indicate the magnitude of each original variable's contribution to the derived principal components. A large loading value suggests that the variable is strongly associated with the corresponding component, whereas coefficients close to zero imply a relatively minor contribution. By introducing sparsity in the loading structure, Sparse PCA enhances the interpretability of the resulting components because only a limited number of variables retain substantial weights. In addition to improving interpretability, sparse loading reduces model complexity when dealing with high-dimensional datasets. Within multivariate analysis, Sparse PCA is widely applied as it preserves a substantial portion of the data variability while simultaneously identifying the most influential variables that shape the principal components [15].

Table 2. Sparse loading

Variable / PC	PC1	PC2
X1	-0.758	0.000
X2	0.000	0.000
X3	0.000	0.332
X4	0.000	-0.157
X5	0.360	0.000
X6	0.000	-0.930
X7	-0.544	0.000
X8	0.000	0.000
X9	0.000	0.000

In the first principal component (PC1), only three variables exhibit substantial loadings, namely X1 (−0.758), X5 (0.360), and X7 (−0.544). This indicates that PC1 is primarily characterized by the combined influence of these three variables, although their directions of association differ. Variables X1 and X7 are negatively associated with the first component, whereas X5 shows a positive relationship. In contrast, the remaining variables (X2, X3, X4, X6, X8, and X9) display zero loadings, indicating that they do not contribute to the formation of this component. This outcome reflects the sparsity constraint imposed by the Sparse PCA approach, which effectively eliminates variables with negligible contributions.



(a) (b)

Figure 6. (a) Heatmap loading sparse PCA (b) Visualization of sparse PCA loading

For the second principal component (PC2), the component structure is determined by three variables: X3 (0.332), X4 (−0.157), and X6 (−0.930). Among these, X6 provides the most dominant contribution to PC2, as reflected by its relatively large loading value. The other variables have zero loadings and therefore do not participate in forming the second component.

The resulting sparse loading pattern demonstrates that SPCA not only performs dimensionality reduction but also effectively highlights the most relevant variables, thereby improving the interpretability of the principal components. Such an approach is particularly advantageous in the analysis of high-dimensional multivariate data, as it enhances model interpretability while preserving essential information contained in the dataset. Overall, the results indicate that six variables (X1, X3, X4, X5, X6, and X7) make substantial contributions to the extracted sparse principal components. In contrast, variables X2 (Population Density), X8 (Agricultural Land Area), and X9 (Number of Forest Fires) exhibit negligible or zero loadings across the retained components. Consequently, these variables do not contribute significantly to the component structure identified by SPCA and were not retained as key variables in the interpretation of the resulting dimensions. This finding suggests that, within the analyzed dataset, their variability is not strongly aligned with the dominant socioeconomic patterns associated with deforestation captured by retained components. The exclusion of X2, X8, and X9 was based on their near-zero loading values in all retained sparse principal components. These findings demonstrate that Sparse PCA not only reduces the dimensionality of the dataset but also performs implicit variable selection, resulting in a more interpretable and parsimonious component structure compared with conventional PCA [8].

3.6 Score Plot of Sparse PCA

The score plot in Sparse Principal Component Analysis (Sparse PCA) is a graphical representation of two or more principal components that illustrates the position of each observation (sample data) within the component space obtained from the dimensionality reduction process. Typically, the plot is constructed using the first sparse principal component (SPC1) and the second sparse principal component (SPC2). Through this visualization, patterns, relationships, and potential groupings among observations can be more clearly identified after the original high-dimensional data have been projected into a lower-dimensional space. This representation facilitates the interpretation of data structure while preserving the most relevant variation captured by the sparse components [16].

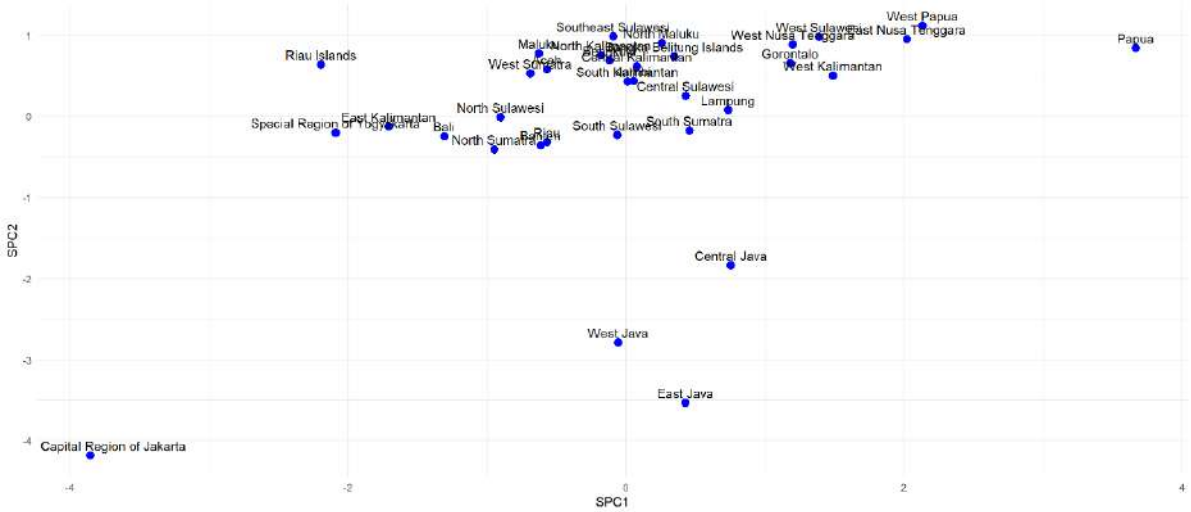


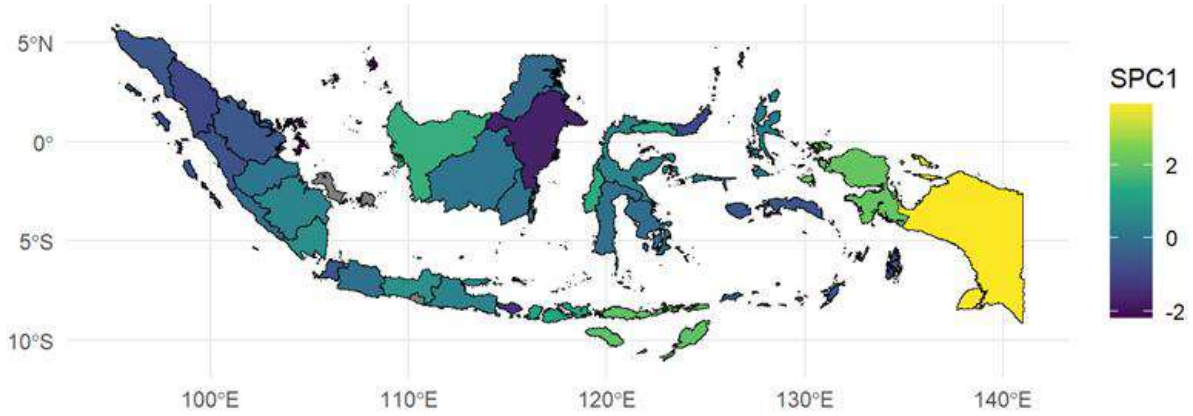
Figure 7. Sparse PCA score plot

In general, most provinces are located near the center of the coordinate system, indicating that their socio-economic characteristics are relatively similar and follow the overall national pattern. These provinces form the main cluster that represents the average condition of the analyzed variables. In contrast, several provinces appear farther from the center, indicating distinct socio-economic profiles. For example, Capital Region of Jakarta is positioned in the negative direction of both SPC1 and SPC2, reflecting a socio-economic structure that differs markedly from most other provinces due to its role as the national economic and administrative center. Some provinces on Java Island, including East Java, West Java, and Central Java, also appear slightly separated from the main cluster, particularly along the SPC2 dimension, indicating certain variations in their socio-economic indicators. Meanwhile, eastern

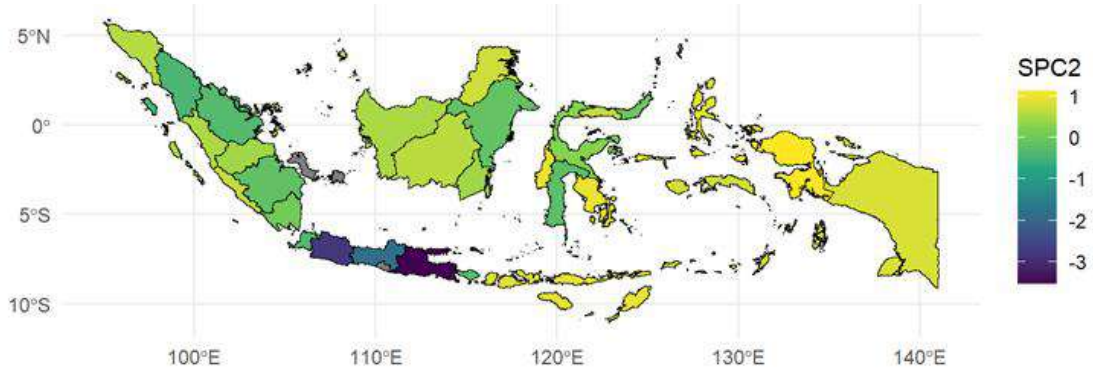
provinces such as Papua and West Papua tend to lie on the positive side of SPC1, suggesting different variable patterns compared to most provinces. Overall, the plot demonstrates that Sparse PCA effectively reduces data dimensionality while clearly illustrating similarities and differences in socio-economic characteristics among provinces. Overall, the score plot demonstrates that Sparse PCA effectively reduces data dimensionality while preserving the underlying variation among observations, thereby facilitating the visualization of similarities and differences in socio-economic characteristics among provinces [17].

3.7 Spatial Distribution of Sparse PCA

The spatial distribution in Sparse PCA refers to the mapping of principal component scores derived from Sparse Principal Component Analysis onto a geographic space. Through this approach, the dominant patterns of variation among multiple variables can be represented spatially. Such visualization facilitates the identification of regions that exhibit similar or contrasting characteristics based on the most influential combination of variables captured by the sparse components. This spatial representation enhances the interpretability of multivariate relationships and provides a clearer understanding of how underlying patterns vary across different geographic areas.



(a)



(b)

Figure 8. Spatial Distribution of (a) Sparse PCA Component 1 (b) Sparse PCA Component 2

The spatial distribution of SPC1 indicates that several provinces in eastern Indonesia, particularly Papua, exhibit relatively high component scores. In contrast, a number of provinces located in Kalimantan and parts of Sumatra show comparatively lower values. This spatial pattern suggests the presence of regional heterogeneity, which is primarily influenced by the combined effect of variables X1, X5, and X7, identified as the dominant contributors to the first sparse principal component. Such regional heterogeneity is consistent with previous studies demonstrating that socio-economic characteristics and development patterns vary considerably among Indonesian provinces, resulting in distinct spatial configurations of environmental and land-use dynamics [18]. The spatial map of SPC2

reveals a different distributional pattern. Several provinces across Sumatra, Kalimantan, and parts of eastern Indonesia display relatively high scores, while provinces on Java tend to exhibit lower values. This indicates that the second component captures an alternative dimension of variation in the dataset, driven by a distinct set of variables compared with those influencing the first component.

To verify whether the spatial patterns observed in Figure 8 reflect statistically meaningful geographic structures, Global Moran's I statistics were calculated for the Sparse PCA scores. The results, presented in Table 3, indicate significant positive spatial autocorrelation for both SPC1 and SPC2, confirming that provinces with similar component scores tend to be spatially clustered rather than randomly distributed. The spatial map of SPC2 reveals a different distributional pattern. Several provinces across Sumatra, Kalimantan, and parts of eastern Indonesia display relatively high scores, while provinces on Java tend to exhibit lower values. This indicates that the second component captures an alternative dimension of variation in the dataset, driven by a distinct set of variables compared with those influencing the first component. Differences in regional socio-economic structures and land-use characteristics have also been reported in previous spatial analyses of Indonesia, highlighting that provinces may exhibit contrasting spatial patterns depending on the underlying socio-economic drivers [19].

Table 3. Global Moran's I statistics for sparse principal components

Component	Moran's I	p-value
SPC 1	0.40229380	3.025070e-06
SPC 2	0.25835357	1.912515e-03

To statistically assess the spatial patterns identified by Sparse PCA, Global Moran's I statistics were computed for the first two sparse principal components. As presented in Table 3, both components exhibited significant positive spatial autocorrelation. SPC1 yielded a Moran's I value of 0.402 (p = 3.025070e-06), indicating a moderate tendency for provinces with similar SPC1 scores to be geographically clustered. Likewise, SPC2 produced a Moran's I value of 0.258 (p = 1.912515e-03), suggesting a positive spatial association, although weaker than that observed for SPC1. These findings indicate that the spatial patterns displayed in Figure 8 are not randomly distributed across Indonesia. Provinces located in close geographic proximity tend to share similar characteristics represented by the sparse principal components. The stronger spatial dependence observed in SPC1 suggests that the dominant variation captured by this component is more spatially structured than the variation represented by SPC2.

Overall, the analysis demonstrates that among the nine initial variables examined, only six variables (X1, X3, X4, X5, X6, and X7) were retained because they exhibited non-zero loadings in the sparse principal components. In contrast, X2, X8, and X9 were excluded due to zero loadings, indicating that these variables contributed negligibly to the extracted component structure. This finding is consistent with the fundamental characteristic of Sparse Principal Component Analysis, which simultaneously performs dimensionality reduction and variable selection by producing sparse loading vectors. As a result, the derived components are more parsimonious and substantially easier to interpret than those obtained from conventional PCA, making SPCA particularly suitable for multivariate datasets with correlated predictors. Although the present study focuses on spatially distributed socio-economic data, the resulting sparse component structure also facilitates the interpretation of regional variation in the underlying factors associated with deforestation [20].

4. Conclusion

This study demonstrates that Sparse Principal Component Analysis (Sparse PCA) serves as an effective technique for reducing the dimensionality of socio-economic variables associated with deforestation in Indonesia, while simultaneously enhancing the interpretability of the analytical results. The findings indicate that two principal components are sufficient to represent the underlying structure of the dataset, capturing the majority of the total variance. The first component is largely characterized by the Human Development Index, poverty rate, and mean years of schooling, whereas the second component is primarily driven by population growth, total population, and GRDP per capita. Among the nine variables initially considered, only six were retained as significant contributors, while population density, agricultural land area, and the number of forest fires were excluded due to their minimal influence. This result underscores the capability of Sparse PCA not only in dimensionality reduction but also in performing implicit variable selection.

Furthermore, the spatial distribution and score plot analyses reveal substantial regional heterogeneity in socio-economic characteristics, particularly between western and eastern regions of

Indonesia. Overall, this study offers a more parsimonious and interpretable analytical framework for identifying key socio-economic drivers of deforestation, and provides meaningful insights to support evidence-based policy development in sustainable forest management. From a practical perspective, future studies are recommended to incorporate additional environmental and institutional variables to obtain a more comprehensive understanding of deforestation dynamics. Moreover, policymakers are encouraged to prioritize region-specific strategies that account for socio-economic disparities in order to enhance the effectiveness of sustainable forest management initiatives.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Mitha Rabiyatul Nufus: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Visualization, Manuscript Writing. **Jenike Gracelya Noke**: Investigation, Data Curation, Visualization, Manuscript Writing and Review. **Eusabius Paul Pega**: Formal Analysis, Investigation, Visualization, Manuscript Editing.

References

[1] M. Leon, G. Cornejo, M. Calderón, E. González-Carrión, and H. Florez, "Effect of Deforestation on Climate Change: A Co-Integration and Causality Approach with Time Series," *Sustainability,* vol. 14, no. 18, 2022, doi: https://doi.org/10.3390/su141811303.

[2] H. M. Njoya, K. Hounkpati, K. Adjonou, K. Kokou, S. Sieber, and K. Löhr, "Socioeconomic analysis of deforestation and economically sustainable farming systems to foster forest landscape restoration in central Togo," *Front. Sustain. Food Syst,* vol. 8, pp. 1-16, 2024, doi: https://doi.org/10.3389/fsufs.2024.1466008.

[3] M. R. Nufus, R. Silaban, E. P. Pega, J. G. Noke, E. F. Karo-Karo, "Analisis nonparametrik laju deforestasi dengan model spline truncated berbasis faktor sosial ekonomi di Indonesia," *Wahana Forestra: Jurnal Kehutanan,* vol. 20, no. 2, pp. 124-134, 2025, doi: https://doi.org/10.31849/f3xg9n24.

[4] Global Forest Watch, Retrieved from Indonesia Deforestation Rates: https://www.globalforestwatch.org/dashboards/country/IDN/?category=forest-change&lang=id&location=WyJjb3VudHJ5IiwiSUROIl0%3D&map=eyJjZW50ZXIiOnsibGF0IjotMi41Nzg0NTMwNjA1NjA0NjU2LCJsbmciOjExOC4wMTUxNTU3ODk5ODA5Mn0sInpvb20iOjIuNDMxNTQ5Njk3NjM2NDc1LCJjYW5Cb3, March. 2026.

[5] A. Tyukavina, P. Potapov, M. C. Hansen, A. H. Pickens, S. V. Stehman, S. Turubanova, D. Parker, V. Zalles, A. Lima, I. Kommareddy, X. P. Song, L. Wang, and N. Harris, "Global Trends

of Forest Loss Due to Fire From 2001 to 2019," *Front. Remote Sens,* vol. 3, pp. 1-20, 2022, doi: https://doi.org/10.3389/frsen.2022.825190.

[6] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *American Journal of Applied Mathematics and Statistics,* vol. 8, no. 2, pp. 39-42, 2020, doi: https://doi.org/10.12691/ajams-8-2-1.

[7] H. I. Dertli, D. B. Hayes, and T. G. Zorn, "Effects of multicollinearity and data granularity on regression models of stream temperature," *Journal of Hydrology,* vol. 639, pp. 1-11, 2024, doi: https://doi.org/10.1016/j.jhydrol.2024.131572.

[8] B. B. Alkan and I. Ünaldi, "Robust sparse principal component analysis: situation of full sparseness," *Journal of Applied Mathematics, Statistics and Informatics,* vol. 18, no. 1, pp. 5-20, 2022, doi: https://doi.org/10.2478/jamsi-2022-0001.

[9] A. Chowdhury, A. Bose, S. Zhou, D. P. Woodruff, and P. Drineas, "A Fast, Provably Accurate Approximation Algorithm for Sparse Principal Component Analysis Reveals Human Genetic Variation Across the World," *Annual International Conference, RECOMB*, pp. 86-106, 2022, doi: https://doi.org/10.1007/978-3-031-04749-7_6.

[10] S. Park, E. Ceulemeans, and K. Van Deun, "Acritical assessment of sparse PCA (research): why (one should acknowledgethat) weights are notloadings," *Behaviour Research Methods,* vol. 56, pp. 1413-1432, 2024, doi: https://doi.org/10.3758/s13428-023-02099-0.

[11] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,"* vol. 15, no. 2, pp. 265-286, 2006, doi: https://doi.org/10.1198/106186006X113430.

[12] Q. Ke, S. Xu, S. Zong, X. Jiang, and S. Li, "Spatiotemporal dynamics and driving mechanisms of agricultural expansion into forests in Southeast Asia," *Journal of Environmental Management,* vol. 393, pp. 1-15, 2025, doi: https://doi.org/10.1016/j.jenvman.2025.127030.

[13] C. A. Asrat, Makkulau, and I. Yahya, "Perbandingan Metode Principal Component Analysis (PCA) dan Partial Least Square (PLS) dalam Penanganan Multikolinearitas pada Kasus Kemiskinan di Provinsi Sulawesi Tenggara Tahun 2023," *Arus Jurnal Sains dan Teknologi,* vol. 3, no. 1, pp. 68-82, 2025, doi: https://doi.org/10.57250/ajst.v3i1.1164.

[14] C.-C. Jeng, "Why a Variance Inflation Factor of 10 Is Not an Ideal Cutoff for Multicollinearity Diagnostics," *Journal of Educational Research and Development*, vol. 57, no. 2, pp.67-92, 2023, doi: https://doi.org/10.53106/199044282023105702004.

[15] F. Chen, and K. Rohe, "A New Basis for Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,* vol. 33, no. 2, pp. 421-434, 2024, doi: https://doi.org/10.1080/10618600.2023.2256502.

[16] S. Widiarto, R. Fauziyah, T. P. Astari, N. L. Juliasih, S. Hadi, L. Zakaria, and I. Saputra, "Authentication of Processed Beef Sausage Products Using Chemometric Analysis Based on FTIR Spectrophotometry Data," *Jurnal Kimia Sains dan Aplikasi,* vol. 28, no. 1, pp. 39-46, 2025.

[17] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, Art. no. 20150202, 2016. doi: 10.1098/rsta.2015.0202.

[18] J. Purwanto, T. Rusolono, and L. B. Prasetyo, "Spatial model of deforestation in Kalimantan from 2000 to 2013," *Jurnal Manajemen Hutan Tropika*, vol. 21, no. 3, pp. 110–118, 2015, doi: 10.7226/jtfm.21.3.110.

[19] M. F. Barri et al., Papua Bioregion: The Forest and Its People. Bogor, Indonesia: Forest Watch Indonesia, 2019. [Online]. Available: https://fwi.or.id/wp-content/uploads/2020/06/FWI-2019-Papua-Bioregion-The-Forest-and-Its-People.pdf

[20] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006, doi: 10.1198/106186006X113430

DOI: https://doi.org/10.34123/jurnalasks.v18i1.971

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Examining the Local Effects of Food Security Index Components Across Kalimantan Using Geographically Weighted Regression

Meirinda Fauziyah[1*], Raditya Arya Kosasih[2], Ayu Bahriah[3], Suyitno[4], Andrea Tri Rian Dani[5]

[1,2,3,4]Statistics Study Program, Department of Mathematics, Faculty of Mathematical and Natural Sciences, Mulawarman University, Samarinda, Indonesia, [5]Doctoral Study Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Airlangga University, Indonesia
**Corresponding Author: E-mail address: meirindafauziyah@fmipa.unmul.ac.id*

ARTICLE INFO

Article history:
Received 4 May, 2026
Revised 12 June, 2026
Accepted 26 June, 2026
Published 30 June, 2026

Keywords:
Adaptive Gaussian; Food Security Index; Geographically Weighted Regression; Haversine Distance; Poverty

Abstract

Introduction/Main Objectives: Food security remains a critical concern across Kalimantan Island, where substantial spatial disparities exist among its 56 regencies and cities, making conventional global regression models inadequate for capturing localized differences. **Background Problems:** This study addresses the limitation of Multiple Linear Regression in accounting for spatial heterogeneity in the relationships between Food Security Index components and the overall index, raising the question of which components exhibit spatially varying local effects across locations. **Novelty:** This study presents the first spatially explicit analysis of food security determinants at the regency and city level across Kalimantan, employing Haversine distance combined with adaptive Gaussian kernel weighting within GWR a combination not previously applied in this context. **Research Methods:** GWR was applied to cross-sectional 2024 data from the Food Security and Vulnerability Atlas, incorporating Cross Validation bandwidth selection and Weighted Least Squares parameter estimation. **Finding/Results:** The GWR model outperformed MLR with an R^2 of 59.63% and MSE of 38.5241. The ratio of population per health worker and average years of schooling for women were the most spatially dominant components, significant in 45 and 43 locations respectively, supporting the need for location-specific policy interventions across Kalimantan.

1. Introduction

Food security is defined as a condition in which food needs are adequately met at both individual and national levels, characterized by the availability of sufficient, safe, diverse, and nutritious food that is fairly distributed, affordable, and aligned with social, cultural, and religious values, thereby enabling people to lead healthy, active, and productive lives in a sustainable manner [1]. Food security stands as one of the central objectives of the Sustainable Development Goals (SDGs), most notably under Goal 2, which is dedicated to eliminating hunger worldwide. In Indonesia, it has become a major strategic focus of the government, with President Prabowo Subianto highlighting the importance of achieving national food self-sufficiency during his administration [2].

Data published by the National Food Agency [3] show that the Food Security Index across provinces in Kalimantan Island is generally classified as food secure or highly food secure. However, at the regency and city level, notable disparities remain. Of the 56 regencies and cities in Kalimantan, six areas including Murung Raya, Mahakam Ulu, Gunung Mas, Lamandau, and Melawi fall into the vulnerable or moderately vulnerable categories, representing approximately 10.7% of all administrative units in the region. Moreover, the food security index ranges from category 2 (highly vulnerable) to category 6 (highly food secure), indicating substantial spatial variation across the island. Geographically, the more vulnerable areas tend to be concentrated in remote, inland regions, suggesting that location-specific characteristics may play a significant role in determining food security outcomes. This spatial heterogeneity implies that the relationships between food security and its determinants may not be uniform across regions some factors may exert stronger or even opposing influences in different localities underscoring the need for an analytical approach that can capture such local variation [4].

The Food Security Index has been widely studied, including research by Evalia et al. [5], which identified determinants of food security in West Sumatra using Multiple Linear Regression. However, MLR assumes that regression relationships are spatially stationary that is, constant regardless of location [6] an assumption that may be unreasonable in a geographically diverse region such as Kalimantan, where socioeconomic and geographic characteristics vary substantially across districts and cities.

GWR addresses this limitation by incorporating geographic coordinates through spatial weighting, allowing regression coefficients to vary across locations and generating local models for each observation point [7]. This makes GWR particularly suitable for detecting spatial variation in the relationships between food security and its determinants. A similar approach was applied by [8], who modelled the food security index in East Java using GWR and found that Rice Production, Adjusted Per Capita Expenditure, Poverty Line, Number of People in Poverty, and Prevalence of Food Inadequacy significantly influenced the food security index in East Java, with spatially varying coefficients. Building on this, the present study extends the application of GWR to Kalimantan Island, a region with distinct geographic and demographic characteristics that have not been previously examined under this framework.

Despite growing interest in spatial approaches to food security analysis, studies applying GWR in the context of Kalimantan Island remain absent from the literature. Previous studies have either relied on global regression models that cannot account for spatial heterogeneity, or have focused on other regions of Indonesia such as Java and Sumatra. This study therefore contributes to the literature by providing the first spatially explicit analysis of food security determinants at the regency and city level across Kalimantan. Furthermore, this study employs Haversine distance as the spatial distance measure, which accounts for the curvature of the earth and has been shown to outperform Euclidean and Manhattan distance measures in terms of accuracy [9], combined with an adaptive Gaussian kernel weighting scheme a combination that has not been previously applied in the context of food security analysis in Kalimantan.

Given the importance of spatial variation in analyzing food security, this study employs GWR to examine the local effects of food security index components at the regency and city level throughout Kalimantan Island. Specifically, this study aims to identify which components of the Food Security Index exhibit spatially varying effects across locations, to examine how the direction and magnitude of these effects differ across districts and cities, and to evaluate the performance of GWR with Haversine distance and adaptive Gaussian kernel weighting for this geographic context. The findings are expected to provide a more spatially nuanced understanding of food security conditions across Kalimantan and to support the development of more targeted and location-specific policy interventions.

2. Material and Methods

2.1. Research Data

This study utilizes cross-sectional data for the year 2024, as the primary focus is on capturing the current spatial distribution of food security conditions across Kalimantan to inform present policy needs. It is acknowledged that cross-sectional data cannot account for temporal dynamics in the relationships among variables, which constitutes a limitation of the present study.

This study examines the Food Security Index (Y) and its eight official component indicators as published in the Food Security and Vulnerability Atlas (FSVA) by the National Food Agency [3]. These components are percentage of population below the poverty line (X_1), percentage of households with food expenditure proportion exceeding 65% of total expenditure (X_2), percentage of households

without access to electricity (X_3), average years of schooling for women aged 15 years and above (X_4), percentage of households without access to clean water (X_5), ratio of population per health worker relative to population density (X_6), percentage of stunted children under five (X_7), and life expectancy at birth (X_8). Rather than treating these components as external determinants of the Food Security Index, this study investigates how each component contributes locally to the overall index, and whether the magnitude and direction of these contributions vary across regencies and cities in Kalimantan. This approach is motivated by the recognition that food security conditions are spatially heterogeneous, and that the relative importance of each component may differ substantially across locations depending on local geographic, demographic, and socioeconomic characteristics [6].

2.2. Multicollinearity Detection

Multicollinearity detection is performed to identify whether interdependencies exist among predictor variables within a regression model. The existence of such dependencies may distort the estimated relationship between predictor variables and the response variable. A widely used measure for assessing multicollinearity is the Variance Inflation Factor (VIF). A predictor variable is regarded as free from multicollinearity when its VIF value does not exceed 10. The formula used to compute the VIF is given as follows [10]:

$$VIF_k = \frac{1}{1 - R_k^2}. \tag{1}$$

where, R_k^2 is the coefficient of determination of the MLR model with $\mathbf{x}_j$ as the response variable and the remaining $\mathbf{x}_i$ as the predictor variables.

2.3. Multiple Linear Regression

Linear regression is a statistical method employed to examine the linear relationship between one or more predictor variables and a response variable. Multiple Linear Regression (MLR) involves the use of two or more predictor variables simultaneously. The standard formulation of the MLR model, which captures the influence of predictor variables on the response variable, can be written as follows [11]:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{2}$$

Where,

y_i : observed value of the response variable for the *i*-th observation
β_0 : intercept (constant) parameter
β_k : regression coefficient associated with the *k*-th predictor variable
x_{ik} : observed value of the *k*-th predictor variable for the *i*-th observation
ε_i : error term for the *i*-th observation

The general model presented in Equation (2) can also be reformulated in matrix notation as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3}$$

2.4. Parameter Estimation of the MLR

Parameter estimation in the MLR model is conducted through the Ordinary Least Squares (OLS) method. This approach determines the model parameters by minimizing the total sum of squared residuals. The estimator for the parameter $\boldsymbol{\beta}$ is derived as follows [12]:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{4}$$

2.5. Simultaneous Significance Testing of MLR Parameters

The simultaneous significance test of MLR model parameters is intended to assess whether the predictor variables collectively exert a significant influence on the response variable. The hypotheses for the simultaneous significance testing of the MLR model parameters are stated as follows:

$H_0 : \beta_1 = \beta_2 = \mathrm{K} = \beta_p = 0$
H_1 : at least one $\beta_k \neq 0, k = 1, 2, \mathrm{K}, p$

The test statistic employed in this test is the F_1 statistic, which is formulated as follows:

$$F_1 = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 \Big/ p}{\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2 \Big/ n-p-1}. \quad (5)$$

Where, n is the number of observation and p is the number of predictor variables. At a given significance level α, the decision criterion for the simultaneous test states that H_0 is rejected when $F_1 > F_{(1-\alpha;p;n-p-1)}$ or equivalently, H_0 is rejected whenever $p_{value} < \alpha$ [13].

2.6. Partial Significance Testing of MLR Parameters

The partial significance test in the MLR model is performed to identify which individual predictor variables have a significant effect on the response variable. The hypotheses for this test are expressed as follows:

$H_0 : \beta_k = 0, k = 1,2,\text{K}\ , p$

$H_1 : \beta_k \neq 0, k = 1,2,\text{K}\ , p$

This test utilizes the t_1 statistic, which is defined as follows:

$$t_1 = \frac{\hat{\beta}_k}{SE\left(\hat{\beta}_k\right)}, \quad (6)$$

Where, $\hat{\beta}_k$ is the estimated regression coefficient and $SE\left(\hat{\beta}_k\right)$ is the standard error of $\hat{\beta}_k$. At a given significance level α, the decision criterion for the partial test states that H_0 is rejected when $|t_1| > t_{(\alpha/2;n-k-1)}$ or equivalently, H_0 is rejected whenever the $p_{value} < \alpha$ [14].

2.7. Homoscedasticity Test

The homoscedasticity test is conducted to identify whether there is variation in the error variance within a linear regression model. Such variation, known as heteroscedasticity, occurs when the error variance differs across observations [15]. One of the widely adopted methods for detecting heteroscedasticity is the Glejser test. This test follows the framework of simultaneous parameter testing in multiple linear regression, in which the absolute values of the residuals serve as the response variable [16]. The hypotheses for the homoscedasticity test are formulated as follows [17]:

$H_0 : \sigma_1^2 = \sigma_2^2 = \text{K} = \sigma_n^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2;\ i = 1,2,\text{K}\ , n$

The test statistic utilized in this test is the G statistic, which is formulated as follows:

$$G = \frac{MSR}{MSE} = \frac{\left(\boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{\varepsilon} - n\bar{\varepsilon}\right)/\ p}{\left(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{e}\right)/\left(n-p-1\right)}. \quad (7)$$

At a significance level the decision rule for the homoscedasticity test is to reject H_0 if $G > F_{(1-\alpha;p;n-p-1)}$ or equivalently, to reject H_0 if and only if the $p_{value} < \alpha$.

2.8. Moran's Index

Moran's Index serves as a widely adopted measure of global spatial autocorrelation, enabling the assessment of how similarly distributed a variable is among geographically neighboring areas. Through this measure, the degree of spatial randomness in the data can be evaluated, where departures from randomness may reveal underlying spatial structures such as geographic clustering or directional trends across the study area. The value of Moran's I ranges from -1 to 1, where values close to 1 indicate positive spatial autocorrelation (clustering), values close to -1 indicate negative spatial autocorrelation (dispersion), and values close to 0 indicate spatial randomness [18]. Moran's Index can be formulated as follows:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - \bar{y}\right)^2} \quad (8)$$

Hypothesis testing for Moran's Index is as follows [19]:

i. Hypothesis used for parameter testing:

$H_0 : I = 0$

$H_1 : I \neq 0$

ii. The test statistic for Moran's Index:

$$Z(I) = \frac{I - E(I)}{\sqrt{\text{var}(I)}} \quad (9)$$

Where:

$$E(I) = l_0 \frac{1}{n-1}$$

$$\text{var}(I) = \frac{n^2 s_1 - n s_1 + 3 s_0^2}{\left(n^2 - 1\right) s_0^2} - E(I)$$

$$s_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$$

$$s_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2$$

$$s_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} ji_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2$$

iii. Decision criterion: H_0 is rejected if $\left|Z(I)\right| > Z_{\frac{\alpha}{2}}$ or $p - value < \alpha$.

2.9. Geographically Weighted Regression Model

Geographically Weighted Regression (GWR) is a spatial statistical approach designed to examine the relationship between predictor variables and a response variable by accounting for spatial effects. This method is especially valuable for capturing spatial heterogeneity, in which the influence of predictor variables differs from one location to another. In GWR, spatial heterogeneity is reflected through the use of spatial weighting. Because each area may have unique characteristics, GWR produces local parameter estimates for every observation point [7]. The GWR model at the *i*-th location can be written as follows:

$$y_i = \beta_0\left(u_i, v_i\right) + \sum_{k=1}^{p} \beta_k\left(u_i, v_i\right) x_{ik} + \varepsilon_i, i = 1, 2, ..., n, \quad (10)$$

The GWR model shown in Equation (8) can likewise be expressed in matrix form as follows [20]:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}\left(u_i, v_i\right) + \varepsilon_i \quad (11)$$

2.10. Gaussian Kernel Weighting Function

Parameter estimation in the GWR model involves spatial weighting to generate local models. Spatial weights represent the influence of one location on another. In GWR, spatial weights can be calculated using kernel weighting functions, one of which is the adaptive gaussian kernel weighting function, expressed as follows [21].

$$w_{ij}(u_i, v_i) = \exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{h_i}\right)^2\right], \quad i, j = 1, 2, ..., n, \tag{12}$$

where w_{ij} denotes the spatial weight attributed to the *j*-th observation within the GWR model at a particular location, and h_i refers to the smoothing parameter or bandwidth for the *i*-th observation. The adaptive Gaussian kernel is selected over the fixed Gaussian kernel due to the uneven spatial distribution of regencies and cities across Kalimantan Island. Unlike fixed kernel approaches that apply a constant bandwidth across all locations, the adaptive kernel adjusts the bandwidth according to local data density narrowing in areas with closely spaced observations and expanding in sparsely distributed regions. This ensures that each local model is estimated using a sufficient and contextually appropriate set of neighboring observations, producing more reliable parameter estimates across all locations regardless of their geographic isolation [7]. The term d_{ij} refers to the distance between observation points, which is computed using the Haversine distance as defined by the following equation [22]:

$$d_{ij} = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{u_j - u_i}{2}\right) + \cos(u_i)\cos(u_j)\sin^2\left(\frac{v_j - v_i}{2}\right)}\right) \tag{13}$$

where, u_i is the longitude coordinate of the *i*-th observation, v_i is the latitude coordinate of the *i*-th observation and r is the earth radius (6,371 km). Unlike Euclidean distance, which assumes a flat surface, Haversine distance accounts for the curvature of the earth by treating it as a sphere with a radius of 6,371 km [23]. This is particularly important for a geographically extensive region such as Kalimantan, where straight-line distance calculations on a flat plane would introduce systematic errors. Bandwidth selection is crucial for obtaining optimal parameter estimates. The optimum bandwidth value can be determined using the Cross Validation (CV) criterion. One advantage of CV is that it considers the balance between prediction error and model complexity. The CV criterion is expressed as follows:

$$CV = \sum_{i=1}^{n}\left[y_i - \hat{y}_{-i}(h_i)\right]^2. \tag{14}$$

Where, $\hat{y}_{-i}(h_i)$ is the predicted value at the *i*-th location obtained from a model fitted without using the *i*-th observation and calculated using bandwidth h_i [24].

2.11. Estimation of GWR Model Parameters

In this research, parameter estimation is performed using the Weighted Least Squares (WLS) method. Similar to OLS, WLS operates by minimizing the sum of squared errors, but additionally incorporates a spatial weighting matrix into the estimation process. The estimator of $\hat{\boldsymbol{\beta}}(u_i, v_i)$ for the *i*-th location is expressed as follows [20]:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}(u_i, v_i)\mathbf{Y}. \tag{15}$$

2.12. Model Adequacy Test for GWR

The model adequacy test is performed to assess whether there is a significant difference between the MLR model and the GWR model. The corresponding hypotheses are formulated as follows:

$H_0 : \beta_k(u_i, v_i) = \beta_k, k = 1, 2, ..., p$ and $i = 1, 2, ..., n$

$H_1 : \beta_k(u_i, v_i) \neq \beta_k, k = 1, 2, ...p$ and $i = 1, 2, ..., n$

The test statistic for assessing the model adequacy is

$$F^* = \left(\frac{SSE_{MLR} - SSE_{GWR}}{\tau_1}\right) \Big/ \left(\frac{SSE_{GWR}}{\delta_1}\right), \tag{16}$$

The rejection region for the model adequacy test at a significance level of α is defined as rejecting H_0 when $F^* > F_{1-\alpha;df_1;df_2}$, where $df_1 = \tau_1^2/\tau_2$ and $df_2 = \delta_1^2/\delta_2$, or equivalently, H_0 is rejected whenever the $p_{value} < \alpha$ [24].

2.13. Simultaneous Significance Test of GWR Model Parameters

The simultaneous significance test of the GWR model parameters is conducted to assess whether the parameters, considered jointly, have a significant effect on the response variable. The hypotheses are formulated as follows:

$H_0 : \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = ... = \beta_p(u_i, v_i) = 0$

H_1 : atleast one $\beta_k(u_i, v_i) \neq 0, k = 0,1,2,...p$ and $i = 1,2,...,n$

The test statistic for the simultaneous significance test of the GWR model is defined as follows:

$$F_2 = \frac{SSE(X_0)}{df_1(X_0)} \Big/ \frac{SSE(GWR)}{df_2(GWR)}. \quad (17)$$

Where, $SSE(X_0)$ is the sum square error of the GWR model without predictor variables. The critical region for the simultaneous test at a significance level of α is defined as rejecting H_0 if $F_2 > F_{(1-\alpha;n-1;\delta_1^2/\delta_2)}$, or equivalently, rejecting H_0 if $p_{value} < \alpha$ [24].

2.14. Partial Significance Test of GWR Model Parameters

The partial significance test is used to identify which parameters have a statistically significant effect on the response variable. The hypotheses are formulated as follows:

$H_0 : \beta_k(u_i, v_i) = 0,\ k = 1,2,...,p$ and $i = 1,2,...,n$

$H_1 : \beta_k(u_i, v_i) \neq 0,\ k = 1,2,...,p$ and $i = 1,2,...,n$

The test statistic for the partial significance test of the GWR model is defined as follows:

$$t_2 = \frac{\hat{\beta}_k(u_i, v_i)}{\sigma_{GWR}\sqrt{c_{kk}}}. \quad (18)$$

The critical region for the partial test at a significance level of α is defined as rejecting H_0 if the test statistic exceeds the critical value, or equivalently, rejecting H_0 if the $p_{value} < \alpha$ [24].

2.15. Model Evaluation

The coefficient of determination (R^2) is a measure employed to assess model performance by indicating the proportion of variability in the response variable that can be accounted for by the fitted model. It ranks among the most frequently used goodness-of-fit measures in regression analysis, owing to its straightforward interpretation and its capacity to reflect how well the model and its predictor variables explain the response variable. The coefficient of determination can be formulated as follows [25]:

$$R^2 = \left(1 - \frac{SSE}{SST}\right) \times 100\%. \quad (19)$$

Mean Square Error (MSE) is the average of the squared differences between the predicted and actual observed values. It is widely used to measure the magnitude of estimation errors produced by a model. A lower MSE value, particularly one approaching zero, signifies that the model's predictions are in close agreement with the actual observations. The MSE can be written as follows [26]:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (20)$$

3. Results and Discussion

3.1. Data Description

The data are summarized through descriptive statistics, encompassing the mean, minimum, maximum, and standard deviation. These statistical measures were computed using R software, and the corresponding results are presented in Table 1.

Table 1. Descriptive statistics

Variable	Mean	Minimum	Maximum	Standard Deviation
Y	75.97	51.29	91.23	9.8572
X_1	5.8891	2.31	11.38	2.1422
X_2	18.4605	2.22	43.54	10.0689
X_3	1.1488	0	6.43	1.7808
X_4	30.3712	0.01	92.51	21.3225
X_5	8.7654	6.58	11.69	1.1170
X_6	8.6057	0.02	43.54	9.2876
X_7	71.1793	64.97	6.43	2.3129
X_8	22.3089	0	92.51	8.1652

Based on Table 1, the predicted food security index in Kalimantan Island is 75.97 with a standard deviation of 9.8572. The lowest food security index is 51.29 in Murung Raya Regency, while the highest index is 91.23 in Balikpapan City. The summary statistics of the other research variables are also presented in Table 1. The spatial distribution of the food security index across regencies and cities in Kalimantan Island is further depicted through a thematic map, as presented in Figure 1.

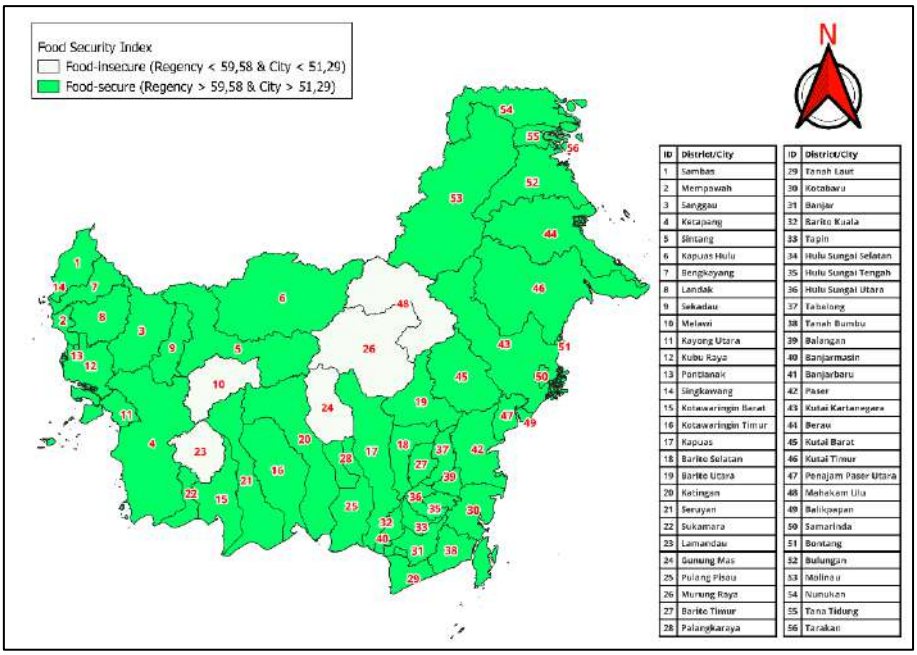


Figure 1. Spatial distribution of the Food Security Index across regencies and cities in Kalimantan Island

Figure 1 presents a map of Kalimantan Island, one of the major islands in Indonesia. The island consists of 56 regencies/cities. The regencies/cities in Figure 1 are classified into two categories, namely food-insecure areas, indicated by white color, and food-secure areas, indicated by green color. The classification of food security status is obtained from the National Food Agency (Badan Pangan Nasional).

Based on Figure 1, the regencies/cities in Kalimantan Island that fall into the food-insecure category include Murung Raya Regency, Melawi Regency, Mahakam Ulu Regency, Gunung Mas Regency, and Lamandau Regency. Information regarding regencies/cities with food security indices below the national average is expected to serve as a basis for government policy formulation and targeted intervention programs aimed at improving food security in these regions.

3.2. Detection of Multicollinearity

Multicollinearity is evaluated through the Variance Inflation Factor (VIF) following Equation (1). The VIF values are computed using R software, and the corresponding results are displayed in Table 2.

Table 2. Detection of Multicollinearity

Variable	VIF
X_1	1.6467
X_2	2.5913
X_3	1.7463
X_4	2.0714
X_5	2.4547
X_6	1.7069
X_7	2.3329
X_8	1.1991

Table 2 reveals that the VIF values for all predictor variables fall below 10. Consequently, it can be concluded that no multicollinearity exists among the predictor variables, and all eight predictor variables can be retained in the linear regression model.

3.3. MLR Modeling

The general MLR model for food security index data, based on Equation (2), is given as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \beta_7 x_{7i} + \beta_8 x_{8i} + \varepsilon_i,\ i = 1,2,...,56 \quad (21)$$

The parameters of the MLR model in Equation (21) are estimated using the OLS approach, as given in Equation (4). Parameter estimation is performed using R software, resulting in the MLR model in Equation (22).

$$\hat{y}_i = 67,7912 - 0,2290x_{i1} + 0,1413x_{i2} - 1,1432x_{i3} - 0,1575x_{i4} - 0,0137x_{i5} - 0,4084x_{i6} + 0,2296x_{i7} + 0,0139x_{i8} \quad (22)$$

The hypotheses for the simultaneous significance testing of the MLR model parameters are stated as follows:

H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1 : at least one $\beta_k \neq 0$; $k = 1,2,3,4,5$

The test statistic F_1 is defined in Equation (5), and the results derived from computations using R software are summarized in Table 3.

Table 3. Results of the simultaneous parameter test of the Multiple Linear Regression (MLR) Model

F_1	$F_{(0.9,8,47)}$	p_{value}	Decision
3.4596	1.8046	0.0033	H_0 is rejected

Table 3 shows that the calculated F_1 statistic satisfies $F_1 = 3.4596 > F_{(0.9,8,47)} = 1.8046$ and $p_{value} = 0.0033 < \alpha = 0.1$. Therefore, it can be concluded that the predictor variables collectively exert a statistically significant influence on the food security index. Next, a partial significance test of the MLR model parameters is carried out using the following hypotheses:

H_0 : $\beta_k = 0$; $k = 1,2,3,4,5$

H_1 : $\beta_k \neq 0$; $k = 1,2,3,4,5$

The test statistic t_1 is defined in Equation (6), and the results of the calculations using R software are presented in Table 4.

As shown in Table 4, average years of schooling for women aged 15 years and above and ratio of population per health worker relative to population density are found to have a statistically significant effect on the food security index. This is evidenced by the test statistics satisfying $|t_1| > t_{(0.95,47)} = 1.6779$ and $p_{value} < \alpha = 0.1$.

Table 4. Results of the partial parameter test of the Multiple Linear Regression (MLR) Model

Parameter	$\|t_1\|$	p_{value}	Decision
β_1	0,335	0,7390	H_0 is not rejected
β_2	0,775	0,4423	H_0 is not rejected
β_3	1,351	0,1833	H_0 is not rejected
β_4	2,045	0,0464	H_0 is rejected
β_5	0,009	0,9932	H_0 is not rejected
β_6	2,545	0,0143	H_0 is rejected
β_7	0,305	0,7619	H_0 is not rejected
β_8	0,091	0,9279	H_0 is not rejected

3.4. Homoscedasticity Test

The homoscedasticity assumption is tested using the Glejser test as a prerequisite for proceeding to spatial modeling. The hypotheses of the Glejser test are formulated as follows.

H_0 : $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_{56}^2 = \sigma^2$

H_1 : at least one $\sigma_i^2 \neq \sigma^2$; $i = 1,2,\text{K},56$

The Glejser test statistic is calculated based on Equation (7), and the corresponding results obtained from R software are shown in Table 5.

Table 5. Results of the Glejser Test

G	$F_{(0.9,8,47)}$	p_{value}	Decision
2,571	1.8046	0,0205	H_0 is rejected

Table 5 shows that the calculated statistic $G = 2.571 > F_{(0.9,8,47)} = 1.8046$ and $p_{value} = 0.0205 < \alpha = 0.1$. Therefore, it can be concluded that heterogeneity exists in the error terms. The existence of heterogeneity in the error terms suggests that the response variable also exhibits heterogeneous behavior. Therefore, spatial analysis is conducted using the GWR model.

3.5. MLR Spatial Autocorrelation Test

The Moran's I test conducted on the MLR residuals yielded a value of -0.0009 with a p_{value} of 0.4239, indicating that the null hypothesis of no spatial autocorrelation cannot be rejected. While this suggests that the residuals are spatially random, the absence of spatial autocorrelation does not preclude the presence of spatial heterogeneity in the relationships between variables a phenomenon that global regression models such as MLR are unable to capture [6]. This provides further justification for the application of GWR in the subsequent analysis.

3.6. Geographically Weighted Regression (GWR) Modeling

The GWR model formulated for the food security index data, derived from Equation (10), is expressed as follows:

$$y_i = \beta_0(u_i,v_i) + \beta_1(u_i,v_i)x_{i1} + \beta_2(u_i,v_i)x_{i2} + \beta_3(u_i,v_i)x_{i3} + \beta_4(u_i,v_i)x_{i4} + \beta_5(u_i,v_i)x_{i5} + \beta_6(u_i,v_i)x_{i6} + \beta_7(u_i,v_i)x_{i7} + \beta_8(u_i,v_i)x_{i8} + \varepsilon_i \;;\; i = 1,2,\ldots,56 \quad (23)$$

The parameter estimation process in the GWR model begins with the calculation of the Haversine distance using Equation (13). Next, the optimal bandwidth is selected through the Gaussian kernel weighting function based on the CV criterion in Equation (14). Once the optimal bandwidth is obtained and the spatial weighting matrix is formed, the GWR parameters are estimated using the WLS method as specified in Equation (15). This process results in local GWR models for each regency/city in Kalimantan Island, producing a total of 56 local models. One example of the estimated GWR models is for Murung Raya Regency. $(i = 27)$.

$$\hat{y}_{27} = 77.4897 - 0.2611x_{27,1} + 0.1108x_{27,2} - 0.9768x_{27,3} - 0.1670x_{27,4} - 0.3322x_{27,5} - 0.4288x_{27,6} + 0.1459x_{27,7} + 0.0156x_{27,8} \quad (24)$$

3.7. Model Adequacy Test for Geographically Weighted Regression (GWR)

The hypotheses for testing the model adequacy of the GWR model against the MLR model can be stated as follows:

H_0 : $\beta_k(u_i,v_i) = \beta_k$; $k = 1,2,\ldots,5$ and $i = 1,2,\text{K},56$

H_1 : $\beta_k(u_i,v_i) \neq \beta_k$; $k = 1,2,\ldots,5$ and $i = 1,2,\text{K},56$

The F^* test statistic is expressed in Equation (16), and the results obtained using R software are presented in Table 6.

Table 6. Results of the GWR Model adequacy test

F^*	$F_{(0.9,9.8430,37.1570)}$	p_{value}	Decision
2.1104	1.7800	0.0493	H_0 is rejected

Table 6 shows that the calculated $F^* = 2.1104 > F_{(0.9,9.8430,37.1570)} = 1.7800$ and the $p_{value} = 0.0493 < \alpha = 0.1$, indicating that the MLR model and the GWR model are not identical. Therefore, the GWR model is appropriate for modeling the food security index across regencies/cities in Kalimantan Island.

3.8. Simultaneous Significance Testing of GWR Model Parameters

The hypotheses for simultaneous significance testing of the GWR model parameters can be stated as follows:

H_0 : $\beta_1(u_1,v_1) = \beta_2(u_1,v_1) = \ldots = \beta_6(u_{56},v_{56}) = 0$

H_1 : at least one $\beta_k(u_i,v_i) \neq 0$; $k = 1,2,3,4,5$ and $i = 1,2,\ldots,56$

The F_2 test statistic is defined in Equation (17), and the corresponding results calculated using R software are shown in Table 7.

Table 7. Results of the simultaneous significance testing of GWR Model parameters

F_2	$F_{(0.9,54.1062,41.8159)}$	p_{value}	Decision
1.8543	1.4667	0.0200	H_0 is rejected

As presented in Table 7, the computed $F_2 = 1.8543 > F_{(0.9,54.1062,41.8159)} = 1.4667$ and the $p_{value} = 0.0200 < \alpha = 0.05$, indicating that the predictor variables collectively exert a statistically significant effect on the response variable.

3.9. Partial Significance Testing of GWR Model Parameters

The hypotheses for the partial significance test of the GWR model parameters are formulated as follows:

H_0 : $\beta_k(u_i,v_i)=0$; $k=1,2,\text{K},5$ and $i=1,2,\text{K}\ 56$

H_1 : $\beta_k(u_i,v_i)\neq 0$; $k=1,2,\text{K},5$ and $i=1,2,\text{K}\ 56$

The t_2 test statistic is formulated in Equation (18), and the outcomes of the partial significance test of the GWR parameters for Murung Raya Regency $(i=27)$ are displayed in Table 8.

Table 8. Results of the partial significance testing of GWR Model parameters

Parameter	$\lvert t_2\rvert$	p_{value}	Decision
$\beta_1(u_{27},v_{27})$	0.4206	0.6762	H_0 is not rejected
$\beta_2(u_{27},v_{27})$	0.6695	0.5069	H_0 is not rejected
$\beta_3(u_{27},v_{27})$	1.2671	0.2121	H_0 is not rejected
$\beta_4(u_{27},v_{27})$	2.3993	0.0210	H_0 is rejected
$\beta_5(u_{27},v_{27})$	0.2288	0.8202	H_0 is not rejected
$\beta_6(u_{27},v_{27})$	2.9485	0.0052	H_0 is rejected
$\beta_7(u_{27},v_{27})$	0.2135	0.8320	H_0 is not rejected
$\beta_8(u_{27},v_{27})$	0.1126	0.9109	H_0 is not rejected

Based on Table 8, it can be concluded that the variables significantly influencing the food security index in Murung Raya Regency $(i=27)$ are average years of schooling for women aged 15 years and above and ratio of population per health worker relative to population density.

The classification of GWR models across all observation locations, based on the significant predictor variables, is illustrated through a thematic map showed in Figure 2.

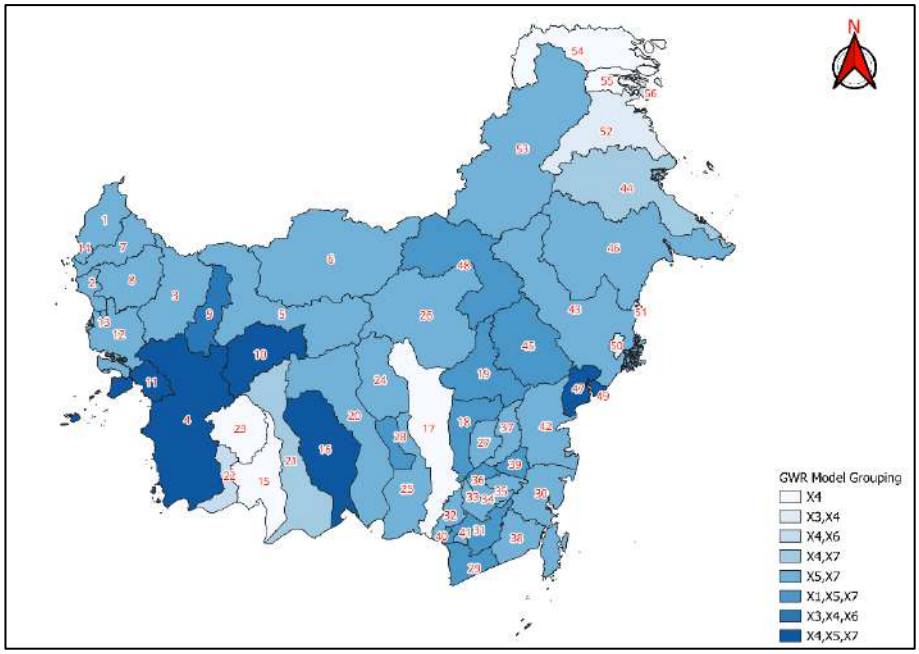


Figure 2. Thematic map of GWR Model grouping

As illustrated in Figure 2, the GWR models for regencies and cities across Kalimantan Island can be classified into eight groups based on the variables that exert a significant influence on the food security index. Each category is indicated by a blue color gradient, ranging from white to dark blue. For instance, regencies/cities colored in dark blue represent GWR models where the significant factors affecting the food security index include average years of schooling for women aged 15 years and above,

percentage of households without access to clean water, and percentage of stunted children under five. The representative model for this group is the GWR model of Ketapang Regency, Melawi Regency, Kayong Utara Regency, Kotawaringin Timur Regency, Penajam Paser Utara Regency, and Balikpapan City.

Three components of the Food Security Index were found to be statistically insignificant across all locations in Kalimantan, namely the percentage of population below the poverty line (X_1), the percentage of stunted children under five (X_7), and life expectancy at birth (X_8), with local coefficients of -0.229, 0.230, and 0.014 and p-values of 0.739, 0.762, and 0.928 respectively.

The non-significance of X_1 may be attributed to its conceptual overlap with X_2, the percentage of households with food expenditure proportion exceeding 65% of total expenditure. Both variables capture dimensions of household economic deprivation, and their high degree of conceptual similarity may have caused the contribution of X_1 to be absorbed by X_2 in the local model estimation, rendering X_1 statistically redundant. This is consistent with the notion that poverty and food expenditure burden are closely interrelated phenomena, particularly in the context of Kalimantan where economic vulnerability tends to manifest simultaneously across multiple dimensions.

The non-significance of X_7 (percentage of stunted children under five) is theoretically interpretable, as stunting is more appropriately regarded as a long-term outcome of chronic food insecurity and malnutrition rather than a direct component that drives the Food Security Index. Its relationship with the index may therefore be indirect and mediated by other variables, reducing its explanatory power in the local regression framework.

Similarly, the non-significance of X_8 (life expectancy at birth) may reflect the fact that life expectancy is a broad health outcome influenced by a wide range of factors beyond food security alone, including healthcare quality, environmental conditions, and genetic factors. Its relationship with the Food Security Index may therefore be too distal and non-specific to yield statistically meaningful local estimates across all regencies and cities in Kalimantan. Nonetheless, the retention of these three components in the model is justified by their status as official indicators of the Food Security and Vulnerability Atlas [3], and their non-significance itself constitutes a meaningful finding that warrants attention in future research.

3.10. GWR Spatial Autocorrelation Test

To validate the GWR model, the Moran's I test was repeated on the GWR residuals, yielding a value of -0.0611 with a p-value of 0.6791. The null hypothesis of no spatial autocorrelation cannot be rejected, confirming that the GWR model has adequately accounted for the spatial structure in the data and that the residuals do not exhibit systematic spatial patterns. The GWR model is therefore considered appropriate for this study.

3.11. Interpretation of the GWR Model

The interpretation is conducted for a single GWR model using only the significant predictor variables. As an example, the GWR model for the observation in Murung Raya Regency $(i=27)$ is presented as follows:

$$\hat{y}_{27} = 77.4897 - 0.2611x_{27,1} + 0.1108x_{27,2} - 0.9768x_{27,3} - 0.1670x_{27,4} - 0.3322x_{27,5} - 0.4288x_{27,6} + 0.1459x_{27,7} + 0.0156x_{27,8} \quad (25)$$

Murung Raya recorded the lowest Food Security Index value in Kalimantan at 51.29, placing it in the highly vulnerable category. The local GWR model for this regency identified two statistically significant components at $\alpha = 0.1$, namely average years of schooling for women aged 15 years and above (X_4) and the ratio of population per health worker relative to population density (X_6), with local coefficients of -0.167 and -0.429 respectively. The negative coefficient of X_6 is consistent with theoretical expectations, suggesting that a higher population-to-health worker ratio in densely populated areas is associated with reduced food security, likely reflecting limited access to health services that support nutritional status and overall household welfare.

However, the negative coefficient of X_4 contradicts the theoretical expectation that higher educational attainment among women would positively contribute to food security. This anomaly may be explained by the phenomenon of educational out-migration, whereby women with higher levels of education tend to migrate to urban centers such as Palangkaraya in search of better employment

opportunities, leaving behind households that are more economically vulnerable and food insecure. Furthermore, in a remote region such as Murung Raya, higher educational attainment does not necessarily translate into improved household income due to the severe scarcity of formal employment opportunities, resulting in a mismatch between education and local labor market conditions.

Among the non-significant components, X_2 (percentage of households with food expenditure proportion exceeding 65% of total expenditure) and X_7 (percentage of stunted children under five) exhibited positive coefficients of 0.111 and 0.146 respectively, contrary to theoretical expectations. The anomalous direction of X_2 may reflect the unique subsistence economy of Murung Raya, where a high proportion of food expenditure does not necessarily indicate food insecurity, as households may supplement their food needs through subsistence farming and forest resources, thereby maintaining food availability despite high expenditure proportions. The positive coefficient of X_7 may be attributed to the endemic nature of stunting in this region, where stunting is more closely associated with long-term sanitation and child-feeding practices rather than current food availability conditions, rendering its relationship with the Food Security Index less straightforward in this specific context.

Similar spatial variation in coefficient signs was also observed across other regencies and cities in Kalimantan, further confirming the presence of spatial heterogeneity in the relationships between food security index components and the overall index, a pattern that would not have been detectable using global regression models.

3.12. Model Evaluation

Model evaluation measures used in this study are the coefficient of determination $\left(R^2\right)$ and the Mean Squared Error (MSE). The obtained values of R^2 and MSE are presented in Table 9.

Table 9. Model evaluation

Model	R^2	MSE
RLB	37.06%	60.0609
GWR	59.63%	38.5241

As shown in Table 9, it can be concluded that the GWR model demonstrates superior performance in modeling the food security index across regencies and cities in Kalimantan Island when compared to the MLR model. This is evidenced by the higher coefficient of determination (R^2) and the lower Mean Squared Error (MSE) of the GWR model relative to the MLR model.

4. Conclusion

This study examined the local effects of Food Security Index components across 56 regencies and cities in Kalimantan Island using Geographically Weighted Regression with Haversine distance and adaptive Gaussian kernel weighting. The results confirm that the GWR model outperforms the MLR model, as evidenced by a higher coefficient of determination $\left(R^2 = 59.63\%\right)$ and a lower Mean Squared Error (MSE = 38.5241), indicating that accounting for spatial heterogeneity substantially improves model fit.

The GWR models across Kalimantan Island can be classified into eight distinct groups based on the components that significantly influence the Food Security Index at each location. Among all components, the ratio of population per health worker relative to population density $\left(X_6\right)$ and average years of schooling for women aged 15 years and above $\left(X_4\right)$ emerge as the most spatially dominant, being significant in 45 and 43 regencies and cities respectively. This finding underscores that access to health services and female educational attainment are the two most critical dimensions of food security across Kalimantan. Three components percentage of population below the poverty line $\left(X_1\right)$, percentage of stunted children under five $\left(X_7\right)$, and life expectancy at birth $\left(X_8\right)$ were not statistically

significant across any location, suggesting that their contributions to the Food Security Index are mediated by other components or reflect long-term outcomes rather than direct drivers.

From a policy perspective, the spatial dominance of (X_6) across 45 regencies and cities suggest that improving the distribution and availability of health workers particularly in remote inland areas should be a priority intervention for strengthening food security across Kalimantan. Special attention should be directed toward the five most vulnerable regencies, namely Murung Raya, Melawi, Mahakam Ulu, Gunung Mas, and Lamandau, where food security conditions remain critically low. In these areas, targeted programs addressing female education, healthcare access, and household economic resilience are recommended, given the unique local characteristics that distinguish these regions from the rest of Kalimantan.

This study has several limitations that should be acknowledged. First, the use of cross-sectional data limits the analysis to spatial variation at a single point in time and does not capture temporal dynamics in food security conditions. Future research is encouraged to employ spatio-temporal approaches such as Geographically and Temporally Weighted Regression (GTWR) when longitudinal data become available. Second, the retention of three statistically non-significant components warrants further investigation into alternative indicators that may better capture the dimensions of food security not explained by the current model. Third, future studies may also consider incorporating additional external variables such as agricultural productivity, infrastructure accessibility, and climate-related factors to provide a more comprehensive understanding of the determinants of food security across Kalimantan.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Meirinda Fauziyah: Conceptualization, Validation, Writing - Review & Editing, Visualization, Supervision. **Raditya Arya Kosasih:** Conceptualization, Methodology, Software, Writing - Original Draft, Visualization. **Ayu Bahriah**: Methodology, Software, Writing - Original Draft. **Suyitno**: Validation, Writing - Review & Editing, Visualization, Supervision. **Andrea Tri Rian Dani**: Writing - Review & Editing, Supervision.

References

[1] Badan Pangan Nasional, *Indeks Ketahanan Pangan*. Jakarta, Indonesia: Badan Pangan Nasional, 2024.

[2] S. Hidayat and A. S. Radyawanto, "Batalyon Infanteri Teritorial Pembangunan TNI AD: Pelopor Ketahanan Pangan Indonesia," *Journal of Industrial Engineering & Management Research*, vol. 6, no. 5, pp. 90–96, 2025.

[3] Badan Pangan Nasional, *Indeks Ketahanan Pangan (IKP) Tingkat Kabupaten/Kota Tahun 2024 (12 Indikator)*, 2025.

[4] M. Syakirotin, T. Karyani, and T. I. Noor, "Model Pengaruh Geographically Weighted Regression dan Strategi Alternatif Ketahanan Pangan Kabupaten Bandung Saat Pandemi COVID-19," in *Prosiding Seminar Nasional Hasil Penelitian Agribisnis VII*, pp. 48–55, 2023.

[5] N. A. Evalia et al., "Faktor-Faktor yang Mempengaruhi Ketahanan Pangan di Sumatera Barat," *AKADEMIK: Jurnal Mahasiswa Humanis*, vol. 5, no. 2, pp. 937–948, 2025.

[6] A. Comber, C. Brunsdon, M. Charlton, G. Dong, R. Harris, B. Lu, Y, Lu, D. Murakami, T. Nayaka, Y. Wang, and P. Harris, "A Route Map for Successful Applications of Geographically Weighted Regression," *Geographical Analysis*, vol. 55, no. 1, pp. 155-178, 2023.

[7] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, UK: John Wiley & Sons, 2002.

[8] T. L. N. Azizah and Suliadi, "Pemodelan Spasial Ketahanan Pangan Jawa Timur Tahun 2023," in *Bandung Conference Series: Statistics,* vol. 5, no. 2, pp. 557-566, 2025.

[9] Y. Miftahuddin, S. Umaroh, and F. R. Karim, "Perbandingan Metode Perhitungan Jarak Euclidean, Haversine, dan Manhattan dalam Penentuan Posisi Karyawan (Studi Kasus : Institut Teknologi Nasional Bandung)," *Jurnal Tekno Insenstif*, vol. 14, no. 2, pp. 69-77, 2020.

[10] Setiawati, "Analisis Pengaruh Kebijakan Dividen terhadap Nilai Perusahaan pada Perusahaan Farmasi di BEI," *JIP: Jurnal Inovasi Penelitian*, vol. 8, no. 8, pp. 1581–1590, 2021.

[11] W. Sulistiyowati and C. C. Astuti, *Statistika Dasar: Konsep dan Aplikasinya*. Sidoarjo, Indonesia: UMSIDA Press, 2017.

[12] A. T. R. Dani, F. B. Putra, V. Ratnasari, and I. N. Budiantara, *BTS (Buku Tentang Statistika): Regresi Parametrik dan Non Parametrik Teori dan Aplikasi dengan Software R*. Tasikmalaya, Indonesia: Perkumpulan Rumah Cemerlang Indonesia, 2024.

[13] R. A. Choerunnisa, R. R. Dewi, M. Bariklana, and E. Widodo, "Analisis Faktor yang Mempengaruhi Tingkat Produksi Jahe di Indonesia Menggunakan Metode Regresi Linier Berganda," *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, vol. 5, no. 2, pp. 231–242, 2021.

[14] A. K. Khotimah, A. A. Rahman, M. Z. Alam, R. Adawiyah, Y. H. Nur, T. R. Aufi, and Sifriyani, "Analisis Regresi Linier Berganda Dalam Estimasi Indeks Pembangunan Manusia di Indonesia," *Jurnal Eksponensial*, vol. 15, no. 2, pp. 90–99, 2024.

[15] N. M. S. Ananda, S. Suyitno, and M. Siringoringo, "Geographically Weighted Panel Regression Modelling of Human Development Index Data in East Kalimantan Province in 2017–2020," *Jurnal Matematika, Statistika dan Komputasi*, vol. 19, no. 2, pp. 323–341, 2023.

[16] A. Miranda, S. Suyitno, and M. Fauziyah, "Modelling the Probability of River Water Pollution Using Geographically Weighted Logistic Regression Model (Case Study: River Water DO Data in East Kalimantan)," *Jurnal Matematika, Statistika dan Komputasi*, vol. 21, no. 2, pp. 408–430, 2025.

[17] S. Novianti, S. Suyitno, and M. Siringoringo, "Model Geographically Weighted Weibull Regression dengan Kriteria Penentuan Bandwidth Optimum Akaike Information Criterion (Studi Kasus: Indikator Pencemaran Air Biochemical Oxygen Demand di Daerah Hutan Tropis Lembab DAS Mahakam Tahun 2019)," in *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya*, pp. 43–68, 2022.

[18] K. D. Lorenza, S. C. Pratiwi, D. Puspita, I. R. A. I, and D. S. Rini, "Penerapan Spatial Autoregressive Model (SAR) untuk Mengetahui Faktor-faktor yang Memengaruhi Indeks Pembangunan Manusia (IPM)," *Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika*, vol. 7, no. 1, pp. 267-279, 2024.

[19] A. Septiani, R. H. Hirzi, and N. U. Fikriah, "Analisis Penyebaran Jumlah Kasus PMK Pada Hewan Ternak Sapi di Kabupaten Lombok Tengah Menggunakan Indeks Moran Tahun 2022," *Variance: Journal of Statistics and Its Application*, vol. 5, no. 2, pp. 159-168, 2023.

[20] R. E. Caraka and H. Yasin, *Geographically Weighted Regression (GWR): Sebuah Pendekatan Regresi Geografis*. Yogyakarta, Indonesia: Mobius, 2017.

[21] S. H. Daulay and E. Simamora, "Pemodelan Faktor-faktor Penyebab Kemiskinan di Provinsi Sumatera Utara Menggunakan Metode Geographically Weighted Regression (GWR)," *Jurnal Riset Rumpun Matematika dan Ilmu Pengetahuan Alam*, vol. 2, no. 1, pp. 47–60, 2023.

[22] A. H. Harahap and M. I. Farza, "Pengaplikasian Rumus Haversine Sebagai Upaya Mengukur Kekuatan Interaksi Antarwilayah Kabupaten Ogan Komering Ilir dan Kota Palembang," *Jurnal Geografi*, vol. 21, no. 1, pp. 90–96, 2025.

[23] V. H. Nabilla, D. Fitria, and F. Fitri, "Comparison of Haversine and Euclidean Distance Formulas for Calculating Distance Between Regencies in West Sumatra," *UNP Journal of Statistics and Data Science*, vol. 1, no. 3, pp. 120-125, 2023.

[24] N. L. A. C. Dewi, M. N. Hayati, and M. Fauziyah, "Pemodelan GWR Menggunakan Fungsi Pembobot Adaptive Box-Car pada Angka Kesakitan DBD di Pulau Kalimantan Tahun 2023," *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 7, no. 1, pp. 55–65, 2025.

[25] N. Y. Adrianingsih, A. T. R. Dani, I. N. Budiantara, D. L. Ull, and R. A. Kosasih, "A Computational Analysis of Kernel-Based Nonparametric Regression Applied to Poverty Data," *Mandalika Mathematics and Education Journal*, vol. 7, no. 3, pp. 1336–1347, 2025.

[26] R. M. Untsa, F. S. Akbar, H. Briantoro, N. Rachmaningrum, and H. U. Mustakim, "Filter Least Mean Square (LMS) untuk Mengurangi Noise pada Sinyal Suara Tembakan," in *CENTIVE Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media*, vol. 3, no. 1, pp. 1–15, 2023.

DOI: https://doi.org/10.34123/jurnalasks.v18i1.972

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Mapping and Modeling Crime Factors in North Sumatra Using GWGPR

Eva Kosasih[1*], Ni Luh Putu Suciptawati[2], Luh Putu Ida Harini[3]

[1,2,3]*Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Udayana, Bali, Indonesia*
**Corresponding Author: E-mail address: kosasih.2208541054@student.unud.ac.id*

ARTICLE INFO

Article history:
Received 4 May, 2026
Revised 15 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Crime Cases; Geographically Weighted Generalized Poisson Regression; North Sumatra; Overdispersion; Spatial Heterogeneity; Spatial Modeling

Abstract

Introduction/Main Objectives: Crime remains a significant social issue influenced by socio-economic factors and exhibiting spatial variation, particularly in North Sumatra Province, which recorded the highest number of criminal cases in Indonesia in 2024. This study aims to identify significant factors affecting crime and examine the spatial variation of their effects across districts/cities in North Sumatra. **Background Problems:** Global regression models often fail to capture crime patterns due to overdispersion and spatial heterogeneity, leading to inconsistent relationships across regions. **Novelty:** This study employs Geographically Weighted Generalized Poisson Regression (GWGPR), which simultaneously addresses overdispersion and spatial heterogeneity, providing a more robust localized analysis than global models. **Research Methods:** Using secondary data from 33 districts/cities in North Sumatra, the variables include population density, open unemployment rate, mean years of schooling, and Gini ratio. The analysis involves Poisson regression, dispersion testing, Generalized Poisson Regression, spatial heterogeneity testing, and GWGPR. **Finding/Results:** The significant factors affecting crime are the open unemployment rate, mean years of schooling, and population density, while the Gini ratio is not significant. **Limitation:** This study is limited by the use of data covering only the year 2024 and a limited set of socio-economic variables, which may not fully capture all factors associated with crime.

1. Introduction

Crime remains a major social problem in both developed and developing countries because it threatens public security, social stability, and community welfare [1]. Various types of crime, such as theft, robbery, fraud, assault, and property-related offenses, continue to occur in many regions. In Indonesia, North Sumatra is one of the provinces with a relatively high crime rate. According to data from the National Criminal Information Center of the Indonesian National Police, North Sumatra recorded 53,897 criminal cases in 2024, the highest number among all provinces in Indonesia [2]. This condition highlights the need for effective policy interventions to address crime in North Sumatra.

Crime is influenced not only by individual behavior but also by socio-economic conditions. From a theoretical perspective, crime can be explained through several socio-economic theories. Social

Disorganization Theory suggests that areas characterized by high population concentration tend to experience weaker informal social control and greater opportunities for criminal activities because increased social interaction and urban concentration may reduce the effectiveness of community supervision [3]. Consistent with this theory, previous studies have reported that population density has a significant effect on crime rates, as densely populated areas increase social interaction and competition for resources [4]. Furthermore, Economic Theory of Crime argues that individuals make rational decisions by comparing the expected benefits and costs of legal and illegal activities. Under unfavorable economic conditions, such as unemployment and limited access to formal employment opportunities, individuals may face stronger incentives to engage in criminal behavior [5]. Consistent with this argument, empirical evidence suggests that open unemployment contributes to increased criminal behavior due to economic pressure and limited job opportunities [6]. Consequently, the open unemployment rate is considered an important explanatory variable.

Human Capital Theory also emphasizes that education improves skills, legal awareness, and employment opportunities, thereby reducing the likelihood of criminal involvement [7]. Consistent with this theory, education level, represented by mean years of schooling, is expected to influence crime rates. As reported in [8], mean years of schooling is associated with lower crime rates, potentially due to improved legal awareness and broader economic opportunities. In addition, Relative Deprivation Theory proposes that income inequality may generate perceptions of social injustice and economic exclusion that encourage criminal behavior [9]. Consistent with this perspective, previous studies have shown that income inequality, measured by the Gini ratio, can lead to social dissatisfaction and increase incentives for criminal behavior [10], [11]. Based on these theoretical arguments and previous findings, population density, open unemployment rate, mean years of schooling, and income inequality are hypothesized to influence crime in this study.

Given the theoretical and empirical evidence discussed above, the selection of population density, open unemployment rate, mean years of schooling, and the Gini ratio is particularly relevant to North Sumatra. These variables represent key demographic, labor market, educational, and inequality dimensions that have been widely associated with crime in previous studies. Since socio-economic conditions differ across regions, the relationships between crime and socio-economic factors may exhibit geographical variation and spatial dependence [12]. Consequently, the effects of these socio-economic factors are unlikely to be uniform across locations, and both the magnitude and direction of their relationships with crime may vary spatially [12], [13]. Therefore, a local modeling approach is more appropriate than a global model for capturing the heterogeneous relationships between crime and the factors influencing it across districts and cities [14]. However, spatial heterogeneity is not the only characteristic that should be considered when modeling crime data.

In addition to spatial heterogeneity, the number of criminal cases also exhibits count-data characteristics that should be considered when selecting an appropriate statistical model. Crime data, specifically the number of criminal cases, are classified as count data. Poisson regression is commonly used for modeling count data. However, it assumes equidispersion, where the variance equals the mean. In practice, crime data often exhibit overdispersion or underdispersion, leading to inefficient estimation and biased statistical inference [15]. Generalized Poisson Regression (GPR) can address this limitation because it includes an additional dispersion parameter [16]. However, GPR is a global model that assumes identical relationships across all regions. This assumption may be unrealistic in practice, as districts and cities in North Sumatra have different socio-economic characteristics, indicating spatial heterogeneity.

To simultaneously accommodate dispersion and spatial heterogeneity, this study applies Geographically Weighted Generalized Poisson Regression (GWGPR). This method enables local parameter estimation for each district and city while handling dispersion in count data. Previous studies have shown that GWGPR performs better than Poisson regression and GPR models in spatial count-data applications [17], [18]. However, no study has specifically applied GWGPR to district-level crime data in North Sumatra. Therefore, this study aims to model criminal cases, identify significant factors influencing crime, and map their spatial variation across districts and cities in North Sumatra in 2024 using GWGPR.

2. Material and Methods

2.1. Poisson Regression

Poisson regression is commonly used for modeling count data, where the response variable is discrete and assumed to follow a Poisson distribution. A discrete random variable Y follows a Poisson distribution with parameter μ if it has the following probability mass function [19]:

$$P(Y = y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}, y = 0,1,2, \ldots \tag{1}$$

Suppose that $Y_i \sim \text{Poisson}(\mu_i)$, for $i = 1,2, \ldots, n$. The Poisson regression model uses the natural logarithm as the link function to relate the expected value of Y_i to explanatory variables x_i. The model can be written as:

$$\ln\big(E(Y_i)\big) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2}$$

or equivalently,

$$E(Y_i) = \mu_i = \exp\big(\mathbf{x}_i^T \boldsymbol{\beta}\big) \tag{3}$$

where $\mathbf{x}_i = [1 \;\; x_{i1} \;\; x_{i2} \;\; \ldots \;\; x_{ik}]^T$ is the predictor vector and $\boldsymbol{\beta} = [\beta_0 \;\; \beta_1 \;\; \beta_2 \;\; \ldots \;\; \beta_k]^T$ is the parameter vector.

Parameter estimation in Poisson regression is performed using Maximum Likelihood Estimation (MLE), with the likelihood function given by:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{4}$$

where $\mu_i = \mathrm{e}^{\mathbf{x}_i^T \boldsymbol{\beta}}$.

2.2. Multicollinearity

Multicollinearity refers to the presence of high correlation among predictor variables, where one predictor variable can be linearly explained by other predictors. The detection of multicollinearity can be carried out using the Variance Inflation Factor (VIF), which is defined as follows [20]:

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1,2, \ldots, k \tag{5}$$

where R_j^2 is the coefficient of determination obtained by regressing the j-th predictor on the remaining predictors. Variables with VIF values greater than 5 indicate multicollinearity.

2.3. Dispersion Test

Poisson regression assumes equidispersion, meaning that the mean equals the variance of the response variable. Violation of this assumption may lead to overdispersion or underdispersion, which can result in inefficient parameter estimation and biased standard errors. To detect dispersion, the deviance and Pearson chi-square statistics are compared with their respective degrees of freedom [21]. The dispersion statistics are defined as follows:

$$\phi_1 = \frac{D}{df}, \qquad D = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right) \tag{6}$$

$$\phi_2 = \frac{\chi^2}{df}, \qquad \chi^2 = \sum_{i=1}^{n}\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i} \tag{7}$$

where $df = n - k - 1$, n is the number of observations, and k is the number of predictor variables. The variable y_i denotes the observed value, while $\hat{\mu}_i$ represents the predicted value. The dispersion condition is assessed based on the value of ϕ. A value of $\phi < 1$ indicates underdispersion, $\phi = 1$ indicates equidispersion, and $\phi > 1$ indicates overdispersion.

2.4. Generalized Poisson Regression (GPR)

Generalized Poisson Regression (GPR) is an extension of the Poisson regression model used to analyze count data that do not satisfy the equidispersion assumption. This model introduces an additional dispersion parameter (ϕ) to accommodate both overdispersion and underdispersion [22]. The probability mass function of the generalized Poisson distribution is given as follows:

$$P(Y_i = y_i) = \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{yi} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left[-\frac{\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right] \quad (8)$$

where $y_i = 0,1,2,\ldots$, $\mu_i = \mathrm{e}^{\mathbf{x}_i^T\boldsymbol{\beta}}$, and ϕ is the dispersion parameter. The mean and variance are given by:

$$E(Y_i) = \mu_i,\ Var(Y_i) = \mu_i(1+\phi\mu_i)^2$$

Parameter estimation in the GPR model is performed using Maximum Likelihood Estimation (MLE). The likelihood function of the GPR model is given by:

$$L(\phi, \boldsymbol{\beta}) = \prod_{i=1}^{n}\left(\frac{\mu_i}{1+\phi\mu_i}\right)^{y_i} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left(-\frac{\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right) \quad (9)$$

Simultaneous testing of model parameters is performed using the Maximum Likelihood Ratio Test (MLRT). The hypotheses are defined as follows [23]:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : at least one $\beta_j \neq 0; j = 1,2,\ldots,k$

The test statistic is given by:

$$D(\hat{\beta}) = -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) = 2\left[\ln L(\hat{\Omega}) - \ln L(\hat{\omega})\right] \quad (10)$$

where $L(\hat{\omega})$ is the likelihood of the restricted model without predictors, and $L(\hat{\Omega})$ is the likelihood of the full model. The decision rule is to reject H_0 if $D(\hat{\beta}) > \chi^2_{(\propto,k)}$.

Partial parameter testing is performed using the Wald test [24]. The hypotheses are:

$H_0 : \beta_j = 0, j = 1,2,\ldots,k$

$H_1 : \beta_j \neq 0, j = 1,2,\ldots,k$

The test statistic is given by:

$$W = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \quad (11)$$

where $\hat{\beta}_j$ is the estimated parameter and $SE(\hat{\beta}_j)$ is its standard error. The decision rule is to reject H_0 if $W > \chi^2_{(\propto;1)}$.

2.5. Spatial Heterogeneity

Spatial heterogeneity is a condition in which the relationships between variables in a regression model vary across observation locations. The presence of spatial heterogeneity is assessed using the Breusch–Pagan (BP) test, with the following hypotheses:

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2,\ i = 1,2,..,n$

The Breusch–Pagan (BP) test statistic is defined as:

$$BP = \left(\frac{1}{2}\right)\mathbf{f}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{f} \quad (12)$$

The elements of vector **f** are defined as $f_i = \left(\frac{\varepsilon_i^2}{\sigma^2} - 1\right)$ where $\varepsilon_i = y_i - \hat{y}_i$, and Z is an $n \times (2k+1)$ matrix consisting of a constant term and interaction terms between predictor variables x_{ij} and spatial coordinates (u_i, v_i), namely $x_{ij}u_i$and $x_{ij}v_i$. The decision rule is to reject H_0 if $BP > \chi^2_{(\alpha,df)}$ where α is the significance level, $df = 2k$ and k is the number of predictors variables.

2.6. Spatial Weight Matrix

Spatial variability across locations is represented through the construction of a spatial weight matrix W. Each element of the matrix is defined based on the Euclidean distance between locations. In GWGPR, the spatial weight matrix is constructed using kernel functions that assign weights to neighboring observations according to their spatial proximity to a target location. Observations located closer to the target location receive larger weights, whereas observations located farther away receive lower weights. The weighting process is controlled by a bandwidth parameter, which determines the spatial extent of neighboring observations involved in local parameter estimation. The choice of bandwidth is important because it affects the balance between local and global information and influences the accuracy of model estimation [14].

Kernel functions can generally be classified into fixed kernels and adaptive kernels. Fixed kernels use a constant bandwidth for all locations, meaning that neighboring observations are weighted based on a fixed geographical distance. Consequently, the number of neighboring observations included in each local model may vary across locations. In contrast, adaptive kernels use a variable bandwidth that adjusts according to the spatial distribution of observations. As a result, the number of neighboring observations remains relatively constant, while the geographical distance represented by the bandwidth may vary across locations. Therefore, the choice of kernel function depends on the spatial characteristics of the data [14]. In the GWGPR model, several commonly used kernel functions are as follows:

a. Fixed Gaussian

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h}\right)^2\right) \tag{13}$$

b. Adaptive Gaussian

$$w_{ij} = \exp\left(-\frac{1}{2}\left(\frac{d_{ij}}{h_j}\right)^2\right) \tag{14}$$

c. Fixed Bisquare

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2, & d_{ij} \le h \\ 0, & d_{ij} > h \end{cases} \tag{15}$$

d. Adaptive Bisquare

$$w_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_j}\right)^2\right)^2, & d_{ij} \le h_j \\ 0, & d_{ij} > h_j \end{cases} \tag{16}$$

The variable d_{ij} represents the Euclidean distance between location i and j, defined as :

$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \tag{17}$$

where (u_i, v_i) denotes the geographical coordinates (latitude and longitude) of location i. The parameter h represents the bandwidth, which determines the spatial range of neighboring observations influencing local parameter estimation. For adaptive kernels, the bandwidth varies across locations and is denoted by h_j. The selection of an optimal bandwidth is crucial because it affects model accuracy as well as the bias–variance trade-off in local estimation [14].

The optimal bandwidth can be determined using the Cross-Validation (CV) method, which is defined as follows [14]:

$$CV = \sum_{i=1}^{n} \left(y_i - \hat{y}_{\neq i}(h)\right)^2 \tag{18}$$

where $\hat{y}_{\neq i}(h)$ is the estimated value of y_i obtained by excluding the observation at location (u_i,v_i), and n is the number of observations. The optimal bandwidth is selected by minimizing the CV value, where a smaller CV value indicates better predictive performance[14].

2.7. Geographically Weighted Generalized Poisson Regression (GWGPR)

Geographically Weighted Generalized Poisson Regression (GWGPR) model is an extension of the Generalized Poisson Regression model that allows parameter estimates to vary across locations. In this model, the regression coefficients are estimated locally for each observation point based on its geographical coordinates (u_i, v_i). The GWGPR model for the i-th location using the log-link function is expressed as:

$$\eta_i = g(\mu_i) = \ln(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)$$
$$\mu_i = g^{-1}(\eta_i) = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)\right) \tag{19}$$

Parameter estimation in the GWGPR model is carried out using Maximum Likelihood Estimation (MLE) method. The likelihood function is given by [10]:

$$L(\boldsymbol{\beta}(u_i, v_i), \phi) = \prod_{i=1}^{n} \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{yi} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left[\frac{-\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right] \tag{20}$$

where $\mu_i = \exp\left(\mathbf{x}_i^T \boldsymbol{\beta}(u_i, v_i)\right)$

Simultaneous testing of model parameters is conducted using the Maximum Likelihood Ratio Test (MLRT), with hypotheses [23]:

$H_0 : \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \cdots = \beta_k(u_i, v_i) = 0$

$H_1 :$ at least one $\beta_j(u_i, v_i) \neq 0;\ j = 1,2,\ldots,k$

The test statistic is defined as:

$$D\left(\hat{\beta}(u_i, v_i)\right) = -2\ln\left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right) = 2\left[\ln L(\hat{\Omega}) - \ln L(\hat{\omega})\right] \tag{21}$$

where $L(\hat{\omega})$ is the likelihood of the restricted model without predictors, and $L(\hat{\Omega})$ is the likelihood of the full model. The decision rule is to reject H_0 if $D(\hat{\beta}) > \chi^2_{(\propto,k)}$.

Partial testing of parameters is conducted using the Wald test [24]. The hypotheses are:

$H_0 : \beta_j(u_i, v_i) = 0,\ j = 1,2,\ldots,k; \quad i = 1,2,\ldots,n$

$H_1 : \beta_j(u_i, v_i) \neq 0,\ j = 1,2,\ldots,k; \quad i = 1,2,\ldots,n$

The test statistic is given by:

$$W = \left(\frac{\hat{\beta}_j(u_i, v_i)}{SE\left(\hat{\beta}_j(u_i, v_i)\right)}\right)^2 \tag{22}$$

where $\hat{\beta}_j(u_i, v_i)$ is the estimated regression coefficient at location (u_i, v_i), and $SE\left(\hat{\beta}_j(u_i, v_i)\right)$ is its standard error. The decision rule is to reject H_0 if $W > \chi^2_{(\propto;1)}$.

2.8. Data Source

This study uses secondary data from 33 districts/cities in North Sumatra Province, Indonesia, consisting of 25 regencies and 8 cities. Crime data for 2024 were obtained from the National Criminal Information Center of the Indonesian National Police (Pusiknas Polri), while predictor variables were obtained from Statistics Indonesia (BPS) of North Sumatra Province. The variables used in this study are summarized in Table 1.

Table 1. Research variables

Variable	Description	Unit
Y	Number of criminal cases	Cases
X_1	Population density	Persons/km^2
X_2	Open unemployment rate	Percent
X_3	Mean years of schooling	Years
X_4	Gini ratio	Index

2.9. Analysis Steps

The data analysis is performed using R and GeoDa software. The analytical procedure is carried out as follows:

1. Descriptive Analysis
   Descriptive statistics are computed to describe the characteristics of criminal cases in North Sumatra Province in 2024, along with the factors suspected of influencing them.
2. Multicollinearity Detection
   Multicollinearity among predictor variables is assessed using the Variance Inflation Factor (VIF) based on equation (5). A predictor variable is considered to exhibit multicollinearity if $VIF > 5$. Variable selection is performed by removing predictors with high VIF values or strong correlations until a set of independent variables is obtained.
3. Poisson regression Modeling
   (a) Estimating the Poisson regression model parameters using the Maximum Likelihood Estimation (MLE) method based on equation (4).
   (b) Conducting dispersion testing using equations (6) and (7) to identify overdispersion or underdispersion. If no dispersion issues are detected, the Poisson regression model is considered adequate; otherwise, the analysis proceeds to Generalized Poisson Regression (GPR).
4. Generalized Poisson Regression Modeling
   (a) Estimating the GPR model parameters using the MLE method based on equation (9).
   (b) Testing parameter significance simultaneously using the Maximum Likelihood Ratio Test (MLRT) (10) and partially using the Wald test (11).
5. Spatial Heterogeneity Testing
   Spatial heterogeneity is tested using the Breusch–Pagan test based on equation (12). If spatial heterogeneity is detected, the analysis proceeds to Geographically Weighted Generalized Poisson Regression (GWGPR) modeling; otherwise, the GPR model is considered sufficient.
6. Geographically Weighted Generalized Poisson Regression Modeling
   (a) Calculating the Euclidean distance between observation locations based on geographic coordinates (latitude and longitude) using equation (17).
   (b) Determining the optimum bandwidth for the kernel weighting function using the Cross-Validation (CV) method (18).
   (c) Constructing the weighting matrix with the selected kernel function using equations (13), (14), (15), and (16).
   (d) Estimating the GWGPR model parameters using the MLE method based on equation (20).
   (e) Testing parameter significance simultaneously using the MLRT (21) and partially using the Wald test (22).

3. Results and Discussion

3.1. Descriptive Analysis

Descriptive statistics of the number of criminal cases and the factors presumed to influence them in North Sumatra Province in 2024 are presented in Table 2, while the spatial distribution of criminal cases is illustrated in Figure 1.

Table 2. Descriptive statistics of research variables

Variable	Min	Max	Mean	Std Dev
Y	13	16,791	1,488.394	2,872.636
X_1	41.16	8,902.16	1,159.908	2,221.838
X_2	0.89	8.13	4.358	2.372
X_3	6.4	11.82	9.503	1.322
X_4	0.206	0.356	0.259	0.040

Based on Table 2, the average number of criminal cases across districts/cities in North Sumatra in 2024 is 1,488.39, with a large standard deviation (2,872.64) and a wide range from 13 to 16,791 cases, indicating substantial variation across regions. This pattern is evident in Figure 1, which shows that Medan City records the highest number of cases, while Gunungsitoli City has the lowest. In particular, criminal cases are concentrated in the administrative and economic center of the province, especially in Medan City, as reflected by the darker orange color, while variations in color gradation suggest that crime is not evenly distributed across districts/cities in North Sumatra.

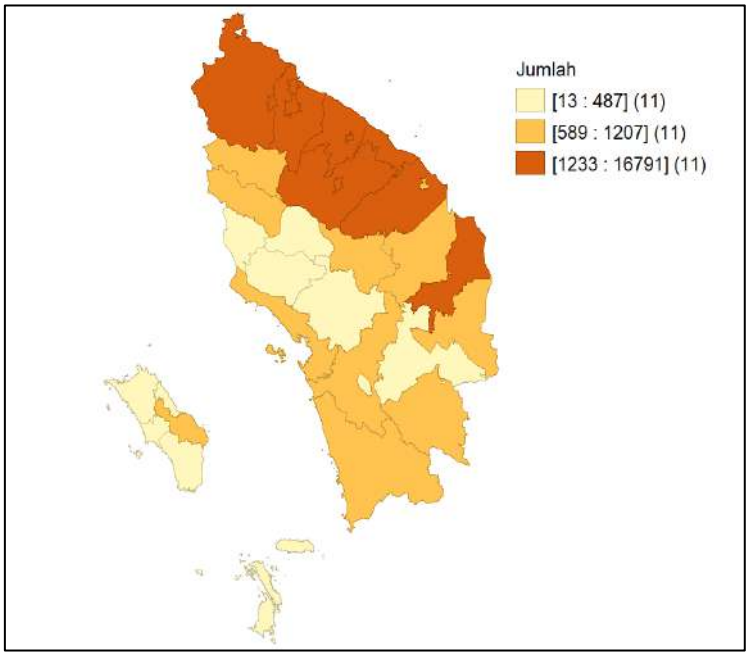


Figure 1. Spatial distribution of the number of criminal cases in North Sumatra

3.2. Multicollinearity Testing

Multicollinearity among predictor variables is assessed using the Variance Inflation Factor (VIF), with a threshold of $VIF < 5$. The results are presented in Table 3.

Table 3. VIF values of predictor variables

Predictor Variable (X)	X_1	X_2	X_3	X_4
VIF Values	4.18	2.46	1.90	3.26

The VIF values in Table 3 indicate that all predictor variables have values below 5. This suggests that there is no multicollinearity among the predictors, and all variables can be included in the subsequent modeling analysis.

3.3. Poisson Regression Modeling

An initial analysis was conducted using Poisson regression, since crime data is count data. Based on the estimated parameters, the Poisson regression model is expressed as follows:

$$\hat{\mu} = \exp(4.286 + 0.0001193X_1 + 0.2577X_2 + 0.007367X_3 + 4.772X_4) \quad (24)$$

The Poisson regression model assumes equidispersion, meaning that the mean equals the variance. To assess this assumption, dispersion testing was performed by comparing the deviance and Pearson chi-square statistics to their respective degrees of freedom. The results are presented in Table 4.

Table 4. Dispersion test result for Poisson Regression Model

Method	Statistic	df	Ratio (Statistic/df)
Deviance	29,126.71	28	1,040.24
Pearson Chi-Squares	26,274.17	28	938.36

As shown in Table 4, both the deviance ratio and the Pearson chi-square ratio are substantially greater than 1. This indicates the presence of overdispersion, where the variance exceeds the mean. Consequently, the Poisson model is not appropriate for this data, and an alternative model that can accommodate overdispersion, such as Generalized Poisson Regression (GPR), is required.

3.4. Generalized Poisson Regression Modeling

Following the detection of overdispersion in the Poisson model, the analysis proceeds with Generalized Poisson Regression (GPR). Using the Maximum Likelihood Estimation (MLE) method, the model parameters were estimated, resulting in the following equation:

$$\hat{\mu} = \exp(5.880 + 0.00001348X_1 + 0.1778X_2 + 0.2753X_3 - 8.171X_4) \quad (25)$$

To evaluate the reliability of the model, parameter significance was assessed both simultaneously and partially. The simultaneous test was conducted using the Maximum Likelihood Ratio Test (MLRT) with the following hypotheses:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
$H_1 :$ at least one $\beta_j \neq 0; j = 1,2,3,4$

The test produced a deviance value of 21.0627, which exceeds the critical chi-square value of $\chi^2_{(0,05;4)} = 9.4877$. Therefore, the null hypothesis (H_0) is rejected, indicating that at least one predictor variable significantly affects the response variable.

Furthermore, the significance of individual predictors was examined using the Wald test. The hypotheses are defined as:

$H_0 : \beta_j = 0, j = 1,2,3,4$
$H_1 : \beta_j \neq 0, j = 1,2,3,4$

The results of the Wald test are presented in Table 5.

Table 5. Wald Test results for the GPR Model

Parameter	Wald Value (W)	$\chi^2_{(0,05;1)}$	Conclusion
β_1	0.031	3.841	Not Significant
β_2	7.272	3.841	Significant
β_3	4.718	3.841	Significant
β_4	2.750	3.841	Not Significant

As shown in Table 5, the open unemployment rate (X_2) and mean years of schooling (X_3) have Wald statistics greater than the critical value of 3.841. Therefore, the null hypothesis (H_0) is rejected for these variables, indicating that they have a statistically significant effect on the number of criminal cases. In contrast, population density (X_1) and the Gini ratio (X_4) have Wald statistics less than the critical value. Thus, the null hypothesis fails to be rejected for these variables, implying that they do not have a statistically significant effect in the model.

3.5. Spatial Heterogeneity Testing

Spatial heterogeneity testing was conducted to determine whether the relationship between predictor variables and the number of criminal cases is consistent across locations or varies spatially. Spatial heterogeneity was assessed using the Breusch–Pagan (BP) test with the following hypotheses:

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{33}^2 = \sigma^2$

H_1: at least one $\sigma_i^2 \neq \sigma^2,\ i = 1,2,..,33$

The analysis produced a BP statistic value of 25.7462, which exceeds the critical chi-square value of $\chi^2_{(0,05;8)} = 15.507$. Therefore, the null hypothesis (H_0) is rejected, indicating the presence of spatial heterogeneity in the data. This result implies that the relationships between predictor variables and criminal cases vary across districts/cities in North Sumatra. Consequently, the GPR model is considered inadequate, and the Geographically Weighted Generalized Poisson Regression (GWGPR) model is employed to account for spatial variation and to provide location-specific parameter estimates.

3.6. Geographically Weighted Generalized Poisson Regression Modeling

The significance of spatial heterogeneity, as identified by the Breusch–Pagan test, indicates that the relationship between predictor variables and the number of criminal cases varies across locations. Therefore, the GPR model is insufficient because it assumes that the effects of explanatory variables are constant across all locations. To accommodate both spatial heterogeneity and overdispersion in count data, the Geographically Weighted Generalized Poisson Regression (GWGPR) model is employed.

The GWGPR modeling process begins by calculating the Euclidean distance between observation locations using geographical coordinates (longitude and latitude). These distances are subsequently used to construct a spatial weight matrix that reflects the influence of neighboring observations on local parameter estimation. The weighting structure is determined through a kernel function and an associated bandwidth parameter. The optimal bandwidth is selected using the Cross-Validation (CV) method. The results of bandwidth selection for each kernel function are presented in Table 6.

Table 6. Optimum bandwidth selection

Kernel Function	Cross Validation (CV)	Bandwidth	Base Bandwidth
Adaptive Gaussian	388,093,455	0.7272668	24 nearest neighbors
Fixed Gaussian	366,850,232	63.57038	63.57 km
Adaptive Bisquare	355,129,647	0.7272577	24 nearest neighbors
Fixed Bisquare	354,705,234	224.5921	224.59 km

Based on Table 6, the adaptive kernel functions (Adaptive Gaussian and Adaptive Bisquare) produce optimal bandwidths corresponding to approximately 24 nearest neighbors. This represents about 72.7% of the 33 districts/cities included in the analysis, indicating that each local model incorporates information from a substantial proportion of the available observations. Consequently, local parameter estimation is based on a relatively large number of neighboring observations. In contrast, the fixed kernel functions (Fixed Gaussian and Fixed Bisquare) use constant bandwidths of 63.57 km and 224.59 km, respectively, indicating that the weighting scheme is determined based on a fixed geographical distance rather than a fixed number of neighboring observations. Among the four candidate kernel functions, the Fixed Bisquare kernel produces the smallest Cross Validation (CV) value of 354,705,234. Since the CV criterion aims to minimize prediction error, the kernel function with the lowest CV value is considered to provide the most appropriate spatial weighting structure for the data. Therefore, the Fixed Bisquare kernel was selected for constructing the spatial weight matrix and estimating the GWGPR model, with an optimal bandwidth of 224.59 km.

Using the Maximum Likelihood Estimation (MLE) method, local parameters are estimated for each district/city. As an example, the GWGPR model for Langkat Regency is expressed as follows:

$$\hat{\mu}_{\text{Langkat}} = \exp(12.9372 + 0.0004X_1 + 0.2387X_2 - 0.4173X_3 - 11.9047X_4) \qquad (26)$$

To validate the model, significance testing is conducted. The simultaneous test using the Maximum Likelihood Ratio Test (MLRT) is defined with the following hypotheses:

$H_0 : \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \beta_3(u_i, v_i) = \beta_4(u_i, v_i) = 0$
$H_1 :$ at least one $\beta_j(u_i, v_i) \neq 0;\ j = 1,2,3,4$

The test produced a deviance value of 203.5904, which exceeds the critical chi-square value of $\chi^2_{(0,05;4)} = 9.4877$. Therefore, the null hypothesis (H_0) is rejected, indicating that at least one predictor variable significantly affects the response variable. Furthermore, the significance of individual predictors is evaluated using the Wald test. The hypotheses are defined as:

$H_0 : \beta_j(u_i, v_i) = 0,\ j = 1,2,3,4; \quad i = 1,2,\dots,33$
$H_1 : \beta_j(u_i, v_i) \neq 0,\ j = 1,2,3,4; \quad i = 1,2,\dots,33$

The Wald test results for Langkat Regency are presented in Table 7.

Table 7. Wald Test results for the GWGPR in Langkat Regency model

Parameter	*W*ald Value (*W*)	$\chi^2_{(0,05;1)}$	Conclusion
β_1	19.7110	3.841	Significant
β_2	6.5652	3.841	Significant
β_3	4.5411	3.841	Significant
β_4	2.8829	3.841	Not Significant

As shown in Table 7, the Wald statistics for population density (X_1), open unemployment rate (X_2), and mean years of schooling (X_3) are greater than the critical value of 3.841. Therefore, the null hypothesis (H_0) is rejected for these variables, indicating that they have a statistically significant effect on the number of criminal cases. In contrast, the Gini ratio (X_4) has a Wald statistic less than the critical value. Thus, the null hypothesis fails to be rejected, implying that this variable does not have a statistically significant effect in the model. Therefore, the local GWGPR model for Langkat Regency can be expressed as follows:

$$\hat{\mu}_{\text{Langkat}} = \exp(12.9372 + 0.0004X_1 + 0.2387X_2 - 0.4173X_3) \tag{27}$$

Based on the estimated model (27), population density (X_1) has a positive effect on the number of criminal cases, with a coefficient of 0.0004, corresponding to $\exp(0.0004) = 1.0004$. This indicates that a one-unit increase in population density (persons/km²), holding other variables constant, increases the expected number of criminal cases by approximately 1.0004 times. Similarly, the open unemployment rate (X_2) shows a positive effect, with a coefficient of 0.2387 ($\exp(0.2387) = 1.2696$), implying that a 1% increase in unemployment increases the expected number of criminal cases by approximately 1.2696 times. In contrast, mean years of schooling (X_3) has a negative effect, with a coefficient of −0.4173 ($\exp(-0.4173) = 0.6588$), indicating that an increase of one year in schooling reduces the expected number of criminal cases to approximately 0.6588 times. These findings are consistent with previous studies, which suggest that higher population density and unemployment increase crime due to intensified social interaction and economic pressure, while higher education reduces crime through improved legal awareness and broader economic opportunities [4], [6], [8]. A summary of significant variables across districts and cities in North Sumatra is presented in Table 8.

As shown in Table 8, the open unemployment rate (X_2) is the most consistently significant variable across regions, indicating that economic pressure plays an important role in driving criminal activity. This finding is consistent with previous research [6], which reported that increases in unemployment tend to be accompanied by increases in criminal activity due to the economic pressures faced by individuals. In contrast, the Gini ratio (X_4) is not significant in any district/city. One possible explanation is that the variation in income inequality across districts/cities is relatively low, as indicated by the low standard deviation reported in Table 2. Consequently, the observed differences in income inequality may not be sufficiently large to explain spatial variations in crime rates.

Table 8. Summary of significant predictors in 33 districts/cities

Significant Variable	Districts/Cities
X_1, X_2	Medan
X_1, X_2, X_3	Langkat
X_2	Asahan, Batu Bara, Binjai, Dairi, Deli Serdang, Humbang Hasundutan, Karo, Labuhan Batu Utara, Nias Selatan, Pakpak Bharat, Pematangsiantar, Samosir, Serdang Bedagai, Simalungun, Tanjung Balai, Tapanuli Utara, Tebing Tinggi, Toba
X_3	Nias Barat, Nias Utara
No Significant Variables	Nias, Mandailing Natal, Tapanuli Selatan, Tapanuli Tengah, Labuhanbatu, Padang Lawas Utara, Padang Lawas, Labuhan Batu Selatan, Sibolga, Padangsidempuan, dan Gunungsitoli

However, this finding does not necessarily imply that income inequality has no relationship with crime in North Sumatra. The insignificant effect may also reflect the presence of other socio-economic factors, such as population density and unemployment, which exhibit stronger associations with criminal activity in the study area. In addition, the Gini ratio is an aggregate indicator that may not fully capture localized socio-economic disparities relevant to criminal behavior.

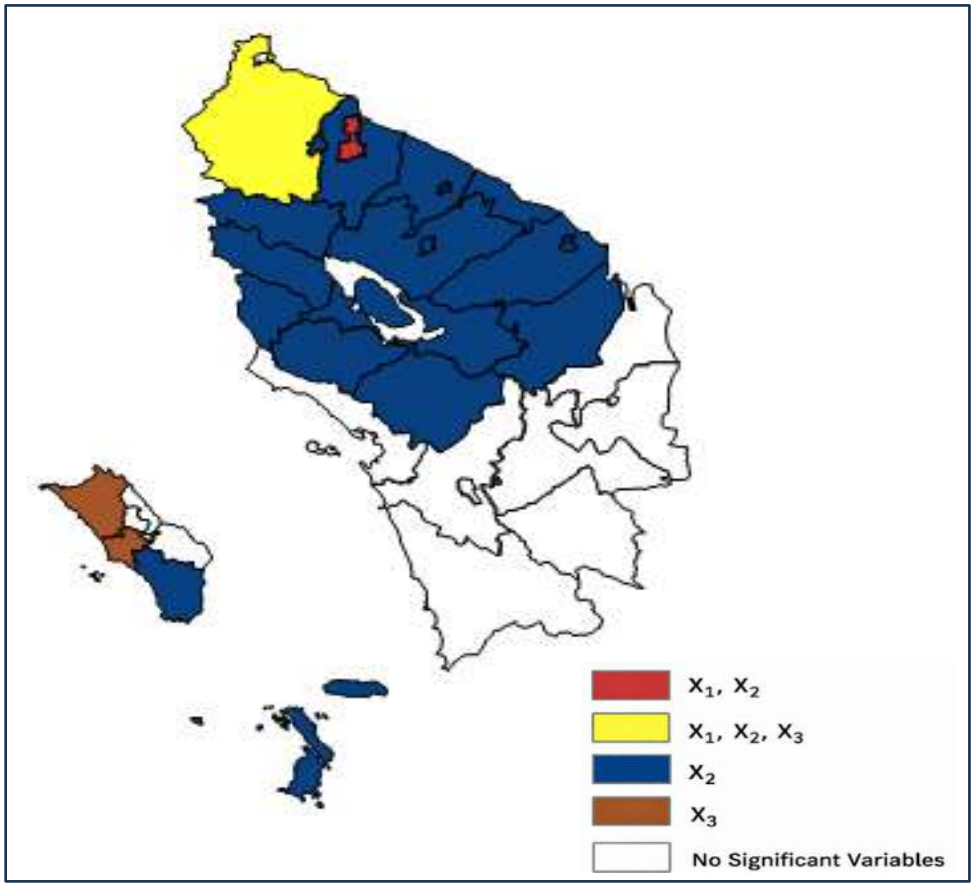


Figure 2. Spatial distribution of significant predictor variables based on the GWGPR model in North Sumatra

This result is consistent with previous findings reported in [25], which indicated that the relationship between income inequality and crime was not always statistically significant. However, it differs from the findings reported in [10], where income inequality was identified as a significant determinant of crime in Indonesia. These differences may be attributable to variations in study area, period of observation, data characteristics, and modeling approaches. Therefore, the relationship between income inequality and crime may vary across regions and depend on local socio-economic conditions.

Additionally, several districts/cities exhibit no significant predictors. This suggests that variations in crime within these areas may be associated with other factors not included in the present model, such as poverty, urbanization, law enforcement effectiveness, demographic structure, or other socio-economic characteristics. However, the absence of significant predictors in some districts/cities does not imply that the GWGPR model is invalid. Instead, it reflects the spatial heterogeneity of crime determinants, indicating that the selected explanatory variables do not have statistically significant local effects in every location. This finding highlights that crime patterns may be influenced by different factors across regions and supports the use of GWGPR, which allows the relationships between crime and explanatory variables to vary across locations. The spatial distribution of significant predictors is illustrated in Figure 2.

4 Conclusion

The results of the GWGPR model indicate that the effects of predictor variables vary across locations, confirming the presence of spatial heterogeneity. Among the predictors, the open unemployment rate (X_2) is the most dominant factor, being statistically significant in 22 districts/cities, followed by mean years of schooling (X_3) in 3 districts/cities and population density (X_1) in 2 districts/cities. These findings demonstrate that the significant factors affecting crime vary across districts and cities in North Sumatra.

The spatial variation identified by the GWGPR model has important policy implications. In districts/cities where unemployment is the only significant factor, policies should focus on employment creation, vocational training, and entrepreneurship programs. In Medan, where both population density and unemployment are significant, crime prevention strategies should combine labor market interventions with urban management and public security measures in densely populated areas. In Langkat, where population density, unemployment, and education are all significant, integrated policies addressing education, employment, and community-based crime prevention are required. Meanwhile, in Nias Barat and Nias Utara, where education is the only significant predictor, efforts to improve educational access and reduce school dropout rates may contribute to crime reduction.

Several districts/cities do not exhibit significant effects for any of the variables included in the model. This suggests that crime in these areas may be influenced by other factors not considered in the present study, such as poverty, urbanization, demographic characteristics, or law enforcement effectiveness. Therefore, for future research, it is recommended to explore additional variables that may influence crime patterns and consider models that integrate both global and local effects, such as Mixed Geographically Weighted Generalized Poisson Regression (MGWGPR).

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Eva Kosasih: Conceptualization, Methodology, Data Collection, Data Analysis, Writing – Original Draft. **Ni Luh Putu Suciptawati:** Research Advisor, Writing – Review. **Luh Putu Ida Harini:** Research Advisor, Writing – Review.

References

[1] S. Bayu and Y. Hudi, "Dampak Kriminalitas Terhadap Kualitas Hidup Masyarakat Urban," Media Hukum Indonesia, vol. 2, no. 6, pp. 308–312, 2025.

[2] Pusiknas Polri, Statistik Kriminal Periode 01 Jan 2024 - 31 Des 2024. Jakarta, Indonesia: Pusiknas Polri, 2025.

[3] C. R. Shaw and H. D. McKay, Juvenile Delinquency and Urban Areas. Chicago, IL, USA: University of Chicago Press, 1942.

[4] J. A. Parenja, M. A.-Z. Salsabila, and Fatmawati, "Dampak Kepadatan Penduduk di Pulau Jawa Terhadap Kesejahteraan dan Kriminalitas," Pediaqu: Jurnal Pendidikan Sosial dan Humaniora, vol. 4, no. 2, pp. 4328–4337, 2025.

[5] G. S. Becker, "Crime and Punishment: An Economic Approach," Journal of Political Economy, vol. 76, no. 2, pp. 169–217, 1968.

[6] K. N. Azmi, S. P. Azzahra, V. K. Dewi, and Y. W. Pertiwi, "Analisis Pengangguran Terhadap Tindakan Kriminalitas di Kota Bekasi," Observasi: Jurnal Publikasi Ilmu Psikologi, vol. 2, no. 3, pp. 223–234, 2024.

[7] T. W. Schultz, "Investment in Human Capital," American Economic Review, vol. 51, no. 1, pp. 1–17, 1961.

[8] D. R. A. Syarifuddin, Pengaruh Rata-Rata Lama Sekolah dan Upah Minimum Terhadap Kriminalitas Melalui Kemiskinan di Indonesia, Bachelor's thesis, Universitas Islam Negeri Alauddin Makassar, Makassar, Indonesia, 2023

[9] R. A. Stouffer, Society and Education. Boston, MA, USA: Allyn & Bacon, 1955.

[10] N. A. Simangunsong, "Pengaruh Ketimpangan Pendapatan Terhadap Kejahatan di Indonesia," Indonesian Journal of Economics and Strategic Management, vol. 2, no. 2, pp. 1612–1620, 2024.

[11] L. Sugiharti, R. Purwono, M. A. Esquivias, and H. Rohmawati, "The Nexus between Crime Rates, Poverty, and Income Inequality: A Case Study of Indonesia," Economies, vol. 11, no. 2, pp. 1–15, 2023.

[12] P. A. Widyastaman and D. Hartono, "Geographic Distribution of Economic Inequality and Crime in Indonesia: Exploratory Spatial Data Analysis and Spatial Econometrics Approach," Applied Spatial Analysis and Policy, vol. 17, no. 2, pp. 547–571, 2024.

[13] L. Sugiharti, M. A. Esquivias, M. S. Shaari, L. Agustin, and H. Rohmawati, "Criminality and Income Inequality in Indonesia," Social Sciences, vol. 11, no. 3, p. 142, 2022.

[14] A. S. Fotheringham, C. Brunsdon, and M. Charlton, Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Chichester, UK: Wiley, 2002.

[15] S. Q. Aina, M. Fauziyah, and W. P. Nurmayanti, "Comparison of Generalized Poisson Regression and Negative Binomial Regression Models Based on Akaike Information Criterion Values," Statistika, vol. 25, no. 1, pp. 86–101, 2025.

[16] M. J. Badriawan and S. Melaniani, "Aplikasi generalized Poisson regression untuk memodelkan faktor yang mempengaruhi jumlah kasus baru difteri di Provinsi Jawa Timur," Media Gizi Kesmas, vol. 12, no. 2, pp. 860–869, 2023.

[17] S. W. Tyas and L. A. Puspitasari, "Geographically Weighted Generalized Poisson Regression Model with the Best Kernel Function in the Case of the Number of Postpartum Maternal Mortality in East Java," MethodsX, vol. 10, p. 102002, 2023.

[18] A. Uswah, J. Rizal, and Y. Fauzi, "Comparison of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression

(GWNBR) Methods in Determining Factors Affecting Tuberculosis Cases in Indonesia," J. Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika, vol. 18, no. 1, pp. 851–865, 2025.
[19] A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data, 2nd ed. Cambridge, UK: Cambridge University Press, 2013.
[20] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 5th ed. Hoboken, NJ, USA: Wiley, 2012.
[21] I. Z. Safira, "Penerapan Metode Generalized Poisson Regression pada Jumlah Kriminalitas di Provinsi Jawa Barat," B.S. thesis, Universitas Islam Negeri Sunan Ampel Surabaya, Surabaya, Indonesia, 2023.
[22] C. C. I. A. Njudang, R. D. Guntur, E. B. Sinu, and J. R. Pahnael, "Perbandingan Model Generalized Poisson Regression dan Negative Binomial Regression dalam Mengatasi Overdispersi pada Kasus Pneumonia di Provinsi NTT Tahun 2023," MATH UNESA: Jurnal Ilmiah Matematika, vol. 13, no. 3, pp. 88–96, 2025.
[23] P. McCullagh and J. A. Nelder, Generalized Linear Models, 2nd ed. London, UK: Chapman and Hall, 1989.
[24] A. Agresti, An Introduction to Categorical Data Analysis, 3rd ed. Hoboken, NJ, USA: Wiley, 2019.
[25] D. T. Anozi and B. Novianda, "Socio-Economic and Property Crime Rate in Indonesia," Economics Development Analysis Journal, vol. 12, no. 3, pp. 305–318, 2023.

DOI: https://doi.org/10.34123/jurnalasks.v18i1.870

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Classification of Village Development Status in Bekasi Regency Using Ensemble Learning and SMOTE-Based Class Balancing

Mochamad Ridwan[1*], Erwin Tanur[2]

[1]BPS-Statistics Bekasi Regency, Bekasi, Indonesia, [2]Training and Education Center, Statistics Indonesia, Jakarta, Indonesia
**Corresponding Author: moch.ridwan@bps.go.id*

ARTICLE INFO

Article history:
Received 25 Nov, 2025
Revised 18 June, 2026
Accepted 26 June, 2026
Published 30 June, 2026

Keywords:
Random Forest, SMOTE, Village Potential Statistics (PODES), Village Development Index (IDM), Classification, Machine Learning, Ensemble Learning, Bekasi Regency.

Abstract

Introduction/Main Objectives: This study aims to classify village development status in Bekasi Regency using machine learning based on the 2024 Village Potential Statistics (PODES) and the Village Development Index (IDM). **Background Problems:** Conventional descriptive assessments ignore complex socio-economic relationships, and class imbalance further reduces model predictive performance. **Novelty:** This study integrates PODES data, ensemble learning, and SMOTE to improve classification, providing a reliable, data-driven framework for village profiling and planning. **Research Methods:** Following preprocessing and a 70:30 split, SMOTE was applied to the training data, and four tree-based models (Decision Tree, Bagging, Random Forest, XGBoost) were evaluated using standard classification metrics. **Finding/Results:** The Random Forest model combined with SMOTE achieved the best classification performance, with an accuracy of 0.7778 and consistently high AUC values across all classes. The most influential predictors were the dominant economic sector, number of farmer groups, availability of basic health services, and presence of micro-business units. These findings demonstrate that combining ensemble learning with SMOTE improves village development classification and provides valuable support for evidence-based rural development planning in Bekasi Regency.

1. Introduction

Villages represent the smallest administrative entities that maintain direct interaction with the community. This position makes the village a strategic foundation for delivering public services and facilitating the fulfillment of basic rights at the local level. As both a community and a government institution, the village plays a vital role in the administrative structure of the Republic of Indonesia [1]. Historically, village communities existed long before the establishment of the modern Indonesian state, and the formation of Indonesia itself emerged from these early rural communities that served as the original basis of governance and social organization [2].

To assess the progress of village development, several measurement tools have been introduced, including the Village Development Index (IDM). IDM categorizes villages into five levels: very underdeveloped, underdeveloped, developing, advanced, and independent [3]. This classification is based on three key dimensions (social resilience, economic resilience, and environmental resilience) [4].

Over time, village development in Indonesia has shown a shift from dependency on central government assistance toward increased capacity for self-governance and local resource management. This transformation makes the study of village independence particularly relevant, especially in regions such as Bekasi Regency. As one of Indonesia's major industrial regions and part of the rapidly expanding Jakarta metropolitan area, Bekasi Regency exhibits substantial variation in village development conditions. Based on the 2024 Village Development Index (IDM), the regency consists of 38 developing villages, 59 advanced villages, and 82 independent villages. This heterogeneity reflects the coexistence of industrial, peri-urban, and agricultural village characteristics, making Bekasi Regency an appropriate empirical setting for examining village development status and evaluating the effectiveness of machine learning approaches for village classification.

The classification of village status serves not only as a measure of development achievement but also as a foundation for evidence-based policymaking. Through this classification, local governments can assess the development level of each village, identify gaps, and set priorities for targeted interventions [5]. For the central government, these results are essential for evaluating the effectiveness of rural development policies, including the allocation of Village Funds. At the same time, local governments benefit from more strategic planning, better budgeting decisions, and the identification of model villages, while other stakeholders, such as NGOs, academic institutions, and private entities, gain reliable information for contributing to rural development initiatives [6]. Therefore, the accuracy of the classification method is critical, as it directly influences the quality of policy recommendations produced.

The primary aim of classifying village development status is to provide a comprehensive overview of village conditions based on social, economic, institutional, and environmental indicators. Beyond current mapping, the classification can also be used to project development trends and support monitoring and evaluation of rural development programs implemented by the government. In addition, classification results can assist policymakers in identifying development disparities among villages and prioritizing interventions based on local needs and characteristics.

In quantitative analysis, decision tree algorithms are among the most widely used classification methods [7]. These models divide the dataset into nodes based on the most informative attributes, producing classification rules that are easy to interpret [8]. Common algorithms include ID3, C4.5, and CART (Classification and Regression Tree).

Tree-based models are highly valued because they can handle both numerical and categorical variables, produce intuitive interpretations, and allow easy visualization of the decision-making process [9]. This makes them well suited for analyzing social, economic, and health data, including studies related to village development, as the resulting rules can be easily understood by policymakers at both district and village levels.

However, decision trees also present limitations. They are prone to overfitting, especially when the number of features is large or the data structure is complex. Additionally, decision trees are highly sensitive to variations in the training data, meaning that small changes can result in significantly different tree structures [10]. These limitations reduce the model’s stability and predictive performance when applied to new data.

To address these challenges, ensemble learning methods have been developed. Ensemble learning combines multiple models to produce predictions that are more accurate, stable, and robust than those of a single classifier [11]. The most widely used examples include Random Forest, which combines many decision trees using bagging, and Gradient Boosting, which improves model performance through iterative learning.

In this research, ensemble learning is expected to enhance the accuracy of classifying village development status in Bekasi Regency. With more reliable classification results, the findings can serve as a stronger foundation for designing targeted and data-driven rural development policies. Furthermore, ensemble methods are expected to improve model stability and predictive performance by reducing the limitations commonly associated with single decision-tree classifiers.

The core problem addressed in this study is how to produce an accurate and reliable classification of village development status using multidimensional data that represent social, economic, institutional, and environmental aspects. Although instruments such as IDM are widely available, traditional analytical methods often struggle to capture the complexity of relationships among variables. Moreover, commonly used tree-based models face risks of overfitting and instability, thereby reducing accuracy when applied to new data. This calls for more adaptive and robust approaches, such as ensemble learning, which is expected to provide better classification performance.

Based on this problem, the study aims to identify factors that influence village development levels in Bekasi Regency, develop classification models using tree-based approaches, and compare their performance with ensemble learning methods. The analysis is limited to villages in Bekasi Regency using 2024 PODES data, with predictor variables representing key social, economic, and institutional indicators. The study evaluates model performance based on common classification metrics such as accuracy, precision, recall, and F1-score.

The novelty of this study lies in the application of ensemble learning for classifying village development levels, an approach that remains relatively limited in studies related to village development. While previous studies have primarily focused on descriptive analyses of village development indicators [12], [13], machine learning research has often relied on single-model classifiers such as Decision Tree algorithms [7], [8], [10]. By combining multiple tree-based models within an ensemble learning framework, this research aims to produce more accurate, stable, and reliable predictive models.

2. Material and Methods

2.1. Village Development Index (IDM)

According to Law No. 6 of 2014, a village is a legal community unit with defined territorial boundaries and the authority to manage governmental affairs and local interests based on ancestral rights and customary practices. Both administrative villages and traditional villages possess the autonomy to regulate and administer their development using local resources and values that exist within the community [13].

To provide a more objective measure of village development performance, the Ministry of Villages, Development of Disadvantaged Regions, and Transmigration introduced the Village Development Index (IDM) in 2015. IDM evaluates villages across three core dimensions (social resilience, economic resilience, and ecological or environmental resilience) which together reflect the village's capacity to manage local potential while responding to development challenges [4].

The indicators that form IDM are constructed based on the principles of sustainable development, in which social, economic, and environmental aspects operate in an integrated manner to ensure long-term continuity. Through this framework, village development is expected not only to support equality and social justice but also to strengthen local values, cultural identity, and environmental stewardship through responsible management of natural resources [14].

Furthermore, IDM also captures progress in village independence following the implementation of the Village Law, supported by Village Funds and the presence of village facilitators. By incorporating local characteristics, village typologies, and community social capital, IDM enables government interventions to become more targeted and impactful. Thus, IDM functions not only as a statistical instrument but also as a strategic planning tool for improving village development effectiveness through collaboration between government institutions and local communities [14].

2.2. Decision tree

Decision trees are a learning method that represent a mapping function from input x to output y in the form of a tree structure. The model operates using a recursive partitioning process, which divides the dataset into smaller subsets based on attribute tests (splits) that provide the highest information gain. This process continues until terminal nodes (leaves) are formed, each representing a final decision or prediction. Decision trees are widely used for both classification (categorical outcomes) and regression (numerical outcomes), with their main advantage being high interpretability, as predictions are generated through a sequence of simple and transparent rules [15].

During tree construction, the selection of the best attribute for splitting is typically based on the information gain metric, which measures the reduction in impurity after the split is performed. The most common impurity measure is entropy, calculated as:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \, log_2(p_i) \tag{1}$$

where S is the dataset, c is the number of classes, and p_i represents the proportion of instances belonging to class i. Once the entropy values are determined, the information gain for each candidate attribute can be computed as follows:

$$Gain(S,A) = Entropy(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{[S]} Entropy(S_v) \quad (2)$$

where A is the candidate attribute, $Values\ (A)$ is the set of possible values for attribute A, and Sv is the subset of data in S for which attribute A takes the value v. The attribute with the highest information gain is selected as the splitting node in the decision tree, as it is considered the most effective in separating the data based on class distinction.

2.3. Ensemble learning

Ensemble learning is a machine learning approach that combines multiple models (commonly referred to as base learners or weak learners) to produce predictive performance that surpasses that of any single model on its own. The core idea is that by integrating models with different strengths and weaknesses, the system can form a composite model that is more accurate, robust, and stable overall.

2.3.1. Bagging (Bootstrap Aggregating)

Bagging predictors, or Bootstrap Aggregating, is an ensemble technique introduced by [16] to improve the predictive accuracy of machine learning models. The method generates multiple replicated training datasets using bootstrap sampling, in which data points are drawn randomly with replacement. A separate model is then trained on each replicated dataset, and the final prediction is obtained through aggregation, typically by averaging for regression tasks or majority voting for classification. This approach has proven effective in reducing variance, especially for algorithms that tend to be sensitive to small changes in the input data [16].

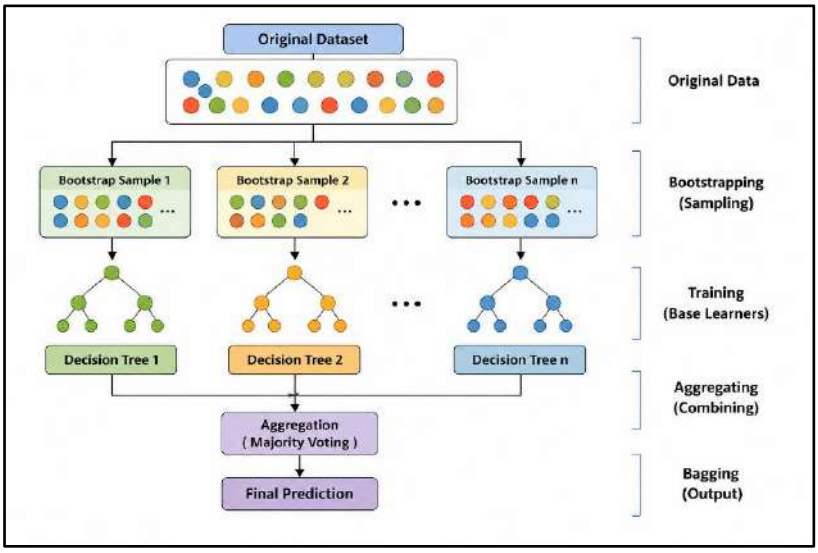


Figure 1. Illustration of the Bagging Process

According to [16], the strength of Bagging lies in its ability to exploit the natural variability in the dataset to enhance model generalization. By training multiple independent models on slightly different versions of the training data and combining their predictions, Bagging can transform an unstable (or weak) predictor into one that is significantly more accurate and robust. This aggregation process minimizes the effect of noise present in any single training sample, making the technique particularly well suited for methods such as decision trees, which are known for their high variance. The illustration of bagging can be seen in Figure 1.

2.3.2. Random Forest

Random Forest is an ensemble-based machine learning algorithm that combines predictions from a large number of decision trees to produce more accurate and stable outputs. Unlike a single decision tree, this algorithm employs the bootstrap aggregating (bagging) technique, in which each tree is trained

on a randomly selected subset of samples and features. This mechanism is specifically designed to reduce correlation among the individual models, thereby effectively minimizing the risk of overfitting and improving predictive validity when applied to new data [17].

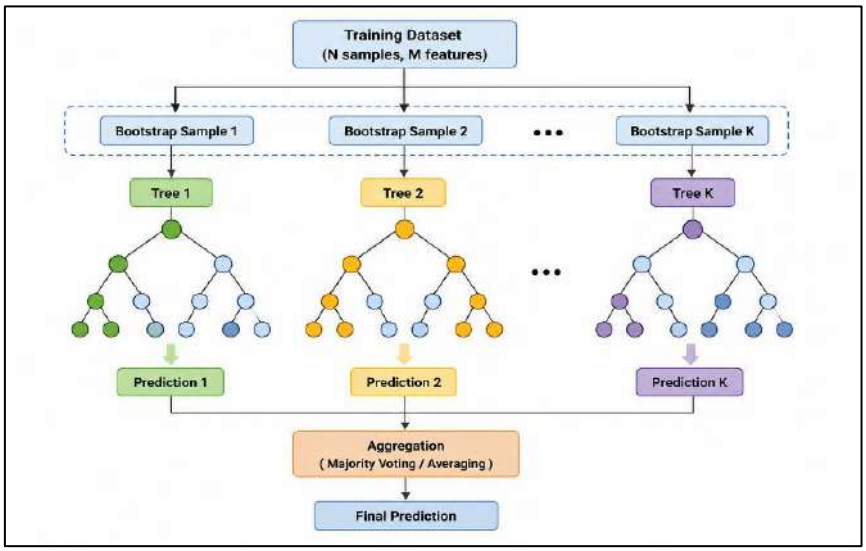


Figure 2. Illustration of the Random Forest Algorithm

In the field of data mining, Random Forest is widely regarded as a powerful predictive method due to its ability to handle both classification and regression tasks using a collective or group-based learning approach. The final prediction is generated by aggregating the outcomes of all decision trees, through majority voting for classification problems or by averaging predictions in regression settings. The illustration of Random Forest can be seen in Figure 2.

2.3.3. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is one of the most advanced ensemble machine learning algorithms, developed as an enhanced implementation of the Gradient Boosting (GB) framework. Functionally, XGBoost combines a series of simple predictive models—typically decision trees—to form a strong learner capable of delivering highly effective performance in both regression and classification tasks [18].

The major advancement of XGBoost compared to conventional Gradient Boosting lies in the introduction of features designed to improve model performance and stability. Among the most notable improvements are the inclusion of regularization and tree pruning mechanisms, which help mitigate the risk of overfitting. In addition, XGBoost is well known for its ability to efficiently process large datasets at high speed, supported by optimizations such as block-based structure and multithreaded CPU execution for parallel computation. The illustration of XGBoost can be seen in Figure 3.

2.4. SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a data resampling algorithm designed to address bias in imbalanced datasets. SMOTE increases the number of samples in the minority class by generating synthetic observations through linear interpolation between an existing minority sample and its k-nearest neighbors. This process helps clarify the decision boundaries between classes and reduces the dominance of the majority class in model training [19]. Technically, SMOTE computes the difference between the feature vector of a minority instance and its nearest neighbor, multiplies this difference by a random value between 0 and 1, and then adds the result to the original feature vector to create a new synthetic sample. By doing so, SMOTE forces the decision region of the minority class to become more generalized, unlike simple duplication methods that tend to create overly specific decision boundaries and increase the risk of overfitting.

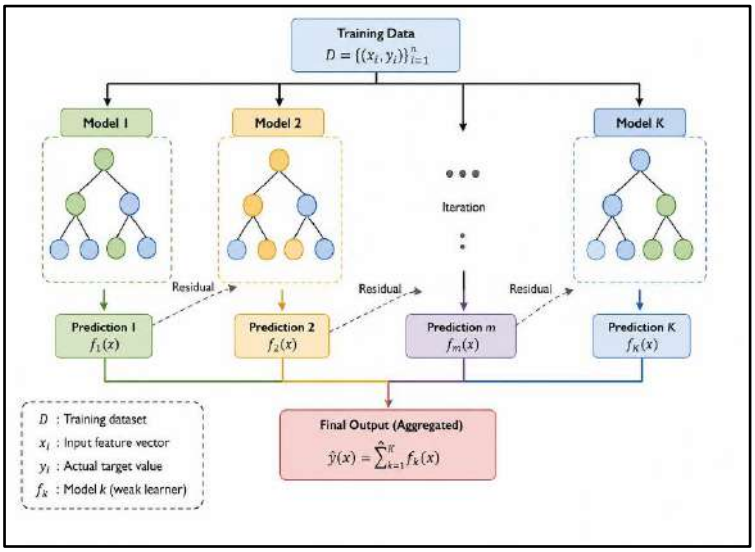


Figure 3. Illustration of the XGBoost Algorithm

2.5. Performance Evaluation

To validate the effectiveness of the algorithms tested, a comprehensive performance evaluation was conducted. The initial assessment is based on the confusion matrix, which maps prediction results into four fundamental components (True Negatives (TN), False Positives (FP), False Negatives (FN), and True Positives (TP)). These components are commonly used to compute standard predictive accuracy. However, relying solely on accuracy as a single metric can be misleading, particularly when dealing with imbalanced datasets or when the cost of misclassification is not uniform. In such situations, a model may appear to perform well simply by prioritizing the majority class, resulting in high accuracy that does not genuinely reflect its ability to detect minority class instances.

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 1. Confussion Matrix

Therefore, following the recommendations of Wu T. et al. (2022), this study adopts the Receiver Operating Characteristic (ROC) curve as a more robust primary performance metric. The ROC curve illustrates the trade-off between the true positive rate (%TP) and the false positive rate (%FP), allowing for an evaluation that is independent of decision thresholds and prior probabilities [20]. According to Fitriani R. et al. (2021), the accuracy value is calculated as follows [21] :

$$\text{Acccuration} = \frac{TP + TN}{TP + FN + TN + FP} \times 100\% \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100\% \quad (5)$$

$$\text{F1} - \text{score} = \frac{2TP}{2TP+FP+FN} \times 100\%$$

As a quantitative indicator, model performance was evaluated using the Area Under the ROC Curve (AUC). Unlike accuracy, AUC provides an unbiased assessment that is not affected by class distribution [22]. An AUC value approaching 1 indicates an ideal classifier with strong ability to distinguish between positive and negative classes, making it the most representative metric for this study. Furthermore, when ROC curves of multiple models intersect, the ROC convex hull analysis can be employed to identify the optimal classifier under specific error-cost constraints.

2.6. Data

This study utilizes the 2024 Village Potential Statistics (PODES) as the primary source of predictor variables. PODES is a nationwide survey conducted periodically by Statistics Indonesia (BPS), covering all villages and urban wards across the country. The dataset provides comprehensive information on demographic, social, economic, infrastructural, institutional, and natural resource conditions within each village. These characteristics make PODES a highly relevant and adequate data source for analyzing variations in development conditions at the village level.

The target variable, on the other hand, is obtained from the 2024 Village Development Index (IDM) published by the Ministry of Village, Development of Disadvantaged Regions, and Transmigration. IDM classifies villages into five levels of development (Independent, Advanced, Developing, Disadvantaged, and Highly Disadvantaged) [23]. However, in the context of Bekasi Regency, none of the villages fall into the Disadvantaged or Highly Disadvantaged categories. Therefore, this study focuses solely on the three relevant categories (Developing, Advanced, and Independent) which serve as the target variable for the classification process. The distribution of IDM categories in Bekasi Regency is illustrated in the figure provided.

Meanwhile, the predictor variables are derived from PODES 2024, as they contain quantitative indicators that objectively describe village conditions and align with the multidimensional assessment framework used in IDM. In addition to comprehensive coverage, PODES is an official dataset produced by the national statistical authority, ensuring high levels of reliability and validity for academic research.

By employing IDM as the dependent variable and PODES as the set of independent variables, this research design is methodologically robust and consistent with the principles of evidence-based policy development. The resulting classification model is expected to support the formulation of more data-driven village development strategies in Bekasi Regency. The complete list of variables used in the analysis is presented in Table 1.

The selection of the 33 predictor variables was guided by theoretical and empirical considerations related to the Village Development Index (IDM). Specifically, the variables were chosen because they represent key aspects of social, economic, and environmental resilience, which constitute the three principal dimensions used in IDM assessment. In addition, only variables that were consistently available in the 2024 PODES dataset, measurable at the village level, and considered relevant to village development conditions were included. Variables that were not directly related to the IDM dimensions or provided redundant information were excluded from the analysis. This theory-driven selection approach was adopted to ensure interpretability and policy relevance while avoiding the inclusion of less informative variables.

Table 1. Research variables

Var	Description	Class	Var	Description	Class
Y	Village Independence Status	1. Developing 2. Advanced 3. Independent	X17	Existence of Featured Products	1. Yes 2. No
X1	Tree Planting on Critical Land, Mangrove Planting, and Similar Activities by the Community	1. Yes 2. No	X18	Existence of Village Cooperatives (KUD)	1. Yes 2. No

Var	Description	Class	Var	Description	Class
X2	Waste/Material Processing or Recycling (Reuse, Recycle) by the Village Community	1. Yes 2. No	X19	Existence of KOPINKRA (Village-level Cooperative Network)	1. Yes 2. No
X3	Promotion of Organic Fertilizer Use in Agricultural Land	1. Yes 2. No	X20	Existence of Savings and Credit Cooperatives (KSP)	1. Yes 2. No
X4	Package A/B/C Education Activities	1. Yes 2. No	X21	Availability of KUR (People's Business Credit) Facilities	1. Yes 2. No
X5	Existence of Community Learning Centers (TBM)	1. Yes 2. No	X22	Availability of KUBE (Group Business Empowerment) Facilities	1. Yes 2. No
X6	Number of Integrated Health Posts (Posyandu)	Numeric	X23	Availability of Banks	1. Yes 2. No
X7	Number of Community Implementers/Leaders	Numeric	X24	Availability of BMT	1. Yes 2. No
X8	Number of Poor Family Certificates (SKTM) Issued by the Village	Numeric	X25	Availability of ATMs	1. Yes 2. No
X9	Number of Persons with Disabilities	Numeric	X26	Availability of Bank Agents	1. Yes 2. No
X10	Number of Family Welfare Movement (PKK) Members	Numeric	X27	Existence of Shop Groups	1. Yes 2. No
X11	Number of Youth Organization (Karang Taruna) Members	Numeric	X28	Existence of Permanent Markets	1. Yes 2. No
X12	Number of Farmer Groups	Numeric	X29	Existence of Semi-Permanent Markets	1. Yes 2. No
X13	Number of Community Groups	Numeric	X30	Number of Minimarkets	1. Yes 2. No
X14	Main Economic Sector of Village Population	Polynomial	X31	Number of Grocery Stores	Numeric
X15	Number of Micro and Small Industries	Numeric	X32	Number of Village-Owned Business Units (Bumdes)	1. Yes 2. No
X16	Number of Industrial Centers	Numeric	X33	Availability of Village Original Revenue (PADes)	1. Yes 2. No

2.7. Workflow

The research process begins with the collection of Potensi Desa (PODES) 2024 data for Bekasi Regency, which contains information on the social, economic, institutional, and environmental characteristics of villages. The collected dataset then undergoes data cleaning and preprocessing procedures, including handling missing values, transforming variables, and encoding categorical attributes to ensure compatibility with machine learning algorithms. In addition, data consistency checks

are performed to improve data quality and ensure the reliability of subsequent analyses.

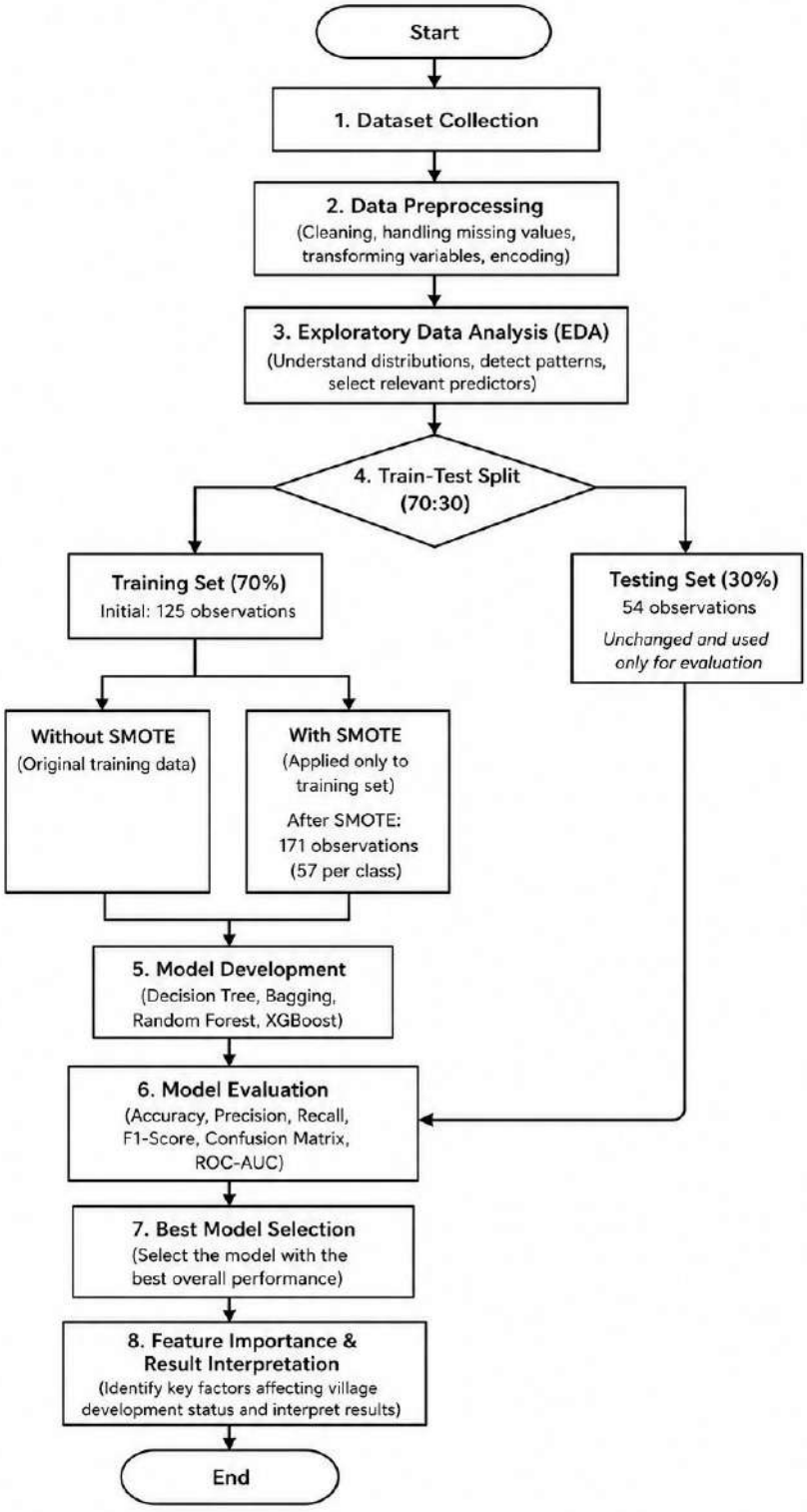


Figure 5. Workflow diagram

The next stage involves Exploratory Data Analysis (EDA) to understand the distribution of variables and identify preliminary patterns within the dataset. Descriptive statistics and visualization techniques are employed to detect anomalies and examine class distribution patterns. The insights obtained from this stage serve as the basis for selecting predictor variables that are conceptually relevant to the Village Development Index (IDM) framework.

The dataset is subsequently divided into training and testing subsets using a 70:30 proportion, where 70% of the observations are allocated for model training and 30% are reserved for model

evaluation. This proportion follows the recommendation of Muraina [25], who emphasizes the importance of an appropriate dataset splitting strategy to reduce modeling bias and improve model generalization capability. Based on this partitioning scheme, the training dataset consisted of 125 observations, while the testing dataset contained 54 observations.

To address the class imbalance problem, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training dataset after the train-test split procedure. This approach was adopted to prevent data leakage and ensure that model evaluation remained unbiased. After applying SMOTE, the number of training observations increased from 125 to 171, resulting in a balanced class distribution of 57 observations for each village development category, while the testing dataset remained unchanged.

Following the data preparation stage, four classification models were developed and compared, namely Decision Tree, Bagging, Random Forest, and XGBoost. The Decision Tree model was used as the baseline classifier, whereas Bagging, Random Forest, and XGBoost represented ensemble learning approaches. All models were trained using the training dataset and subsequently evaluated using the testing dataset. To ensure reproducibility and consistency across experiments, each classification model was implemented using a predefined set of hyperparameters. The hyperparameter settings used in this study are presented in Table 2.

Table 2. Hyperparameter Model

Model	Main Hyperparameters
Decision Tree	max_depth = 10 random_state = 42
Bagging	n_estimators = 200 base_estimator = Decision Tree max_depth = 10 random_state = 42
Random Forest	n_estimators = 200 max_depth = 10 random_state = 42
XGBoost	objective = multi:softprob max_depth = 10 eval_metric = mlogloss random_state = 42
SMOTE	k_neighbors = 5 random_state = 42

The selected hyperparameter values were intended to provide a fair comparison among classification methods while minimizing excessive model complexity. For ensemble-based methods, 200 estimators were employed to improve predictive stability and reduce variance. In addition, SMOTE was implemented using the default value of k_neighbors = 5 to balance the minority classes within the training dataset.

The classification performance of each model was assessed using multiple evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrix, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These complementary metrics provide a comprehensive assessment of classification performance across different village development categories. Comparative evaluation results were subsequently used to identify the model that delivered the most reliable classification performance.

The final stage involves interpreting the classification results and deriving substantive conclusions from the analysis. The best-performing model is further examined to identify the most influential variables affecting village development status and to explain the observed classification patterns. The findings are expected to provide evidence-based insights that can support village development planning and policy formulation in Bekasi Regency. The complete research workflow is illustrated in Figure 5.

3. Results and Discussion

3.1. Overview of Data and Class Distribution

This study utilizes the 2024 Potensi Desa (PODES) dataset for Bekasi Regency, comprising 179 villages. The variables used in the analysis include social, economic, institutional, and environmental indicators, as listed in Table 1. These variables were selected to capture key dimensions of village development and to support the classification of village development status based on the Village Development Index (IDM).

The distribution of village development status in Bekasi Regency shows that the majority of villages fall under the Independent category, with a total of 82 villages. This represents the largest proportion among all categories and indicates that most villages in Bekasi Regency possess relatively strong socio-economic and institutional capabilities. In addition, 59 villages are classified as Advanced, reflecting that a considerable number still have potential for further development toward full independence, particularly through the enhancement of basic services, economic opportunities, and community empowerment. Meanwhile, the developing category includes 38 villages, indicating that a smaller share of areas still require more intensive attention in terms of basic infrastructure, social services, and institutional strengthening in order to catch up with the more advanced villages. IDM Distribution in Bekasi Regency can be seen in Figure 6.

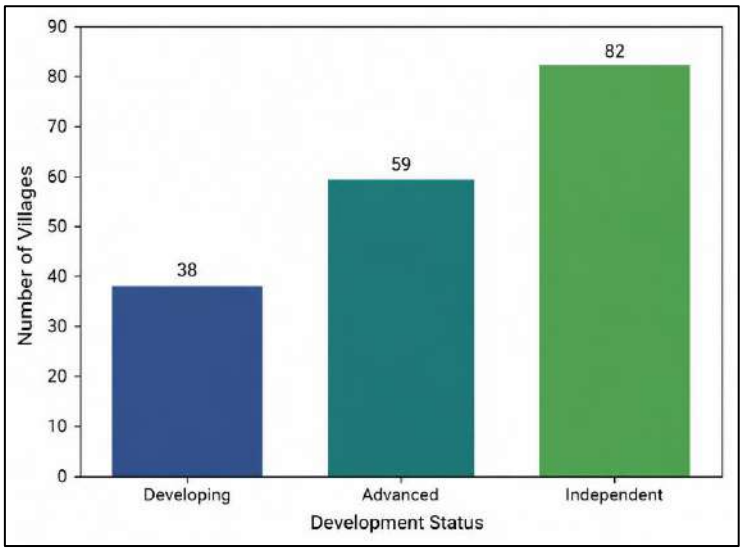


Figure 6. IDM distribution in Bekasi Regency

Overall, this distribution pattern indicates that Bekasi Regency is in a relatively strong position in terms of village development, with a predominance of independent villages reflecting progress across key development sectors. Nevertheless, the presence of advanced and developing villages remains a strategic policy concern to prevent widening disparities and to ensure that development progresses in an equitable and inclusive manner throughout the region.

During the initial exploratory analysis, class distribution revealed an imbalance among the categories (Developing, Advanced, and Independent). The *Independent* category accounted for the largest proportion of observations, while the *Developing* category represented the smallest share.

The correlation heatmap further shows that most numerical variables in the PODES dataset exhibit low to moderate correlations. This indicates that there are no overly strong relationships among variables, suggesting a low risk of multicollinearity and implying that each variable contributes relatively independent information to the village classification model. The highest correlation was observed between X10 and X11 ($r = 0.64$), which reflects a natural association between indicators that tend to develop simultaneously for example, improvements in basic services alongside supporting institutional structures. Moderate correlations were also found between X6–X7 ($r = 0.54$) and X15–X31 ($r = 0.46$), which may represent linkages between basic public service indicators and economic activity within villages. These patterns are reasonable, as more developed villages typically possess more complete public facilities and exhibit more dynamic economic conditions.

Most other variables demonstrated weak correlations (r = 0.00–0.30), including X12, X13, X14, and X16, showing that these indicators capture different dimensions of village conditions. This diversity strengthens the predictive power of machine learning algorithms when the variables are used collectively. Moreover, since no extremely high correlation (r > 0.80) was observed, all variables were retained in the analysis. Overall, the correlation pattern indicates that the numerical variables complement one another and do not exhibit redundancy. Heatmap variabel Numeric can be seen in Figure 7.

3.2. Initial Model Evaluation

The initial evaluation conducted prior to handling class imbalance shows that the performance of the four classification models was still limited due to the unequal distribution of the village development categories. As presented in Table 3, the Decision Tree model achieved an accuracy of 0.6667, reflecting the basic ability of a tree-based classifier to capture relationships among the PODES variables, although it remains sensitive to variance and prone to overfitting. The Bagging model demonstrated an improvement in accuracy to 0.7037, indicating that bootstrap aggregation helped reduce model variance and produce more stable predictions when compared with a single decision tree.

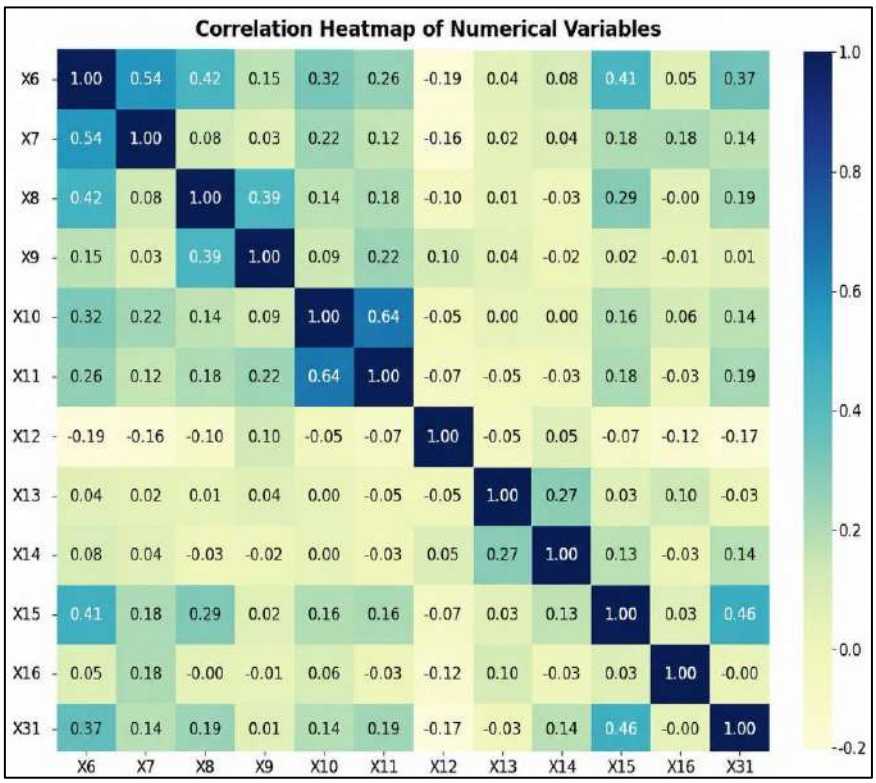


Figure 7. Heatmap of numerical variable

The Random Forest model demonstrated the best performance in the initial stage, achieving an accuracy of 0.7407. This result is consistent with the characteristics of the Random Forest algorithm, which combines hundreds of decision trees to provide more robust predictions, particularly for data with complex and heterogeneous structures. This indicates that the combination of feature randomness and bootstrap sampling is quite effective in enhancing the model's ability to capture inter-variable patterns, even when the data remain imbalanced.

In contrast, the XGBoost model yielded the lowest accuracy, 0.6296, in the initial evaluation. This performance may be attributed to XGBoost's sensitivity to learning rate, max depth parameters, and class distribution imbalance. Since XGBoost optimizes errors iteratively through boosting, class imbalance can cause the model to focus more on the majority class while neglecting minority classes, thereby reducing overall accuracy.

Tabel 3. Initial model accuracy

No	Model	Model Accuracy
1	Decision Tree	0.6667
2	Bagging	0.7037
3	Random Forest	0.7407
4	XGBoost	0.6296

Overall, the evaluation prior to applying SMOTE indicates that class imbalance has a significant impact on model performance, as reflected by the relatively moderate accuracy achieved across all models. These results underscore the importance of implementing imbalance-handling techniques to improve classification accuracy and reduce bias toward the majority class. This finding is consistent with previous studies showing that ensemble methods such as Random Forest generally provide more robust and stable predictive performance than single Decision Tree models [17], [24]. However, even ensemble models may experience performance degradation when trained on imbalanced datasets, making class-balancing techniques such as SMOTE essential for achieving optimal classification performance [19], [20], [25].

3.3. Improving Model Performance Using SMOTE

The application of class balancing techniques using SMOTE improved the performance of most classification models, although the single Decision Tree model experienced a decline in accuracy after oversampling. As shown in Table 4, Bagging, Random Forest, and XGBoost achieved higher accuracy after class balancing, indicating that the initial class imbalance negatively affected predictive performance. These findings suggest that ensemble-based models were better able to benefit from the balanced class distribution generated by SMOTE.

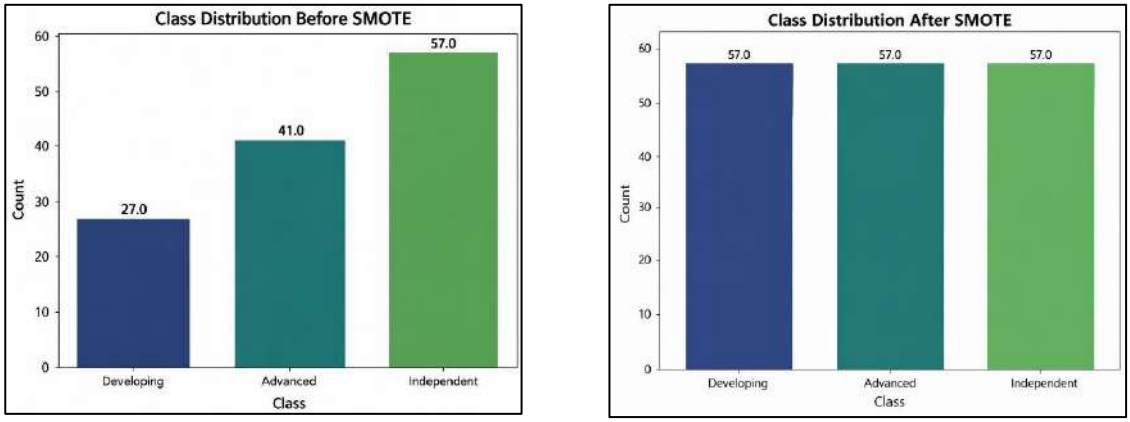


Figure 8. Class distribution of the target variable before and after handling class imbalance

The Decision Tree model experienced a decrease in accuracy from 0.6667 to 0.5741 after applying SMOTE. This decline may be attributed to the sensitivity of single decision trees to changes in data distribution introduced by synthetic samples. Since SMOTE generates new minority-class observations, the resulting data distribution may alter tree-splitting structures and decision boundaries, which can adversely affect the performance of individual tree-based classifiers [19]. In contrast, the Bagging model showed an increase in accuracy from 0.7037 to 0.7407, indicating that ensemble-based approaches are better able to leverage balanced class distributions. This finding is consistent with previous studies suggesting that ensemble methods such as Bagging and Random Forest are generally more robust to data variation and class imbalance because prediction errors are distributed across multiple trees rather than relying on a single classifier [19], [20].

Table 4. Comparison of model accuracy before and after SMOTE

No	Model	Accuracy Before SMOTE	Accuracy After SMOTE
1	Decision Tree	0.6667	0.5741
2	Bagging	0.7037	0.7407
3	**Random Forest**	**0.7407**	**0.7778**
4	XGBoost	0.6296	0.6481

The Random Forest model remained the best-performing model, improving from an initial accuracy of 0.7407 to 0.7778 after applying SMOTE. This indicates that Random Forest is not only robust to data variance but also capable of leveraging a more balanced class distribution to produce more accurate and stable classifications. Improvements were also observed in the XGBoost model, whose accuracy increased from 0.6296 to 0.6481. Although its initial performance was low due to boosting's sensitivity to the majority class, the results after SMOTE demonstrate that XGBoost can function more effectively when class proportion differences are minimized.

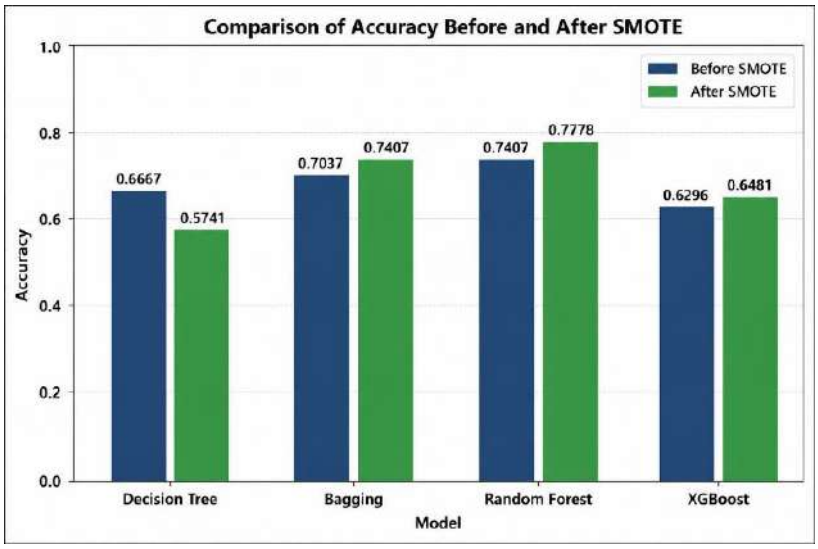


Figure 9. Model accuracy plot

Overall, the application of SMOTE improved the performance of most classification models, particularly Bagging, Random Forest, and XGBoost. Although the Decision Tree model experienced a decline in accuracy, the overall results indicate that class-balancing techniques can substantially enhance predictive performance in imbalanced datasets. These findings highlight that data balancing techniques are a crucial step in classification modeling, particularly when class distributions are uneven [25]. The use of SMOTE significantly contributes to improving classification model accuracy, especially in datasets with imbalanced classes. As reported by Wu et al. (2022), integrating SMOTE into machine learning architectures can achieve higher detection accuracy compared to conventional methods [20]. This improvement occurs because SMOTE enriches minority-class representations, enabling models to learn decision boundaries more effectively [19]. Model Accuracy Plot can be seen in Figure 9.

3.4. Best Model Selection

As presented in Table 4, the classification report indicates that the Random Forest model, applied after SMOTE, achieved good and balanced classification performance across all categories of village independence status. It should be noted that the precision, recall, and F1-score values are identical for each class because the confusion matrix exhibits a nearly symmetric misclassification pattern. Under this condition, the numbers of false positives and false negatives become equal for each class, resulting in identical precision and recall values and consequently identical F1-scores.

Table 5. Random Forest model classification report

Class	Precision	Recall	F1-Score	Support
Developing	0.8182	0.8182	0.8182	11
Advanced	0.6667	0.6667	0.6667	18
Independent	0.8400	0.8400	0.8400	25
accuracy			0.7778	54
macro avg	0.7749	0.7749	0.7749	54
weighted avg	0.7778	0.7778	0.7778	54

In the Developing category, the model achieved precision, recall, and F1-score values of 0.8182, indicating that it can accurately identify villages classified as Developing. In the Advanced category, all three metrics reached 0.6667. Although most villages predicted as Advanced were correctly classified, several Advanced villages were misclassified as either Developing or Independent. This outcome may occur because Advanced villages represent a transitional development stage and therefore share characteristics with both neighboring categories. The Independent category demonstrated the strongest classification performance, with precision, recall, and F1-score values of 0.8400. The high recall indicates that the model successfully identified most villages that truly belong to the Independent category. Meanwhile, the less-than-perfect precision suggests that some villages from other classes, particularly the Advanced category, were incorrectly classified as Independent. This finding reflects the socio-economic similarity between Advanced and Independent villages, making the boundary between these two categories less distinct.

Overall, the model achieved an accuracy of 0.7778, with macro-average and weighted-average F1-scores of 0.7749 and 0.7778, respectively. The closeness of these values suggests that the model provides relatively balanced classification performance across all classes without substantial bias toward any particular category. These findings indicate that the integration of SMOTE-based class balancing and the Random Forest ensemble algorithm is effective in improving predictive performance on datasets with imbalanced class distributions, making it a suitable approach for modeling village independence status classification in Bekasi Regency.

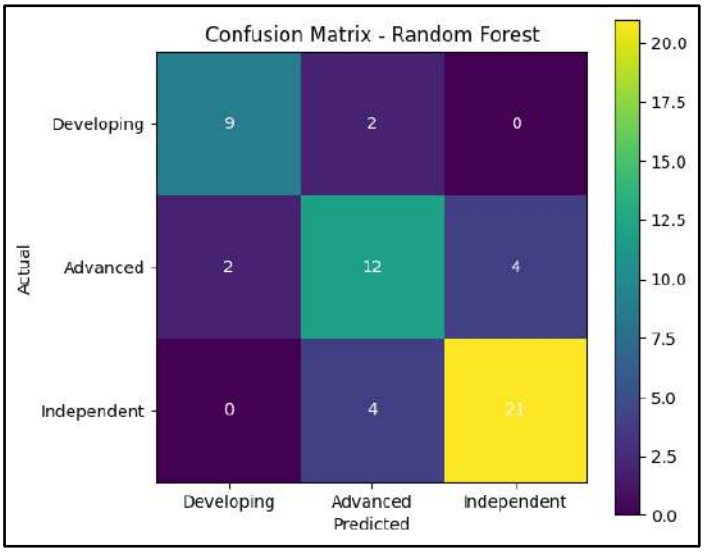


Figure 10. Confusion matrix of the Random Forest Model

The confusion matrix indicates that the Random Forest model (Figure 10), after applying SMOTE, was able to classify the majority of villages into the correct categories. In the Developing category, the model correctly identified 9 out of 11 villages, while 2 villages were misclassified as Advanced, and none were predicted as Independent. For the Advanced category, the model correctly classified 12 out of 18 villages, with 2 misclassified as Developing and 4 misclassified as Independent. This aligns with the relatively lower recall for the Advanced category, reflecting overlapping characteristics with the other two categories.

Meanwhile, in the Independent category, the model correctly identified 21 out of 25 villages, with 4 misclassified as Advanced and none classified as Developing. Overall, these results confirm that most prediction errors occurred between the Advanced and Independent categories, which substantively share

similar socio-economic characteristics, making the boundary between them not always distinct. The multi-class ROC curve illustrates the performance of the Random Forest model in distinguishing among the different village categories (Figure 11). Overall, all three ROC curves lie above the diagonal baseline, indicating that the model has good classification capability across all classes.

In the Developing category, the AUC value is 0.9387, the highest among the three classes. This demonstrates that the model is highly effective in differentiating Developing villages from the other two categories, with a relatively low error rate. The Independent category also shows strong performance, with an AUC of 0.9297. This success is consistent with the confusion matrix results, where most Independent villages were correctly classified. The high AUC indicates that the model can consistently and reliably recognize the characteristic patterns of Independent villages. Meanwhile, the Advanced category has the lowest AUC at 0.8148, though it still falls within the good range. This value confirms that the model's ability to separate Advanced villages from the other two classes is lower compared to the other categories. This aligns with the classification results, which show that some Advanced villages tend to be misclassified as either Developing or Independent, reflecting overlapping characteristics among the categories [26]. Overall, AUC values above 0.80 across all three classes confirm that the Random Forest model exhibits strong and reliable classification performance in predicting village independence levels based on the selected predictor variables.

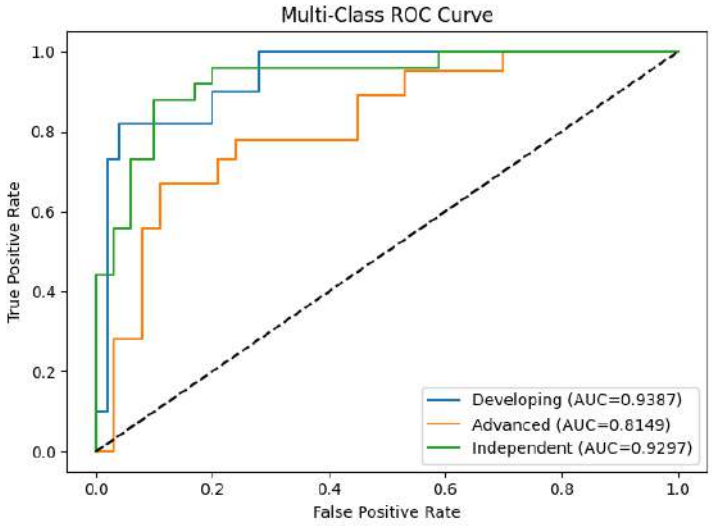


Figure 11. ROC Curve of the best model

3.5. Prediction on Test Data

To provide a clearer picture of model performance on previously unseen data, Table 5 presents a comparison between the actual classification of village independence status and the predictions generated by the selected model for 20 test observations. The table also includes the associated prediction probabilities for each class (Developing, Advanced, and Independent) allowing a more detailed interpretation of the model's level of confidence in each decision.

Overall, the model demonstrates a strong ability to identify the correct class for most observations. Several data points, such as those with indices 15, 117, and 104, were correctly classified as independent, with the highest probability also assigned to that category. This pattern suggests that the model is capable of distinguishing distinct class characteristics when the underlying feature patterns are well represented.

Nevertheless, some misclassifications remain, for example at indices 130 and 14, where the model predicted a different class from the actual label. In many of these instances, the predicted probabilities for the competing classes were relatively close, indicating that the observations may possess overlapping characteristics or fall near the boundary between class definitions. This highlights the inherent complexity of the classification problem and the possibility that some villages share transitional attributes between development categories.

Despite these occasional errors, the predictions on the test dataset reaffirm the model's overall reliability and support the earlier evaluation metrics. The results demonstrate that, after applying

SMOTE to address class imbalance, the model learns meaningful patterns and can generalize well when applied to real or future data.

Table 6. Comparison of actual and predicted classifications on 20 test observations

Index	Actual Class	Predicted Class	Probability Developing	Probability Advanced	Probability Independent
130	Advanced	Developing	0.7362	0.2403	0.0235
15	Independent	Independent	0.0200	0.1850	0.7950
117	Independent	Independent	0.0204	0.3521	0.6275
14	Advanced	Independent	0.0717	0.3038	0.6246
152	Developing	Developing	0.7025	0.2608	0.0367
165	Advanced	Advanced	0.1610	0.6490	0.1900
169	Developing	Developing	0.7035	0.2459	0.0506
104	Independent	Independent	0.0870	0.3072	0.6058
167	Advanced	Advanced	0.1938	0.6112	0.1950
13	Advanced	Advanced	0.1550	0.4888	0.3563
53	Advanced	Advanced	0.3716	0.5482	0.0802
144	Developing	Developing	0.5041	0.4649	0.0310
79	Advanced	Advanced	0.0920	0.7260	0.1820
175	Developing	Developing	0.9000	0.0800	0.0200
115	Independent	Independent	0.0589	0.3935	0.5476
106	Advanced	Advanced	0.0666	0.4690	0.4644
2	Advanced	Independent	0.3150	0.3200	0.3650
36	Developing	Developing	0.6324	0.3674	0.0002
9	Advanced	Independent	0.0354	0.4371	0.5275

3.6. Analysis of Dominant Factors (Feature Importance)

As can be seen in Figure 12, the feature importance analysis indicates that the most influential variable in determining village independence levels is X14 (Main Economic Sector of Village Population). The high importance of this variable suggests that the dominant economic structure of a village community plays a central role in shaping whether a village is Developing, Advanced, or Independent. Villages with productive economic bases, such as trade, industry, or services, tend to have stronger economic capabilities compared to those still reliant on traditional sectors. The next significant contributor is X12 (Number of Farmer Groups), highlighting that the presence of agricultural institutions remains an important foundation for strengthening village economies. A higher number of farmer groups suggests a more structured agricultural production system, better access to training, and stronger networks for community-based development.

Additionally, X6 (Number of Integrated Health Posts/Posyandu) also plays a major role in shaping village independence status. This indicates that the availability of basic health services, particularly for mothers and children, not only influences public health quality but also affects productivity and the economic competitiveness of the village. Meanwhile, the high importance of X31 (Number of Grocery Stores/Warung Kelontong) underscores that local economic circulation through small businesses is a strong indicator of community economic activity. A large number of grocery stores reflects increased purchasing power and dynamic trade activity.

Several other variables, such as X8 (Number of Poor Families/SKTM), X27 (Number of Shop Groups), and X15 (Number of Micro and Small Industries), also contribute meaningfully to the model's decision-making. These variables essentially reflect the dynamics of social welfare and local economic activity. In terms of social institutions, X7 (Number of Community Leaders/Executors) is significant in representing the community's capacity to manage village empowerment programs. Meanwhile, variables such as X9 (Number of Persons with Disabilities) and X30 (Number of Minimarkets) provide

supplementary information regarding social service challenges and the modernization of village commerce.

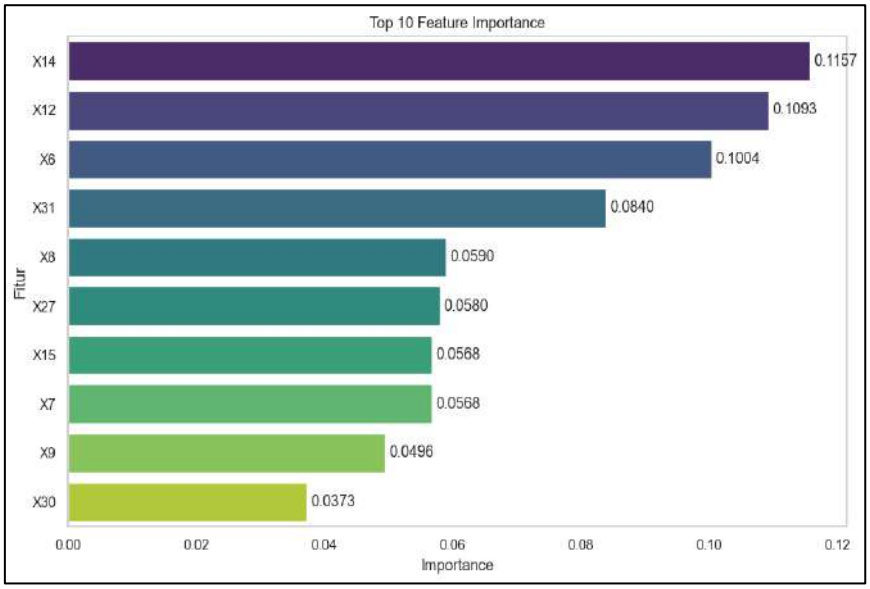


Figure 12. Feature importance best model

Overall, this pattern confirms that village independence is influenced not only by economic factors but also by the presence of social infrastructure, institutions, and robust basic service support. The stronger the local economic structure and community empowerment capacity, the higher the likelihood that a village can sustainably progress toward Independent status.

3.7. Substantive Interpretation of Model Results

The study results indicate that the Random Forest model, after applying SMOTE, was able to deliver strong classification performance in mapping village independence levels in Bekasi Regency. The model achieved an accuracy of 0.7778 with a macro-average F1-score of 0.7749, suggesting that it is not only accurate but also relatively balanced in recognizing all three class categories (Developing, Advanced, and Independent). These findings indicate that the application of data balancing techniques can enhance the algorithm's effectiveness in handling imbalanced class distributions.

Substantively, the model's performance reflects patterns consistent with the realities of village development in Bekasi Regency. In the Independent category, the model achieved its best performance, with precision, recall, and F1-score values of 0.8400. This suggests that Independent villages are easier to distinguish and exhibit more consistent variable patterns compared to the other two categories. Independent villages generally have strong local economic structures, well-developed institutional capacity, and adequate public facilities, making their distinguishing patterns more easily captured by the model.

In contrast, the Advanced category showed the lowest performance, with precision, recall, and F1-score values of 0.6667. This aligns with the confusion matrix results, where some Advanced villages were misclassified as Developing or Independent. This phenomenon reflects that Advanced villages are in a transitional phase and share socio-economic characteristics with both other categories, making the boundaries between classes less distinct. Developing villages exhibited better performance, with an F1-score of 0.8182, indicating that their characteristics can still be consistently recognized by the model.

The findings of this study are consistent with previous research emphasizing the advantages of ensemble learning over single-tree classification approaches. Studies by Joses et al. [11] and Breiman [24] reported that ensemble-based methods generally provide higher predictive stability and better generalization performance because prediction errors are distributed across multiple learners. Furthermore, the observed improvement after applying SMOTE supports the findings of Wu et al. [20] and Pulungan et al. [25], who demonstrated that class-balancing techniques can substantially improve classification effectiveness on imbalanced datasets. The accuracy achieved in this study (0.7778) therefore indicates that the proposed Random Forest–SMOTE framework provides competitive

performance for village development status classification while maintaining balanced predictive capability across all classes.

Further examination of the AUC values in the ROC curve reinforces these findings. The Developing category achieved the highest AUC of 0.9387, followed by Independent at 0.9297, while Advanced had the lowest AUC at 0.8148. Nevertheless, all three AUC values exceed 0.80, indicating that the model possesses strong discriminative ability across all classes. Overall, the model is capable of providing a substantive mapping of village independence status that is consistent with the development patterns observed in Bekasi Regency.

3.8. Strengths & Limitations

This study has several methodological and empirical strengths worth noting. First, the use of an ensemble learning approach, particularly Random Forest, has been shown to provide better classification performance than a single Decision Tree model. This finding is consistent with previous studies demonstrating that ensemble methods reduce model variance, improve predictive stability, and generally achieve better generalization performance by aggregating multiple learners [11], [24]. Second, the application of the SMOTE class-balancing technique effectively improved the model's ability to recognize minority classes, resulting in more equitable classification outcomes that are less biased toward dominant classes. Similar findings have been reported by Chawla et al. [19], Wu et al. [20], and Pulungan et al. [25], who showed that oversampling techniques can substantially improve classification performance in imbalanced datasets. Third, the model leveraged official PODES and IDM data, which are highly valid, comprehensive, and objectively reflect village-level empirical conditions. This strengthens the relevance of the study findings in the context of evidence-based policy and development planning.

However, several limitations should also be acknowledged. First, the model was evaluated using data from a single year (cross-sectional), which limits its ability to capture the longitudinal dynamics of village independence. Time-series analysis could provide deeper insights into village development trajectories over time. Second, the predictor variables were primarily derived from PODES indicators, which are largely quantitative and administrative in nature. Consequently, dimensions such as social capital, community participation, and institutional quality that are not directly captured in numerical form may not be fully represented in the model. Third, the classification results still reveal overlap between the Advanced and Independent categories, primarily due to the similarity of characteristics between villages in these two development stages. This suggests that additional discriminative variables or alternative modeling approaches may be required to improve class separation.

Moreover, this study selected Random Forest as the best-performing model based on empirical evaluation results. Nevertheless, future research could explore other boosting-based algorithms optimized through hyperparameter tuning, which may yield higher predictive performance. The implementation of Explainable Artificial Intelligence (XAI) approaches, such as SHAP or LIME, may also provide deeper insights into the contribution of individual variables to model predictions. Overall, this study provides a robust foundation for applying machine learning to village independence mapping while highlighting opportunities for future methodological improvements, richer data integration, and the incorporation of spatial and temporal dimensions into village development analysis.

4. Conclusion

This study demonstrates that the integration of Village Potential Statistics (PODES) 2024 and the Village Development Index (IDM) through a Random Forest classification model combined with the Synthetic Minority Over-sampling Technique (SMOTE) provides a reliable approach for mapping village independence status in Bekasi Regency. The findings confirm that machine learning methods can support a more objective, consistent, and data-driven assessment of village development by effectively utilizing multidimensional village-level indicators.

he analysis reveals that village independence is closely associated with factors related to local economic structure, agricultural institutions, access to basic health services, and community economic activities. These findings suggest that village development is not determined solely by economic performance, but also by the availability of supporting institutions and public services that strengthen community capacity and resilience. Consequently, policies aimed at accelerating village development should prioritize strengthening local economic sectors, expanding access to essential services, and enhancing community-based institutions.

From a practical perspective, the proposed classification framework can support the Bekasi Regency Government in identifying villages that require targeted interventions, monitoring development progress, and allocating resources more efficiently based on empirical evidence. In particular, villages classified as Advanced may require focused policy attention because they represent a transitional stage with characteristics overlapping those of Developing and Independent villages. Targeted programs designed to strengthen economic opportunities, institutional capacity, and public service provision may help accelerate their transition toward higher levels of village independence.

Furthermore, the study highlights the potential of predictive analytics as a complementary tool for evidence-based rural development planning. Beyond describing current village conditions, machine learning models can assist policymakers in identifying development patterns and anticipating future needs, thereby improving the effectiveness of planning and evaluation processes.

Nevertheless, this study is limited by its reliance on cross-sectional data from a single year and by the availability of predominantly quantitative indicators. Future research is therefore encouraged to incorporate longitudinal datasets, additional social and institutional variables, and advanced analytical approaches, including spatial analysis and explainable artificial intelligence (XAI). Such developments would provide deeper insights into village development dynamics and further strengthen the contribution of data-driven methods to rural policy formulation and decision-making.

Ethics approval

Not required.

Acknowledgments

The authors would like to express their sincere gratitude to Mr. Krido Saptono, Head of the BPS-Statistics Bekasi Regency, for his encouragement and institutional support throughout this research. The authors also thank the BPS- Statistics Bekasi Regency for providing access to the 2024 Village Potential Statistics (PODES) data used in this study. Furthermore, the authors sincerely appreciate the anonymous reviewers and the Editor of the Jurnal Aplikasi Statistika & Komputasi Statistik for their valuable comments and constructive suggestions, which have significantly improved the quality of this manuscript.

Competing interests

The author declares that there are no conflicts of interest related to this study.

Funding

This research received no external funding.

Underlying data

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Credit Authorship

Mochamad Ridwan: Conceptualization, Methodology, Data Curation, Software, Writing (Original Draft, Writing, Review and Editing), Visualization. **Erwin Tanur**: Supervision, Validation, Writing – Review & Editing.

References

[1] B. Rahmasari, "Paradigma Pembangunan Desa Dalam Pengelolaan Keuangan Desa Berdasarkan Undang-Undang Nomor 6 Tahun 2014 Tentang Desa," *Volksgeist: Jurnal Ilmu Hukum dan Konstitusi*, vol. 3, no. 2, 2020, doi: 10.24090/volksgeist.v3i2.4001.

[2] S. Agung, *Pemerintahan asli masyarakat adat: sebuah studi kepemimpinan adat di Lembah Timur Ciamis, Jawa Barat*. Deepublish, 2020.

[3] P. Hendrarso, P. Handoko, M. Faiz Ali Ramdhani, N. Andayani, and R. Tania, "Kajian Pengentasan Desa Tertinggal Melalui Pendekatan Indeks Desa Membangun," *Transparansi : Jurnal Ilmiah Ilmu Administrasi*, vol. 4, no. 1, 2021, doi: 10.31334/transparansi.v4i1.1607.

[4] S. Sriningsih, E. Astuti, B. Ismiwati, and F. Ekonomi, "Implementasi PERMENDESAPDTTRANS NO. 2 Tahun 2016 Terkait Status Desa di Desa Sukarara Lombok Tengah," *Jurnal Kompetitif: Media Informasi Ekonomi Pembangunan, Manajemen dan Akuntansi*, vol. 6, no. 1, 2020.

[5] A. M. Gai, A. Witjaksono, and R. R. Maulida, "Perencanaan dan Pengembangan Desa," 2020, *Dream Litera Buana*.

[6] A. Amka, M. Anshar Nur, and J. Jamalluddin, "Kebijakan dan Strategi Pembangunan Daerah untuk Masa Depan," *CV BRAVO PRESS Indonesia*, (n.d.).

[7] A. H. Nasrullah, "Implementasi algoritma Decision Tree untuk klasifikasi produk laris," *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, vol. 7, no. 2, pp. 45–51, 2021.

[8] H. Kurniawan, "Deteksi Twitter Bot menggunakan Klasifikasi Decision Tree," *Jurnal Sustainable: Jurnal Hasil Penelitian dan Industri Terapan*, vol. 9, no. 1, pp. 31–37, 2020.

[9] I. Setiawan, R. F. A. Cahyani, and I. Sadida, "Exploring complex decision trees: Unveiling data patterns and optimal predictive power," *Journal of Innovation And Future Technology (IFTECH)*, vol. 5, no. 2, pp. 112–123, 2023.

[10] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024.

[11] S. Joses, D. Yulvida, and S. Rochimah, "Pendekatan metode ensemble learning untuk prakiraan cuaca menggunakan soft voting classifier," *Journal of Applied Computer Science and Technology*, vol. 5, no. 1, pp. 72–80, 2024.

[12] T. H. Handoko, *Manajemen personalia dan sumberdaya manusia*. Bpfe, 2016.

[13] D. S. Lindawaty, "Pembangunan desa pasca Undang-Undang No. 6 Tahun 2014 tentang desa [Village development post Law No. 6 of 2014 on villages]," *Jurnal Politica Dinamika Masalah Politik Dalam Negeri Dan Hubungan Internasional*, vol. 14, no. 1, pp. 1–21, 2023.

[14] Direktorat Jenderal Pembangunan Desa dan Perdesaan | KDPDTT, "Tentang Indeks Desa Membangun.", (n.d.).

[15] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: from efficient prediction to responsible AI," *Front. Artif. Intell.*, vol. 6, p. 1124553, 2023.

[16] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[17] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/BJML/2024/007.

[18] M. Niazkar *et al.*, "Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023)," *Environmental Modelling & Software*, vol. 174, p. 105971, Mar. 2024, doi: 10.1016/j.envsoft.2024.105971.

[19] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[20] T. Wu, H. Fan, H. Zhu, C. You, H. Zhou, and X. Huang, "Intrusion detection system combined enhanced random forest with SMOTE algorithm," *EURASIP J. Adv. Signal Process.*, vol. 2022, no. 1, p. 39, 2022.

[21] R. D. Fitriani, H. Yasin, and T. Tarno, "Penanganan klasifikasi kelas data tidak seimbang dengan random oversampling pada naive bayes (Studi kasus: Status peserta KB IUD di Kabupaten Kendal)," *Jurnal Gaussian*, vol. 10, no. 1, pp. 11–20, 2021.

[22] Ş. K. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turk. J. Emerg. Med.*, vol. 23, no. 4, pp. 195–198, 2023.

[23] D. S. Lindawaty, "Pembangunan desa pasca Undang-Undang No. 6 Tahun 2014 tentang desa [Village development post Law No. 6 of 2014 on villages]," *Jurnal Politica Dinamika Masalah Politik Dalam Negeri Dan Hubungan Internasional*, vol. 14, no. 1, pp. 1–21, 2023.
[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[25] M. P. Pulungan, A. Purnomo, and A. Kurniasih, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, 2024.
[26] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Min.*, vol. 16, no. 1, p. 4, 2023.

DOI: https://doi.org/10.34123/jurnalasks. v18i1.895

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Analyzing Medium and Long Text Indonesian Tourism Feedback Using Topic Modeling and Sentiment Analysis

Sulisetyo Puji Widodo[1*], Isnaeni Noviyanti[2]

[1]*BPS-Statistics Indonesia, Jakarta, Indonesia,* [2]*Universitas Indonesia, Depok, Indonesia*
**Corresponding Author: E-mail address:* sulisetyo.widodo@bps.go.id

ARTICLE INFO

Article history:
Received 2 Dec, 2025
Revised 20 Dec, 2025
Accepted 2 June, 2026
Published 30 June, 2026

Keywords:
Feedback, Indonesian Tourism, Natural Language Processing, Sentiment Analysis, Topic Modeling

Abstract

Introduction/Main Objectives: Tourism is a vital sector supporting Indonesia's economic growth, making the effective utilization of public feedback essential for improving service quality. Most feedback is collected through web-based forms in the form of open-text responses that provide rich insights but remain underutilized due to their unstructured nature. Background **Problems**: This study examines the challenge of identifying the most suitable topic modeling and sentiment analysis techniques for analyzing medium- and long-text feedback in the Indonesian tourism context. **Novelty**: The novelty lies in the comparative evaluation of classical topic modeling algorithms against modern embedding-based approaches combined with multiple Indonesian transformer models, which has not been extensively explored in tourism-related datasets. **Research Methods**: The research compares LDA and NMF with BERTopic, Top2Vec, kBERT, and kUSE using coherence scores, and evaluates sentiment analysis using majority voting across transformer architectures. **Finding/Results**: The results show that BERTopic performed best for medium-length text, while NMF was optimal for long text, and a RoBERTa-based model achieved the highest sentiment agreement. Positive sentiment often appeared in feedback on facilities and fees, whereas negative sentiment dominated topics on environmental and governance issues. These findings offer valuable insights for tourism managers and policymakers in prioritizing improvements and refining strategies.

1. Introduction

Tourism is a crucial sector for Indonesia's economic development. Therefore, improving and strengthening this sector requires a data-driven approach. In recent years, web-based digital surveys have successfully collected information on travel patterns and tourist experiences at specific destinations. The information collected is generally structured quantitative data, typically multiple-choice or rating scales.

Furthermore, surveys are often accompanied by feedback forms designed to capture respondents' aspirations, complaints, or input regarding their experiences. Feedback can be about the survey process, application usage, or general views on tourism in Indonesia. The data obtained from these forms is generally unstructured and tends to be narrative, but it contains more in-depth and contextual information. Unfortunately, this type of data is often underutilized in policy formulation and tourism

service development. Yet, open-ended user feedback has strategic potential to uncover real-world issues and capture community expectations that may not be reflected in quantitative survey results.

Processing feedback data can benefit various parties. For survey organizers, user feedback is useful for evaluating and refining survey instrument design, improving questionnaire flow, and identifying technical issues that may have gone undetected during the testing phase. Meanwhile, for governments or authorities managing the tourism sector, feedback can provide a direct picture of public perceptions of existing services, infrastructure, and tourist attractions. This information can be used as a basis for developing more appropriate policies based on community needs. Therefore, utilizing feedback can encourage more targeted and participatory interventions.

However, the narrative and unstructured nature of feedback data makes manual analysis inefficient, especially when the volume of data collected is large. Variations in respondents' communication styles, including language style, length of writing, and focus on specific issues, pose challenges in consistently extracting relevant information. This situation demands an approach capable of classifying and capturing public perceptions without losing context.

In this context, Natural Language Processing (NLP)–based approaches offer an effective solution for analyzing large-scale textual feedback. Techniques such as topic modeling and sentiment analysis are particularly useful for uncovering latent thematic structures and sentiment orientations in user-generated content [1]–[11], [12]–[16], [20]. Topic modeling enables the extraction of hidden thematic patterns from text, including issues related to cleanliness, accessibility, services, and environmental sustainability at tourist destinations [1]–[11]. Meanwhile, sentiment analysis is employed to identify users' attitudes toward these themes, categorizing opinions as positive, negative, or neutral based on contextual cues [12]–[16], [20]. To ensure robust evaluation, model predictions are assessed using majority vote and agreement-based measures, allowing the selection of models that best align with dominant labeling patterns [17]–[19]. The integration of topic modeling and sentiment analysis thus provides a more comprehensive understanding of public perception and supports evidence-based decision-making in the tourism sector.

This study aims to (1) explore the potential use of user feedback data in the context of tourism surveys in Indonesia, focusing on identifying key topics emerging from feedback content and analyzing the accompanying sentiment trends. Furthermore, (2) this study evaluates and selects the best models for both topic modeling and sentiment analysis to recommend the most appropriate model for feedback with medium to long text lengths. Furthermore, (3) by applying topic modeling and sentiment analysis approaches, this study is expected to produce an analytical framework that helps map strategic issues of public concern. Finally, (4) the results of this analysis are not only useful for survey managers in improving the quality of instruments and services, but can also be used as considerations for policymakers in developing more participatory, adaptive, and evidence-based tourism development programs.

2. Material and Methods

The workflow in this study is illustrated in Figure 1, which consists of several main stages. (1) Dataset Preparation, where feedback data collected in CSV format is preprocessed, and only medium-length text (11–30 words) and long text (>30 words) are retained, while short text (<11 words) is removed. (2) Topic Modelling, in which the prepared datasets are processed using various algorithms such as GSDMM, BERTopic, Top2Vec, kBERT, kUSE, NMF, Agglomerative, and LDA. (3) Coherence Score Evaluation, conducted to identify the best-performing topic model for each dataset category. (4) Sentiment Analysis, where the best models are integrated with transformer-based classifiers such as RoBERTa, DistilBERT, BERT, ALBERT, and XLM-RoBERTa. (5) Model Agreement, which ensures consistency and reliability of the sentiment results across models. Finally, the process produces the output in the form of sentiment per topic for both medium and long text datasets.

2.1. Dataset

The data used in this study were obtained from feedback forms distributed concurrently with the Nusantara Tourist Survey (Survei Wisatawan Nusantara) conducted by Badan Pusat Statistik (BPS – Statistics Indonesia) throughout 2024. The Nusantara Tourist Survey focuses on domestic tourism activities, namely Indonesian residents traveling within the country. The feedback forms were administered separately from the main survey questionnaire and were intended to capture respondents'

opinions, experiences, and aspirations regarding domestic tourism and the use of the survey application. A summary of the collected feedback data is presented in Table 1.

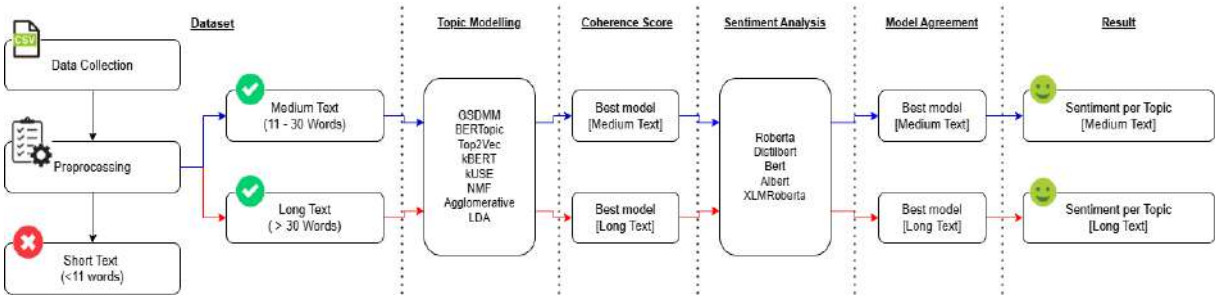


Figure 1. Research workflow

The primary variable in the dataset is "feedback," which contains text entries in the form of open-ended sentences or paragraphs with no length limit or specific structure. All entries were written in Indonesian, reflecting the respondents' population of domestic tourists. Due to its free-form and non-standardized nature, this data displays a variety of language usage, including non-standard words, abbreviations, and informal expressions. Therefore, a preprocessing stage based on Natural Language Processing (NLP) techniques is crucial for filtering, normalizing, and structuring the data before further analysis, such as topic modeling and sentiment analysis.

Table 1. Details of Feedback data

Stage	Feedbacks	Max Words	Avg Words
Data Collection	28.996	1.183	10 - 11
Preprocessing	21.171	851	9 - 10

2.2. Preprocessing

Before further analysis, the text data from the feedback column underwent a pre-processing phase to remove irrelevant elements and adjust the text formatting for uniformity. The first step was to reduce all letters to lowercase to avoid differences in word representation caused solely by capitalization. Next, the text was cleaned of non-linguistic elements such as emojis, symbols, numbers, and links—including internal URLs of the survey system—that were deemed not to contribute to the sentence's core semantic meaning. Redundant characters, such as repeated vowels or consonants, were also removed to normalize variations in informal expressions commonly found in open-ended input data.

Table 2. Details of Feedback data after data separation

Length Category	Feedbacks	Max Words	Avg Words
Short	16.255	10	4
Medium	4.084	30	16
Long	832	851	70

After the structural cleaning phase was completed, the text was separated into word units and filtered using a list of common stopwords that tend not to carry significant semantic weight in the analysis. The remaining words were then stemmed, reducing words to their base or lexical form to unify various morphological variations of the same word. The final results of the pre-processing process were then counted and grouped based on the length of each feedback. Based on a literature review, this study divides feedback into three categories: small (≤10 words), medium (11–30 words), and long (>30 words). This categorization aims to facilitate comparative analyses, such as identifying topic patterns or sentiment tendencies based on feedback length. Details of the dataset grouping are presented in Table 2. Furthermore, Chen et al. [6] highlighted that a large proportion of short texts with generic content may introduce noise and obscure meaningful patterns found in more contextually rich feedback.

Therefore, this study focuses only on the medium and long categories, as these are considered to contain more comprehensive contextual information than short texts.

2.3. Topic Modelling

This stage aims to obtain a thematic structure so that feedback can be grouped based on common themes. In this study, various topic modeling approaches were compared to determine the most effective method for extracting key topics from the text data in the feedback column. Classic statistical models such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) were applied with word frequency-based text representations or TF-IDF, each of which groups documents into topics based on the distribution of dominant words. Meanwhile, more modern, semantic distribution-based approaches, such as Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM), are designed to handle short and varied documents—such as open-ended feedback data—without requiring predefined topic distributions.

This study also employed deep learning-based approaches (neural embedding), including BERTopic, Top2Vec, kBERT, and kUSE, which utilize transformer models such as Sentence-BERT and Universal Sentence Encoder to obtain vector representations of documents. These models enable more contextual topic grouping based on semantic proximity and generate topics with higher meaning cohesion through clustering algorithms such as K-Means and Agglomerative Clustering.

All models are then evaluated using a coherence score to assess the semantic consistency of the resulting topics. In this study, the C_V metric was used to evaluate and identify the best-performing models for both medium-length text (11–30 words) and long-length text (>30 words). C_V was chosen because this metric has proven to be stable, interpretable, and capable of capturing semantic relationships between topics across both text size categories. The use of a uniform metric across both datasets also facilitates objective and consistent performance comparisons between models.

To improve the relevance and accuracy of topic extraction, the entire topic modeling process in this study was applied separately to only two text length categories: medium (11–30 words) and long (>30 words). This separation was made because the linguistic and semantic characteristics of short feedback tend to differ significantly compared to longer text. By dividing the topic modeling process into length categories, each model can be calibrated to suit the structure and density of information within each data set. This strategy is expected to produce more precise and applicable topic mapping and enable the identification of thematic differences or similarities between categories, which will be useful for supporting decision-making in managing the national tourism sector.

2.4. Sentiment Analysis

The sentiment analysis in this study was conducted using a transformer-based approach, a state-of-the-art deep learning model that has proven effective in understanding semantic and syntactic context in natural text. Five pre-trained Indonesian language models from the Huggingface website were utilized. Each selected model had a different architecture, as shown in Table 3, to determine the best model based on its architecture. At this stage, each entry in the feedback data was modeled separately to identify sentiment polarity, such as positive, neutral, or negative for each model. This process was automated using a sentiment classification pipeline, which generated labels and confidence scores for each prediction, enabling the visualization of opinion trends and analysis of public opinion about the tourism sector.

Table 3. Overview of Indonesian Sentiment Analysis Models

Model	Architectures	Language
w11wo/indonesian-roberta-base-sentiment-classifier	**RoBERTa** For Sequence Classification	Indonesian
afbudiman/distilled-optimized-indobert-classification	**DistilBert** For Sequence Classification	Indonesian
ayameRushia/bert-base-indonesian-1.5G-sentiment-analysis-smsa	**Bert** For Sequence Classification	Indonesian
tyqiangz/indobert-lite-large-p2-smsa	**Albert** For Sequence Classification	Indonesian
cardiffnlp/twitter-xlm-roberta-base-sentiment-multilingual	**XLM-RoBERTa** For Sequence Classification	Multilingual

2.5. Model Agreement

This study used a dataset without ground truth labels, so model evaluation for sentiment analysis was conducted using a model agreement approach to assess the level of prediction agreement between models. This study used majority voting, a method that determines the final label based on the highest consensus among models. The level of agreement between a model and the majority label indicates the consistency of predictions on the same data. The model with the highest agreement with the majority label is considered the most stable and representative of the data characteristics and is therefore selected as the primary candidate for further sentiment analysis. This evaluation also considered variations in text length (medium and long) to determine its effect on prediction consistency.

3. Results and Discussion

3.1. Medium dataset

3.1.1 Topic Modelling

At this stage, all feedback that falls into the medium text length category will be assigned topics using each model (GSDMM, BERTopic, Top2Veck, BERT, kUSE, NMF, Agglomerative, and LDA). Each model is tested with several topics and word count settings. Table 4 shows an example of feedback with 5 topics and 10 words assigned topics.

Table 4. Example of Topic Modeling Results Using Various Models (Medium Dataset)

Model	Topic	Feedback
GSDMM	4	Prices for tourists are often excessively high and are deliberately inflated for profit. In some tourist destinations in Indonesia, visitors may also encounter coercive practices and irresponsible behavior from local service providers. (*Harga untuk wisatawan yang sangat mahal dan di sengaja mengambil kesempatan untuk meraih keuntungan dan terkadang tempat wisata di indonesia orang-orang di sana suka memaksa dan tidak bertanggung jawab*)
BERTopic	**0**	
Top2Vec	0	
kBERT	0	
kUSE	0	
NMF	4	
Agglomerative	0	
LDA	3	

3.1.2 Coherence Score

Evaluation results for the medium-length feedback data category show that BERTopic dominates as the model with the best performance across most configurations, Table 5. This model recorded the highest coherence scores for the configurations of 5 topics, 5 words (0.819), 5 topics, 10 words (0.808),

10 topics, 5 words (0.819), and 10 topics, 10 words (0.761). This consistent score demonstrates that the contextual embedding-based approach used by BERTopic is effective in capturing semantic relationships between words, facilitating theme interpretation in medium-sized data.

On the other hand, Non-negative Matrix Factorization (NMF) demonstrated very competitive and even superior performance for the configuration of 15 topics, 5 words (0.816), while maintaining a high score for the configuration of 15 topics, 10 words (0.745). This confirms that NMF has an advantage when the number of topics is expanded, although overall it remains slightly behind BERTopic in terms of consistency across configurations.

Table 5. Coherence Scores of Topic Models (Medium Dataset)

Model	5 Topics, 5 Words	5 Topics, 10 Words	10 Topics, 5 Words	10 Topics, 10 Words	15 Topics, 5 Words	15 Topics, 10 Words
GSDMM	0.597	0.464	0.563	0.470	0.576	0.489
BERTopic	***0.819**	**0.808**	**0.819**	**0.761**	0.762	0.760
Top2Vec	0.453	0.334	0.453	0.344	0.384	0.291
kBERT	0.574	0.405	0.608	0.467	0.604	0.468
kUSE	0.542	0.388	0.574	0.433	0.588	0.458
NMF	0.775	0.697	0.766	0.698	**0.816**	**0.745**
Agglomerative	0.560	0.392	0.497	0.417	0.511	0.402
LDA	0.511	0.414	0.531	0.416	0.574	0.463

Other models, such as GSDMM (maximum 0.576), kBERT (0.608), and kUSE (0.588), produced moderate performance—better than both LDA (0.574) and Agglomerative Clustering (0.560)—but still lagged far behind BERTopic and NMF. Top2Vec, on the other hand, achieved the lowest score (maximum 0.453), making it less reliable in this context.

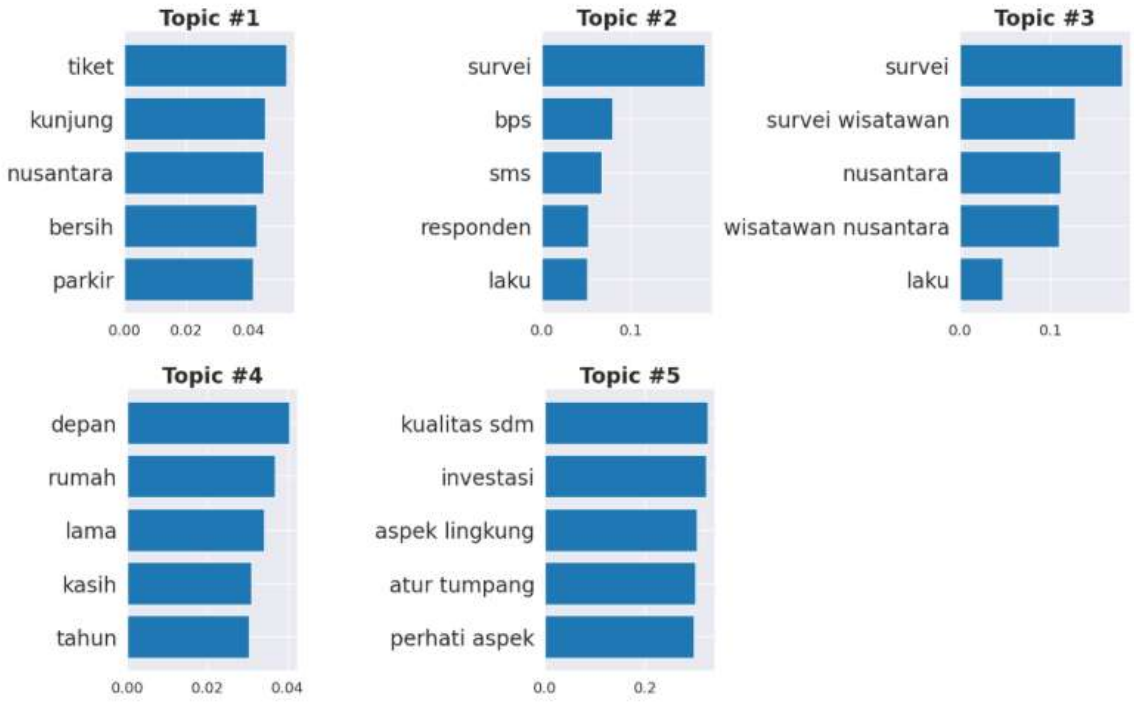


Figure 2. Top Terms per Topic (5 Topics × 5 Words) by BERTopic on Medium Dataset

Overall, these results confirm that BERTopic is recommended as the primary model due to its high and consistent coherence across various configurations. However, NMF can be a strong alternative, especially for scenarios with a larger number of topics, while the other models are more appropriate for comparison or complementarity in exploratory analysis.

Based on the evaluation in Table 5, BERTopic with a configuration of five topics and five keywords successfully identified distinct themes. Figure 2 illustrates that Topic #1 includes terms such as ticket, visit, and parking, which relate to tourist activities and destination facilities. Topic #2 emphasizes survey, BPS, and SMS, reflecting methods of data collection and respondent outreach. Topic #3 highlights survey, tourist survey, and nusantara, indicating the participation of domestic tourists and their travel behavior, and showing thematic connections with Topic #2.

Meanwhile, Topic #4 is less specific, containing terms such as front and house, making its interpretation more abstract and requiring further contextual analysis. Topic #5 underscores governance and development issues, represented by human resource quality, investment, and environmental aspects. While these keywords are useful for distinguishing topics, thematic overlaps remain between some categories, particularly between Topics #2 and #3, requiring manual interpretation. To clarify the thematic context, Table 6 presents topic interpretations and descriptions, formulated with expert input and intended as a reference for further analysis.

Table 6. Topic Interpretations and Top Keywords from BERTopic (Medium Dataset)

Topic	Interpretation	Description	Keywords
1	Tourist Facilities and Comfort	Tourist experiences related to facilities, cleanliness, tickets, and parking at tourist destinations.	Ticket, Visit, Nusantara, Clean, Parking (*Tiket, Kunjung, Nusantara, Bersih, Parkir*)
2	Survey and Data Collection Methods	Survey data collection methods via BPS and SMS to reach respondents.	Survey, BPS, SMS, Respondent, Sold (*Survei, BPS, SMS, Responden, Laku*)
3	Survei Wisatawan Nusantara	Participation of domestic tourists in the survey to understand travel behavior.	Survey, Tourist Survey, Nusantara, Domestic tourist, Sold (*Survei, Survei Wisatawan, Nusantara, Wisatawan Nusantara, Laku*)
4	Respondents' Perceptions and Preferences	Respondents' perceptions and preferences related to daily experiences.	Front, House, Long, Give, Year (*Depan, Rumah, Lama, Kasih, Tahun*)
5	Investment and Environmental Management	Management of human resources, investment, and environmental aspects with potential overlaps.	Human Resource Quality, Investment, Environmental Aspect, Overlap, Attention aspect *(Kualitas SDM, Investasi, Aspek Lingkung, Tumpang Tindih, Perhati Aspek*)

After identifying the themes, the next step was to examine the distribution of feedback on each topic. Figure 3 shows that the distribution of responses was uneven: Topic #1 (Tourist Facilities and Convenience) was the top performer, followed by Topic #2 (Survey and Data Collection Methods). Conversely, topics like Topic #4 (Respondent Perceptions and Preferences) and Topic #5 (Investment and Environmental Management) received relatively few responses. This imbalance indicates that issues related to tourist facilities and the survey process garnered more attention, while personal perceptions and environmental governance received less attention. This could signal that low-intensity topics require further review, especially if they are assumed to be hidden complaints.

3.1.3 Sentiment Analysis

In this stage, feedback belonging to the medium text length category is analyzed using several sentiment analysis models (RoBERTa, DistilBert, Bert, Albert, XLM-RoBERTa). These models are employed to classify the feedback into three categories: positive, negative, and neutral. Table 7 provides an example of sentiment analysis results, showing how each model assigns different sentiment labels to the same feedback text.

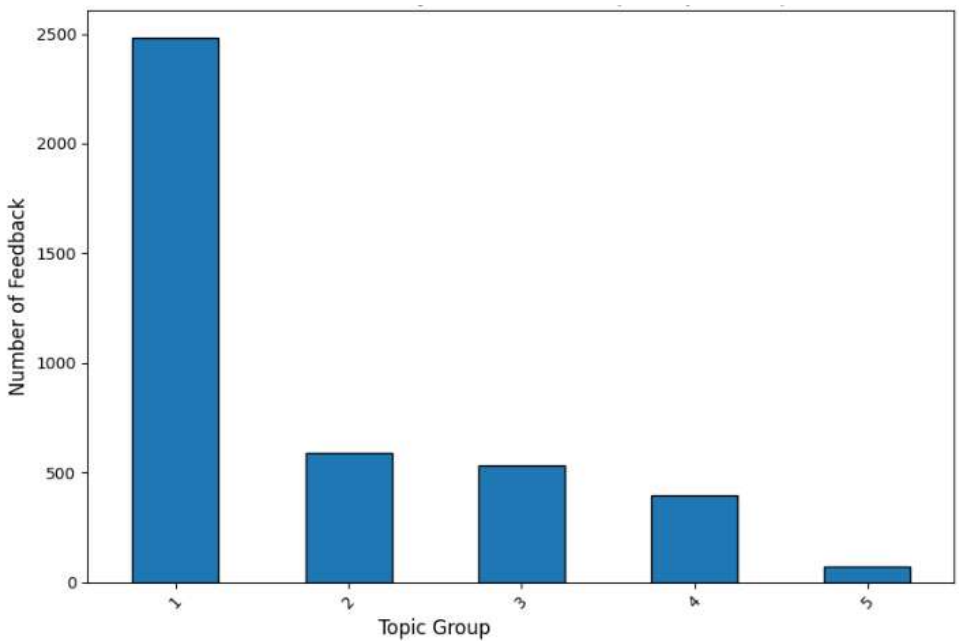


Figure 3. Number of Feedback Entries per Topic by BERTopic (Medium Dataset)

Table 7. Example of Sentiment Analysis Results Using Various Models (Medium Dataset)

Model	Sentiment	Feedback
RoBERTa	(-)	Prices for tourists are very expensive and are deliberately taken advantage of to gain profits and sometimes in tourist attractions in Indonesia the people there like to force and are irresponsible.
DistilBert	(+)	
Bert	(+)	(*Harga untuk wisatawan yang sangat mahal dan di sengaja mengambil kesempatan untuk meraih keuntungan dan terkadang tempat wisata di indonesia orang-orang di sana suka memaksa dan tidak bertanggung jawab*)
Albert	(-)	
XLM-RoBERTa	(-)	

3.1.4 Model Agreement

The evaluation results in Table 8 show that the RoBERTa model achieved the highest agreement with the majority predictions at 82.32%, closely followed by ALBERT at 82.07%. BERT and XLM-RoBERTa also performed competitively, with agreement scores of 79.43% and 78.26%, respectively, while DistilBERT, as a more lightweight model, obtained the lowest agreement at 68.29%.

Table 8. Model Agreement Results for Sentiment Analysis on the Medium Dataset

Model	Agreement with the majority	%
RoBerta	**3.362**	**82.32**
DistilBert	2.789	68.29
Bert	3.244	79.43
Albert	3.352	82.07
XLM-RoBerta	3.196	78.26

These results suggest that architectures with enhanced contextual modeling capabilities, such as RoBERTa and ALBERT, provide more consistent predictions than baseline or distilled models. While BERT and XLM-RoBERTa remain reliable, architectural refinements appear to offer a clear advantage in agreement stability. The lower performance of DistilBERT reflects the trade-off inherent in model

distillation, where reduced complexity improves efficiency but limits the capacity to capture deeper semantic nuances. Overall, model selection for sentiment analysis should balance computational efficiency with the need for contextual depth and prediction stability.

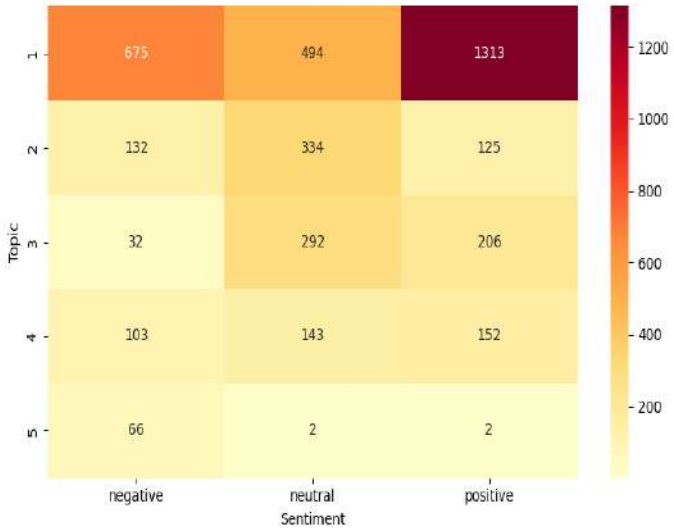


Figure 4. Sentiment Distribution Across Topics by BERTopic (Medium Dataset)

After identifying the most stable sentiment model, the next step was to analyze sentiment distribution across topics using predictions from the RoBERTa model, which achieved the highest agreement score of 82.32%. As shown in Figure 4, the results indicate that:

- Topic 1 (Tourist Facilities and Convenience) was the most frequently discussed topic by respondents and a key focus of attention. This topic was characterized by keywords such as "tickets," "visit," "Indonesian archipelago," "clean," and "parking," and was dominated by positive sentiment, indicating satisfaction with the tourist facilities provided. However, a significant amount of negative sentiment also indicated that there were still a number of complaints regarding these facilities.
- Topic 2 (Survey and Data Collection Methods) and Topic 3 (Domestic Tourist Survey) were dominated by neutral sentiment, with keywords such as "survey," "bps," "sms," "respondents," and "domestic tourist survey." These topics were more informative because many respondents referred to the survey process itself, rather than services or direct experiences.
- Topic 4 (Respondent Perceptions and Preferences) showed a relatively balanced distribution of sentiment. Keywords such as "front," "home," "long," "love," and "year" indicated a variety of opinions regarding respondents' personal perceptions or preferences. This topic reflects a mix of satisfaction and dissatisfaction.
- Topic 5 (Investment and Environmental Management) has the least amount of data, but almost all of the sentiment is negative. Keywords such as human resource quality, investment, environmental aspects, overlapping, and attention to aspects indicate that this topic relates to criticism of environmental management or policy aspects that are not yet optimal. Therefore, this topic needs to be prioritized for future improvement.

Overall, Topic 1 is an area that needs to be maintained and continuously improved because it indicates satisfaction with the majority of respondents, while Topic 5 is the most critical area and requires immediate follow-up improvements.

3.2. Long dataset

3.2.1 Topic Modelling

In this stage, all long text category feedback is classified as assigned to topics using various models, including GSDMM, BERTopic, Top2Vec, kBERT, kUSE, NMF, Agglomerative, and LDA. Each model is evaluated with different configurations of topic number and word count. Table 9 presents an example of feedback from the long data set that has been assigned to five topics with ten representative words.

Table 9. Example of Topic Modeling Results Using Various Models (Long Dataset)

Model	Topic	Feedback
GSDMM	4	This feedback is provided in response to the 2024 Digital Nusantara Tourist Survey conducted by Badan Pusat Statistik (BPS – Statistics Indonesia). The survey is planned to be implemented by BPS in July 2024. The survey is expected to generate comprehensive information related to tourism activities in Indonesia, including detailed tourist profiles, travel motivations, purposes of travel, types of accommodation used during trips, … (*Masukan terhadap Survei Digital Wisatawan Nusantara 2024 oleh Badan Pusat Statistik (BPS), namun berikut adalah beberapa informasi mengenai survei tersebut: Survei Digital Wisatawan Nusantara 2024 akan dilakukan oleh BPS pada Juli 2024. Survei ini akan menghasilkan informasi terkait wisata, seperti profil wisatawan, maksud perjalanan, akomodasi yang digunakan, dan rata-rata lama perjalanan.*)
BERTopic	0	
Top2Vec	0	
kBERT	0	
kUSE	0	
NMF	4	
Agglomerative	0	
LDA	3	

3.2.2 Coherence Score

Referring to Table 10, the topic modeling evaluation results for the long feedback data category show that Non-negative Matrix Factorization (NMF) consistently performed best compared to the other seven models. NMF recorded the highest coherence scores in almost all configurations: 0.680 for 5 topics of 5 words, 0.698 for 10 topics of 5 words, and 0.696 for 15 topics of 5 words. Even in other configurations, NMF remained above 0.58, indicating consistent performance in maintaining semantic relationships between words. These values strengthen evidence that NMF is effective in mapping thematic structures in long texts, which generally have higher vocabulary variety and meaning complexity.

Table 10. Coherence Scores of Topic Models (Medium Dataset)

Model	5 Topics, 5 Words	5 Topics, 10 Words	10 Topics, 5 Words	10 Topics, 10 Words	15 Topics, 5 Words	15 Topics, 10 Words
GSDMM	0.522	0.441	0.519	0.453	0.529	0.481
BERTopic	0.602	0.473	0.576	0.473	0.602	0.473
Top2Vec	0.418	0.333	0.444	0.329	0.444	0.329
kBERT	0.569	0.435	0.561	0.446	0.563	0.450
kUSE	0.551	0.451	0.541	0.437	0.545	0.463
NMF	**0.680**	**0.581**	***0.698**	**0.625**	**0.696**	**0.609**
Agglomerative	0.476	0.385	0.538	0.453	0.534	0.419
LDA	0.563	0.450	0.528	0.391	0.541	0.419

Under NMF, BERTopic and kBERT performed quite competitively. BERTopic achieved the highest score of 0.602 for the 5 topics of 5 words and 15 topics of 5 words. Furthermore, kBERT achieved a maximum score of 0.569 for the 5 topics of 5 words. Both models demonstrated the ability to capture semantic representations, although their consistency was still not on par with NMF. Meanwhile, GSDMM recorded the highest score of 0.529 on 15 5-word topics, but its performance remained below the three main models. Models such as Top2Vec, kUSE, and Agglomerative Clustering showed fluctuating results, with average scores below 0.55, indicating limitations in maintaining coherence across configurations. LDA, while relatively stable, only achieved a maximum score of 0.563 on 5 5-word topics, making it less than ideal for analyzing long text data.

Overall, for the long feedback category, NMF is the most recommended choice because it consistently provides the highest scores across almost all configurations. A configuration of 10 5-word topics can be considered the most optimal representation, while 5 5-word topics and 15 5-word topics remain strong options for maintaining a balance between topic number and semantic cohesion. Models such as BERTopic and kBERT can be considered as secondary alternatives or comparisons in exploratory analysis.

Topic modeling results on the long dataset yielded ten main topics with more targeted keywords that illustrate diverse issues in tourism, as shown in Figure 5. Topic #6 highlights travel costs with the terms "ticket prices," "airplane tickets," and "expensive flights," while Topic #2 focuses on destination quality and accessibility with the terms "tourist attractions," "Indonesian tourism," and "access to places." Topic #3 addresses environmental and governance issues, Topic #4 emphasizes the digitalization of tourism, while Topic #5 addresses geographic aspects with the terms "Java island," "island country," and "sea route."

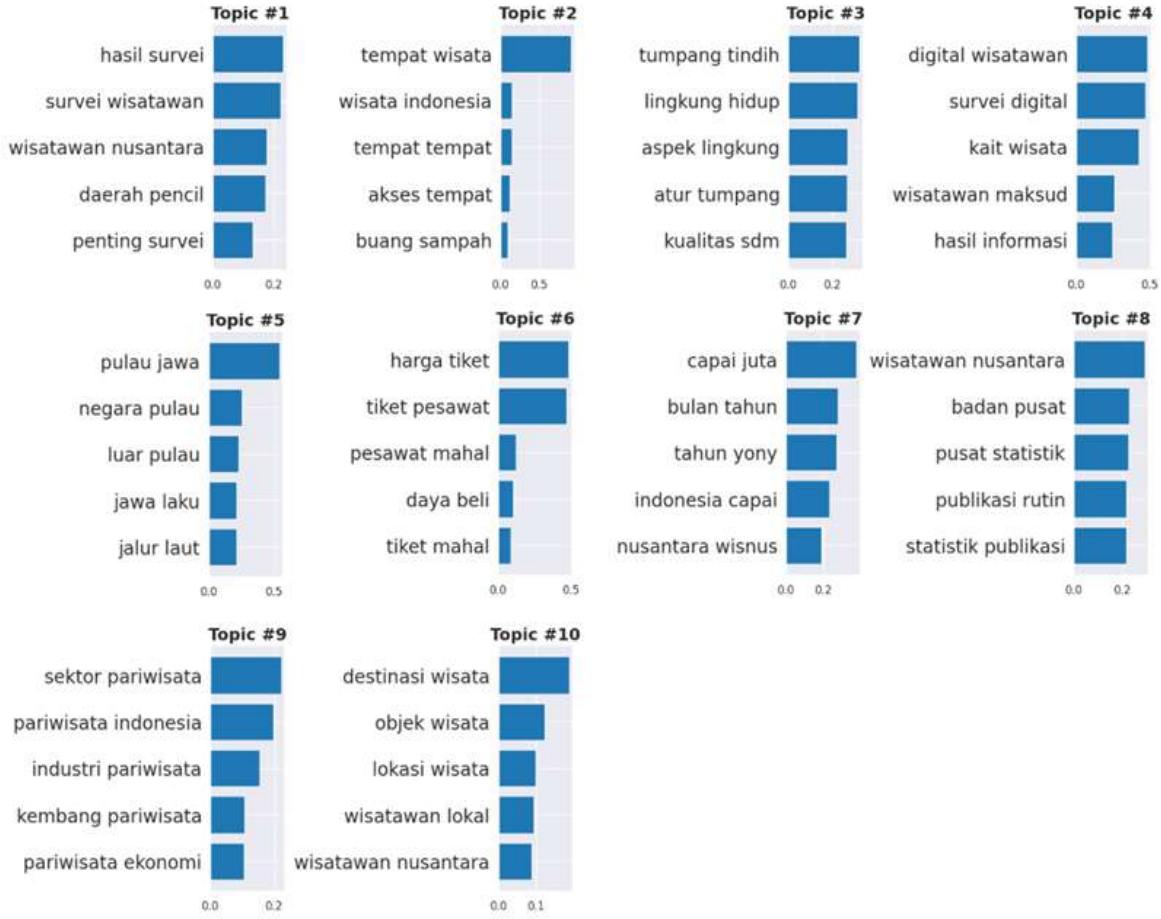


Figure 5. Top Terms per Topic (10 Topics × 5 Words) by NMF on Long Dataset

In contrast, Topics #9 and #10 reflect a macro perspective on tourism development. The first emphasizes industry aspects with the terms "tourism sector," "tourism industry," and "economic tourism," while the second highlights destination issues and the role of domestic tourists with the terms

"tourist destinations," "tourist attractions," and "local tourists." This separation of themes is quite clear, although there is potential for overlap, for example, between Topics #2 and #10, which both discuss destinations. These findings indicate that the model successfully maps major issues into more specific topics, which is further strengthened by expert interpretations in Table 11 to provide a more comprehensive context for subsequent analysis.

Table 11. Topic Interpretations and Top Keywords from NMF (Long Dataset)

Topic	Interpretation	Description	Keywords
1	Tourist Survey and Data Collection	Highlights the implementation of the domestic tourist survey, including the importance of reaching remote areas for data completeness.	survey results, tourist survey, domestic tourists, remote areas, survey importance (*hasil survei, survei wisatawan, wisatawan nusantara, daerah terpencil, penting survei*)
2	Quality and Accessibility of Tourist Destinations	Discusses access to destinations, cleanliness of tourist environments, and waste issues affecting tourist experiences.	tourist sites, Indonesian tourism, places, access to places, throw garbage. (*tempat wisata, wisata Indonesia, tempat tempat, akses tempat, buang sampah*)
3	Environment and Governance	Raises issues of tourism governance related to human resource quality, environmental sustainability, and regulatory overlap.	overlap, environment, environmental aspects, overlap regulation, human resource quality (*tumpang tindih, lingkungan hidup, aspek lingkungan, atur tumpang, kualitas SDM*)
4	Digitalization in Tourist Surveys	Focuses on the use of digital technology to facilitate tourist surveys and information processing.	digital tourists, digital survey, related tourism, tourist intention, survey results (*digital wisatawan, survei digital, kait wisata, wisatawan maksud, hasil informasi*)
5	Geographical Dimension (Islands & Transportation Routes)	Emphasizes the importance of geographical factors such as Java island, outer islands, and sea routes in tourist mobility.	Java Island, island nation, outside islands, Java trend, sea routes (*pulau Jawa, negara pulau, luar pulau, Jawa laku, jalur laut*)
6	Travel Costs & Transportation	Highlights the high price of airline tickets and their relation to tourists' purchasing power.	ticket price, plane ticket, expensive plane, purchasing power, expensive ticket (*harga tiket, tiket pesawat, pesawat mahal, daya beli, tiket mahal*)
7	Tourist Visit Statistics	Presents tourist visit data based on numbers, time periods, and annual achievement trends.	reach millions, month year, year on year, Indonesia reached, domestic tourists (*capai juta, bulan tahun, tahun yoy, Indonesia capai, nusantara wisnus*)
8	Data Publication & Official Statistics	Emphasizes the role of BPS in providing and publishing tourism statistics regularly.	domestic tourists, central agency, statistics center, regular publication, statistics publication (*wisatawan nusantara, badan pusat, pusat statistik, publikasi rutin, publikasi statistik*)
9	Tourism Industry & Economy	Describes tourism's contribution as a growing industry sector supporting the national economy.	tourism sector, Indonesian tourism, tourism industry, tourism development, tourism economy (*sektor pariwisata, pariwisata Indonesia, industri pariwisata, kembang pariwisata, pariwisata ekonomi*)
10	Destinations & Local Tourists	Discusses visits by local tourists to various objects and tourist locations in the archipelago.	tourist destinations, tourist attractions, tourist locations, local tourists, domestic tourists (*destinasi wisata, objek wisata, lokasi wisata, wisatawan lokal, wisatawan nusantara*)

Once the main topics were identified, the next step was to analyze the feedback on each topic group. Figure 6 shows an uneven distribution of respondents' responses. Topic #10 (Destinations & Local Tourists) topped the list with the highest number of feedbacks, followed by Topic #1 (Tourist Surveys

and Data Collection) and Topic #2 (Destination Quality and Accessibility). In contrast, several other topics such as Topic #4 (Digitalization of Tourist Surveys), Topic #5 (Geographic Dimension), and Topic #7 (Tourist Visit Statistics) received only a few responses, each below 50. This imbalance indicates that issues related to tourist destinations, service quality, and tourist surveys received more frequent attention in the feedback, thus requiring more attention as they likely represent some of the most relevant to respondents.

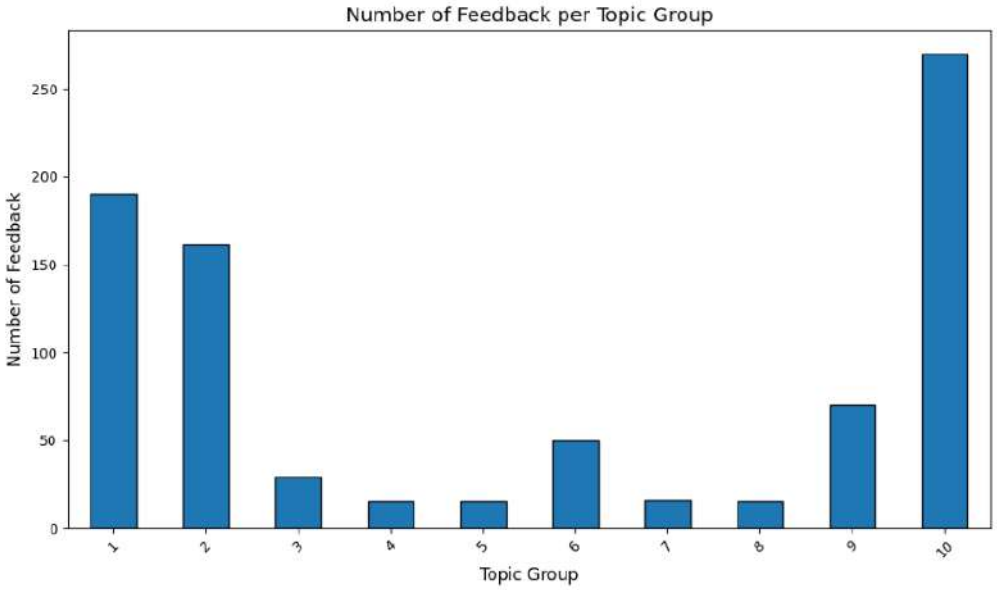


Figure 6. Number of Feedback Entries per Topic by NMF (Long Dataset)

3.2.3 Sentiment Analysis

Similarly to the medium dataset, feedback classified within the long text length category is analyzed using a range of sentiment analysis models, namely RoBerta, DistilBert, Bert, Albert, XLM-RoBERTa. The purpose of employing these models is to systematically categorize the feedback into three sentiment classes: positive, negative, and neutral. Table 12 provides an illustrative example of the analysis results, highlighting how each model may produce different sentiment labels when applied to the same feedback text.

Table 12. Example of Sentiment Analysis Results Using Various Models (Long Dataset)

Model	Sentiment	Feedback
RoBERTa	(-)	MasyaAllah, since 2024, the results of the BPS tourism to Madiun have been satisfactory. We took our family on a trip to Madiun. Enjoying the beauty of Madiun. The food and recreation areas are really economical and pocket-friendly. We are truly satisfied. Every trip in the city center of Madiun, we are treated to several beautiful statues that are truly similar to the original. The government is truly top notch. Vacations now don't have to go far anymore.
DistilBert	(+)	
Bert	(+)	
Albert	(-)	(*MasyaAllah dari tahun 2024 hasil BPS wisata ke Madiun hasilnya memuaskan. Kami mengajak keluarga jalan jalan ke madiun.. Menikmati indahnya kota madiun. Makanan dan tempat rekreasinya bener bener hemat dan ramah dikantong. Kami bener bener puas.setiap perjalanan dikota pusat madiunnya disuguhi beberapa patung patung yg bagus dan bener bener mirip banget dengan aslinya. Benerbener pemerintahanya Top markotop dah. Liburan sekarang gak perlu jauh jauh lagi*)
XLM-RoBERTa	(-)	

3.2.4 Model Agreement

In the sentiment analysis of the long-category dataset, the results presented in Table 13 demonstrate clear performance differences among the evaluated models. RoBERTa achieved the highest agreement with the majority vote at 84.38%, indicating its strong capability in capturing sentiment nuances in long Indonesian texts. BERT followed with an agreement score of 82.09%, showing that it remains effective in modeling contextual information, although its performance is slightly lower than that of RoBERTa. ALBERT and XLM-RoBERTa exhibited moderate performance, with agreement levels of 75.96% and 75.24%, respectively, reflecting a balance between model efficiency and, in the case of XLM-RoBERTa, multilingual generalization. DistilBERT, as a lightweight distilled model, recorded the lowest agreement at 65.75%, suggesting that while computationally efficient, model distillation reduces sentiment classification accuracy for longer texts.

Table 13. Model Agreement Results for Sentiment Analysis on Long Dataset

Model	Agreement with the majority	%
RoBERTa	**702**	**84.38**
DistilBert	547	65.75
Bert	683	82.09
Albert	632	75.96
XLM- RoBERTa	626	75.24

Figure 7. Number of Feedback Entries per Topic by NMF (Long Dataset)

After identifying the most consistent sentiment model, the next step was to examine the distribution of sentiment within each topic based on the predictions of the RoBERTa model. The analysis revealed the following results:

- Topic 10 (Destinations and Local Tourists) generated the most positive sentiment, reflecting respondents' appreciation for domestic tourist destinations. This topic was characterized by keywords such as tourist destinations, tourist attractions, and local tourists, confirming enthusiasm for the domestic tourism sector.
- Topic 2 (Quality and Accessibility of Tourist Destinations) and Topic 6 (Travel & Transportation Costs) also showed predominantly positive sentiment. However, in Topic 6,

complaints persisted regarding high travel costs or ticket prices, which impact respondents' purchasing power.

- Topic 3 (Environment and Governance) and Topic 5 (Geographic Dimension/Islands & Transportation Routes) displayed largely negative sentiment, indicating concerns regarding environmental issues, unequal access between regions, and limited infrastructure.
- Topic 1 (Tourist Survey & Data Collection) and statistical topics such as Topic 7 (Tourist Visit Statistics) and Topic 8 (Official Data & Statistics Publication) tended to be neutral, with respondent feedback being more informative regarding the survey mechanism and tourism data.

Overall, Topics 10, 2, and 6 were perceived positively and should be maintained, while Topics 3 and 5 should be prioritized for improvement due to numerous complaints regarding governance and limited access to the tourism sector.

Limitations

However, this study has several limitations that should be acknowledged. First, the dataset used in this research exhibited an imbalance across both sentiment categories and topic groups, which may have influenced the representativeness and stability of the sentiment distribution. Second, short texts were deliberately excluded from the analysis due to their sparse linguistic features and limited contextual information. Although this decision helped improve topic coherence and sentiment consistency, it may limit the generalization of findings to shorter forms of feedback. Third, the sentiment evaluation relied solely on the model agreement approach since no human-annotated ground truth was available for validation. While this approach provides a reliable estimation of model consistency, it cannot fully replace human judgment in assessing nuanced emotional expressions. Lastly, all models were trained and tested using domain-specific data related to the tourism sector, which may constrain the transferability of the results to other domains or datasets.

4. Conclusion

On the medium-sized dataset, the evaluation results showed that BERTopic was the superior model with the highest coherence score. This model was able to group respondent feedback into more structured topics, such as tourist facilities, surveys, domestic tourists, and tourism governance. The distribution of feedback revealed that the topic of tourist facilities received the most attention and was largely positive, indicating respondents' appreciation for the well-established destination management.

Furthermore, on the long-sized dataset, NMF performed more consistently than the other models. NMF was more effective at capturing a more complex vocabulary, resulting in a wider variety of topics, including travel costs, destination quality and accessibility, environmental issues, and tourism digitalization. Among these topics, tourist destinations were the most frequently discussed theme and were generally viewed positively, although some complaints about travel costs and access were still found.

It is also noteworthy that several topics within the medium and long text datasets contained only a small number of feedback entries because the phenomena they represented naturally occurred less frequently in the survey. Removing these topics could risk eliminating minor issues that are substantively significant, such as specific complaints or technical suggestions from respondents. Therefore, these topics were retained as part of the thematic diversity in the medium and long text datasets.

In terms of sentiment analysis, RoBERTa proved the most stable across both datasets, followed closely by ALBERT and BERT, while DistilBERT was the least consistent. The combination of topic models (BERTopic for medium-length text and NMF for long-form text) with RoBERTa as the sentiment model represents the most optimal configuration.

In conclusion, this study (1) demonstrates the usefulness of user feedback data in the context of Nusantara Tourist Survey, showing its potential to capture valuable insights for evaluation and development. Furthermore, (2) the analysis successfully identified key topics emerging from feedback content along with their sentiment trends, where tourist destinations and facilities were generally assessed positively, while travel costs, environmental concerns, and governance issues generated recurring negative sentiments. In addition, (3) the evaluation of several models showed that RoBERTa was the most consistent sentiment model, and when combined with BERTopic for medium-length texts and NMF for long-form texts, it provided the most optimal configuration. Finally, (4) these findings establish an analytical framework that can guide survey managers in improving instruments and

services, while also supporting policymakers in prioritizing actions such as enhancing accessibility, controlling costs, and strengthening tourism governance to ensure more participatory, adaptive, and evidence-based tourism development.

Ethics approval

The study was conducted in accordance with the ethical guidelines, and informed consent was obtained from all individual participants included in the study.

Acknowledgments

The authors would like to thank BPS-Statistics Indonesia for providing the data utilized in this research.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

The research received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Sulisetyo Puji Widodo: Conceptualization, Methodology, Investigation, Data Curation, Software Development**. Isnaeni Noviyanti:** Formal Analysis, Writing – Original Draft Preparation.

References

[1] D. Angelov, "Top2Vec: Distributed representations of topics," arXiv preprint arXiv:2008.09470, 2020.

[2] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in Proc. First Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA, Jul. 2010, pp. 80–88.

[3] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in Proc. 22nd Int. World Wide Web Conf. (WWW '13), Rio de Janeiro, Brazil, May 2013, pp. 1445–1456.

[4] J. Qiang, Z. Qian, Y. Li, Y. Yuan, and X. Wu, "Short text topic modeling techniques, applications, and performance: A survey," IEEE Trans. Knowl. Data Eng., vol. 34, no. 3, pp. 1427–1445, Mar. 2022.

[5] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in Advances in Information Retrieval (ECIR 2011), LNCS, vol. 6611. Berlin, Germany: Springer, 2011, pp. 338–349.

[6] Z. Ji, Z. Lu, and H. Li, "An information retrieval approach to short text conversation," arXiv preprint arXiv:1408.6988, 2014.

[7] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," in Proc. 20th ACM Int. Conf. Inf. Knowl. Manag. (CIKM '11), Glasgow, U.K., Oct. 2011, pp. 2333–2336.

[8] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF–IDF procedure," arXiv preprint arXiv:2203.05794, 2022.
[9] B. Bianchi, G. Lami, and F. Sebastiani, "CombinedTM: Combining topic models for improved short text modeling," Inf. Process. Manage., vol. 58, no. 2, 2021.
[10] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "The embedded topic model," arXiv preprint arXiv:1707.01417, 2020.
[11] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM), Shanghai, China, 2015, pp. 399–408.
[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[14] [P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[15] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv preprint arXiv:1908.10063, 2019.
[16] M. A. Jahin, M. N. Uddin, and M. A. Hossain, "TRABSA: Transformer and attention-based bidirectional LSTM for sentiment analysis," Sci. Rep., vol. 14, 2024.
[17] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," Comput. Linguist., vol. 34, no. 4, pp. 555–596, 2008.
[18] K. Gwet, Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Gaithersburg, MD: Advanced Analytics, LLC, 2014.
[19] K. Krippendorff, Content Analysis: An Introduction to Its Methodology. Thousand Oaks, CA: SAGE Publications, 2018.
[20] J. M. Serrano-Guerrero, J. A. Olivas, F. P. Romero, and E. Herrera-Viedma, "Sentiment analysis: A review and comparative analysis of web services," Inf. Sci., vol. 311, pp. 18–38, 2015.

DOI: https://doi.org/10.34123/jurnalasks.v18i1.973

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Ensemble Boosting Models for Forecasting Rice Prices in Indonesia

Muhammad Jimmy Saputra[1], Yeni Rahkmawati[2*], Selvi Annisa[3], Anne Mudya Yolanda[4]

[1,2,3]Univeristas Lambung Mangkurat, Banjarbaru, Indonesia, [4]University of Leeds, Leeds, United Kingdom
**Corresponding Author: E-mail address: yeni.rahkmawati@ulm.ac.id*

ARTICLE INFO

Article history:
Received 6 May, 2026
Revised 15 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Rice Price; Ensemble Boosting; GBM; LightGBM; Forecasting

Abstract

Introduction/Main Objectives: Rice is a key staple commodity influencing food security and inflation in Indonesia, making accurate price forecasting essential. In this study, we aim to compare ensemble boosting models and identify the best-performing model for rice price prediction. **Background Problems**: Notably, rice prices exhibit non-linear patterns over time, while classical statistical methods have limitations in capturing such complexities, resulting in suboptimal forecasting performance. **Novelty**: This study proposes a lag-based approach that uses lag variables as the only predictors, arranged across multiple input schemes to flexibly capture historical patterns without external variables. **Research Methods**: Daily national medium rice price data (Jan 2021–Jan 2026) from the National Food Agency are modeled using Gradient Boosting Machine (GBM) and LightGBM, with hyperparameter tuning via Optuna. The forecasting framework relies exclusively on significant lag variables without incorporating exogenous factors. Model performance is evaluated using RMSE, MAE, and MAPE. **Findings/Results**: LightGBM with optimized hyperparameters achieves the best performance (RMSE = 66.389; MAE = 50.213; MAPE = 0.362%). Furthermore, forecasts for the next 89 days indicate stable prices around Rp13,360–Rp13,395/kg, with no significant fluctuations.

1. Introduction

Rice is a fundamental food source for a majority of global population. Its consumption are found is concentrated in Asia, Sub-Saharan Africa, and South America, where it plays a crucial role in meeting dietary energy and carbohydrate requirements. Indonesia ranks fourth globally among rice-consuming countries, with consumption reaching 35.367 million tons in 2020 [1]. Moreover, data from the National Socioeconomic Survey (Susenas) conducted in March 2025 indicate that rice, including local rice, premium-quality rice, and imported rice, has the highest consumption participation rate in Indonesia at 99.07%, highlighting the strong reliance of Indonesian society on rice as a daily staple food [2].

Rice is a strategic food commodity whose prices are classified within the volatile food group and are highly sensitive to supply disruption and market dynamics, thereby directly influencing year-on-year inflation in Indonesia [3]. This relationship is illustrated in Figure 1, where rice consistently appears as one of the main contributors to inflation, although its impact may fluctuate overtime. Typically, increases in rice prices are driven by imbalances between supply and demand, which are caused by factors such as El Niño, rising non-subsidized fertilizer prices, limited availability of subsidized fertilizers, and irrigation damage [4]. In response to these challenges and to maintain price stability and

food security, the government implements the second mission of Asta Cita through strategies such as agrarian reform, fertilizer assurance, irrigation improvement, distribution efficiency, import policies, and food reserve management [5].

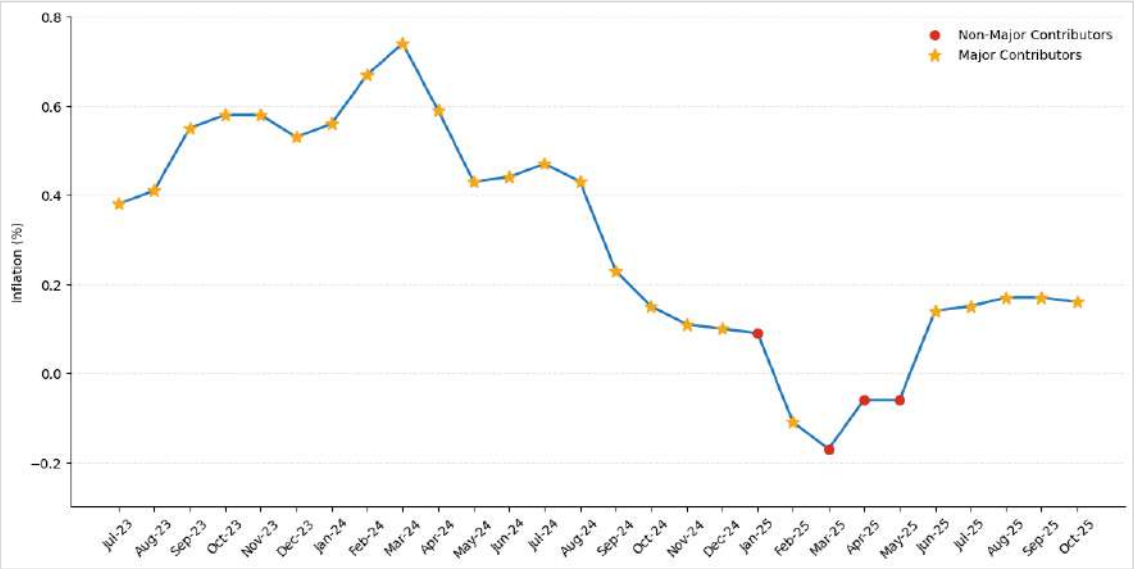


Figure 1. Contribution of rice commodity to year-on-year inflation

Government efforts to maintain long-term price stability can be supported by rice price forecasting, which enables the anticipation of future trends and the formulation of more effective policy responses [6]. Previous studies on rice price forecasting have employed various traditional statistical models, including Double Exponential Smoothing [7], Linear Regression [8], ARIMA [9], and SARIMA [10]. However, advancements in science and technology have led to the increasing adoption of modern statistical approaches, particularly machine learning, due to their ability to capture non-linear relationships and complex interactions among variables that are often difficult to handle using traditional methods [11]. Several machine learning models applied in forecasting studies include Random Forest, Support Vector Machine (SVM) [12], Decision Tree (DT), Gradient Boosting Machine (GBM) [13], and Extreme Gradient Boosting (XGBoost) [14].

One of the most widely developed machine learning approaches is ensemble boosting, a sequential learning technique that combines multiple weak learners to construct a strong predictive model, where each successive model focuses on correcting the errors of its predecessor [15]. Common implementations of this method include Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (LightGBM). GBM builds decision trees iteratively to improve accuracy [16], but its performance becomes less efficient when applied to large datasets due to the need to processes the entire dataset at each iteration. As an advancement of GBM, LightGBM enhances computational efficiency by focusing only on important data and combining certain features without sacrificing accuracy [17].

Various studies have explored the capability of ensemble boosting models across different fields. For instance, GBM has been shown to produce wind power generation predictions with the lowest nRMSE among boosting models [18]. Similarly, in the agricultural commodity sector, LightGBM has consistently outperformed six other models, namely ARIMA, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN), in predicting agricultural product prices [19]. Taken together, these findings justify the selection of GBM and LightGBM as a suitable methods for forecasting rice prices, which exhibit substantial volatility.

This study adopts a distinct approach by utilizing lag variables as the sole predictors, which are combined into multiple input schemes to represent historical information. Each scheme is evaluated to examine its effect on model performance in capturing rice price movement patterns. To further enhance model performance, hyperparameter tuning is performed using Optuna. This framework allow for flexible exploration of temporal relationships without relying on a single variable combination. This study intends to support the government in formulating food self-sufficiency policies through more accurate rice price forecasting.

2. Material and Methods

2.1. Data Source

The data used in this study are secondary data obtained from the official website of the National Food Agency. The dataset is a daily time series covering the period from January 2021 to January 2026, with a total of 1,857 observations. The variable used is the national medium rice price at the consumer level, measured in Indonesian Rupiah per kilogram (IDR/kg), which is subsequently modeled and forecasted. Medium rice refers to rice with a moderate quality level that meets the standards set by the National Food Agency Regulation No. 2 of 2023. Its main characteristics include a minimum milling degree of 95%, a maximum moisture content of 14%, a maximum of 2% broken fragments (menir), a maximum of 25% broken grains, a maximum of 4% other rice grains, a maximum of 1 unhusked grain per 100 grams, and a maximum of 0.05% foreign matter [20] The selection of medium rice is based on its role as the most widely consumed type of rice among low- to middle-income populations, particularly households slightly above the poverty line [21]. Therefore, fluctuations in medium rice prices can reflect changes in society's purchasing power and are directly linked with economic stability and food policy.

2.2. Research Framework

This study aims to develop a rice price forecasting model to understand price movement patterns and support food policy decision-making. The research stages include data collection, data preprocessing, construction of input variable, variable selection, model development, and evaluation and interpretation of results. Visualization, analysis, and modeling were conducted using Python 3.12.13 in the Google Colab environment, utilizing libraries such as NumPy, pandas, scikit-learn, LightGBM, statsmodels, Optuna, and Matplotlib. The expected outcome is the best-performing model for rice price forecasting, along with prediction results that illustrate future price trends, as presented in the research framework in Figure 2.

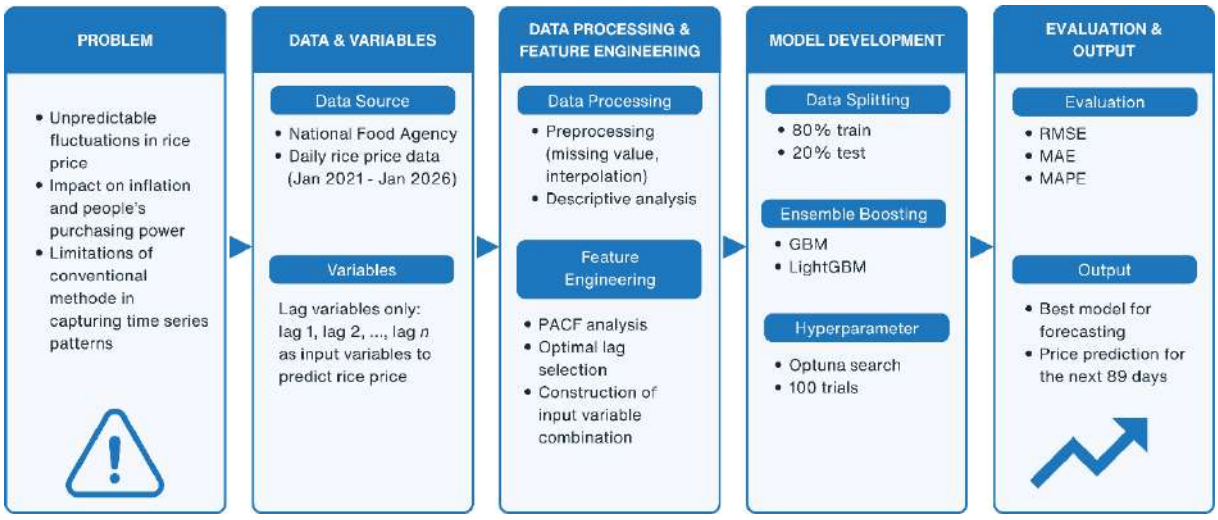


Figure 2. The research framework

2.3. Linear Interpolation

Linear interpolation is a commonly used method for handling missing data, with its primary advantage being its ability to preserve the underlying trend without introducing unnecessary fluctuations [22]. By applying this approach, data consistency can be maintained while reducing potential bias that may affect subsequent analysis, particularly in time series contexts where continuity between observations is essential. The method operates by constructing a straight line connecting two known data points surrounding the missing value [23], allowing the unknown value to be estimated as a proportion between them. Despite its simplicity, this approach is effective because it assumes a linear relationship between the observations before and after the missing interval.

$$y_t = y_{t-1} + \frac{y_{t+1} - y_{t-1}}{x_{t+1} - x_{t-1}}(x_t - x_{t-1}) \quad (1)$$

In Equation (1), y_t represents the estimated value at time t, y_{t-1} and y_{t+1} denote the observed values at the previous and subsequent time points, respectively, while x_t, x_{t-1}, and x_{t+1} correspond to the time indices of the missing, previous, and subsequent observations.

2.4. Partial Autocorrelation Function

The Partial Autocorrelation Function (PACF) measures the correlation between observations Y_t and Y_{t+k} after removing the linear influence of all intermediate observations, namely $Y_{t+1}, Y_{t+2}, \dots, Y_{t+k-1}$. In other words, PACF captures the direct correlation between two data points separated by lag k, while controlling for the linear dependencies of the intervening variables [24]. Mathematically, this function is expressed as the following conditional correlation:

$$\phi_{kk} = \text{corr}(Y_t, Y_{t+k} \mid Y_{t+1}, \dots, Y_{t+k-1}) \tag{2}$$

Here, ϕ_{kk} denotes the PACF coefficient at lag k, Y_t represents the time series value at time t, and Y_{t+k} is the value at time $t+k$, while k indicates the lag and t represents time. The sample PACF can be mathematically defined as:

$$\hat{\phi}_{kk} = \frac{\hat{\rho}_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_j} \tag{3}$$

with the recursive relation:

$$\hat{\phi}_{kj} = \hat{\phi}_{k-1,j} - \hat{\phi}_{kk}\hat{\phi}_{k-1,k-j}, j = 1,2, \dots, k-1. \tag{4}$$

In these expressions, $\hat{\phi}_{kk}$ represents the estimated PACF value at lag k, while $\hat{\phi}_{k-1,j}$ denotes the estimated PACF at lag j from lag $k-1$. Furthermore, $\hat{\rho}_k$ is the estimated ACF at lag k, $\hat{\rho}_{k-j}$ is the estimated ACF at lag $k-j$, and $\hat{\rho}_j$ is the estimated ACF at lag j. Here, k indicates the lag, and j is the lag index.

The selection of input variables is crucial for forecasting model performance. Before training, an initial set of input must be determined. The PACF is used to identify these initial inputs, where variables at specific lags are considered potential predictors if their PACF values fall outside the confidence interval bounds, indicating statistical significance. The corresponding lags highlight significant autocorrelation, which can be visualized with the PACF plot [25]. By leveraging PACF and considering other relevant information, an initial set of input variables for different subseries can be approximated [26].

2.5. Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a supervised learning algorithm based on ensemble boosting that builds models sequentially to optimize a specific objective function. At each iteration, the model improves previous prediction errors by fitting weak learners, typically decision trees, to the residual errors. A weak learner is a simple model with predictive performance slightly better than random guessing, but through sequential boosting its performance can be significantly improved [27]. The main objective of GBM is to minimize the loss function by combining weak learners in an additive manner [28].

Figure 3 illustrates that GBM constructs a new weak learner by following the gradient direction to reduce residuals. This process uses multiple regression trees as weak learners and applies a forward stagewise learning approach to iteratively minimize residuals from the previous iteration. At the same time, the algorithm adjusts sample weights across iterations to generate subsequent weak learners, allowing the model to better capture patterns in the data and improve prediction accuracy at each step. To obtain more accurate predictions, all weak learners produced in each iteration are then combined linearly through an additive model [29].

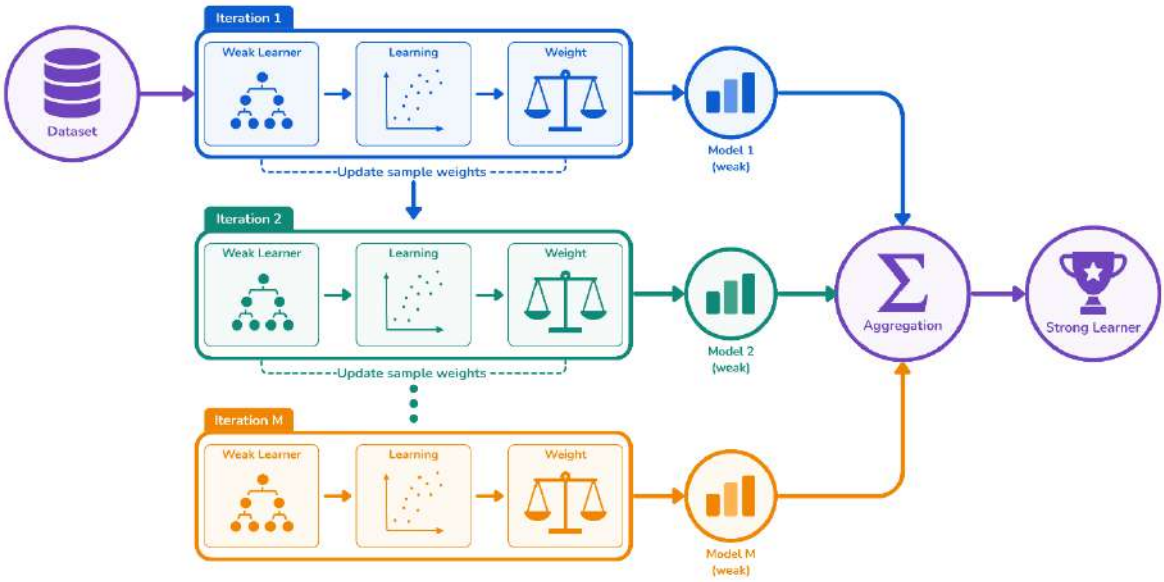


Figure 3. GBM workflow

The process begins with the initialization of a base model that minimizes error on the training data [30], followed by the computation of pseudo-residuals to construct weak learners at each iteration. The model is updated iteratively by adding the contribution of each new weak learner until a stable final model is obtained in the form of an additive regression function:

$$\hat{y}_t = F_0(x_t) + \sum_{m=1}^{M} \nu \cdot h_m(x_t) \tag{5}$$

In equation (5), $\hat{y}_t$ represents the predicted value for the t-th observation, $F_0(x)$ is the initial prediction before boosting begins, ν is the learning rate that controls the contribution of each weak learner, $h_m(x_t)$ denotes represents the weak learner at iteration m evaluated using the input features at time t, and M is total number of boosting iterations used to construct the model.

2.6. Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is a decision tree-based model that extends the ensemble boosting approach with high computational efficiency. Its main advantage lies in the use of a *leaf-wise tree growth* strategy, which expands the node that yields the largest error reduction in a greedy manner, allowing faster training compared to conventional methods [31]. In contrast to *level-wise tree growth*, which grows the tree layer by layer where the number of nodes at depth D reaches 2^D, the leaf-wise approach focuses on error optimization, leading to improved speed and model performance [32].

During the learning process, each tree is constructed to minimize the loss function while controlling model complexity through regularization, ensuring a balance between accuracy and model stability. This approach also utilizes gradient and Hessian information to improve the quality of model updates at each iteration. In addition, LightGBM incorporates several acceleration techniques such as histogram-based algorithms, Gradient-Based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB), which enable efficient processing of large-scale data without degrading model performance [33].

In general, the final prediction of LightGBM is obtained by summing the contributions of all decision trees built sequentially, as follows:

$$\hat{y} = \sum_{i=1}^{I} T_i\,(X, \theta_i) \tag{6}$$

In this equation, $\hat{y}$ represents the final predicted value, T_i denotes the i-th decision tree, X is the set of input variables, θ_i represents the parameters learned by the i-th tree, and I is the total number of trees in the model.

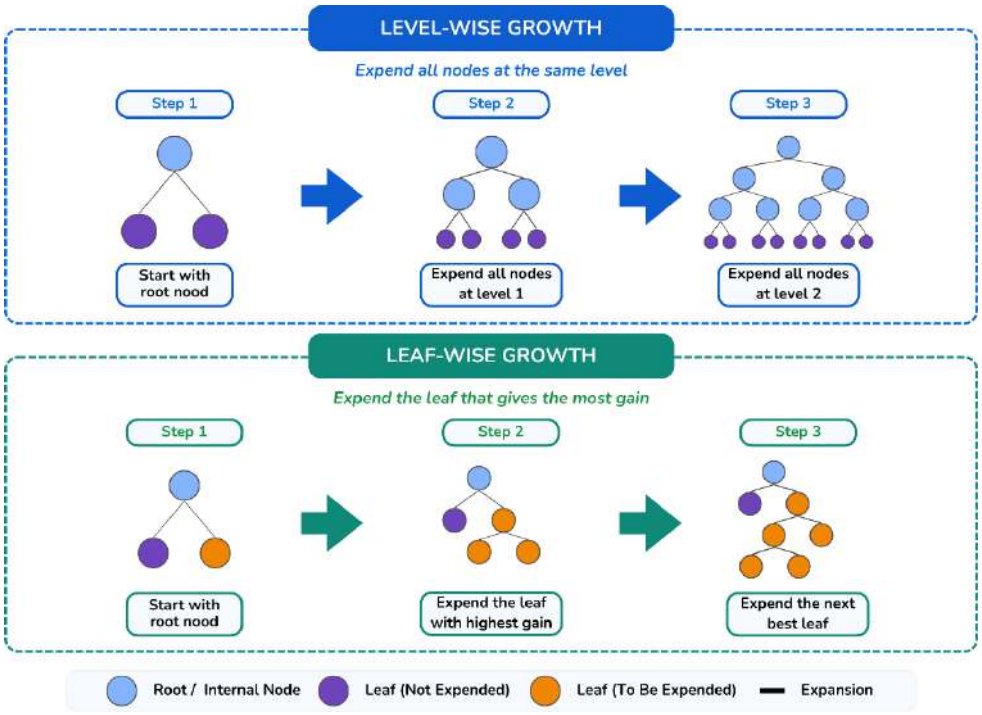


Figure 4. Level-wise and leaf-wise growth mechanism

2.7. Optuna Search

Optuna is an automated framework for optimization hyperparameters in machine learning models. This library enables a flexible hyperparameter search process through a define-by-run approach, allowing the search space to be dynamically defined within the code according to the desired conditions. Hyperparameters themselves are parameters that are set prior to the model training process and remain constant throughout training. These parameters play a crucial role in determining the model's learning capacity, generalization ability, and computational cost. Compared to traditional optimization methods such as grid search and random search, Optuna offers a more flexible and efficient approach to exploring complex search spaces [34].

In its optimization process, Optuna commonly employs the Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian optimization method operating within the Sequential Model-Based Optimization (SMBO) framework. This algorithm divides the parameter space into two probability distributions: one representing trials with good performance and another representing the remaining trials. Candidate parameters are then selected by maximizing the ratio between these two distributions, guiding the search toward more promising regions of the parameter space. This approach enables more efficient exploration, particularly in high-dimensional optimization problems [35]. Furthermore, Optuna provides several advantages that enhance optimization efficiency. The Bayesian optimization algorithm leverages results from previous trials to guide the search process, allowing optimal solutions to be identified with fewer trials compared to conventional methods. Optuna also incorporates a pruning mechanism that monitors the performance of each trial in real time and terminates underperforming trials at an early stage. As a result, computational resources can be focused on the most promising hyperparameter combinations, making the optimization process more efficient [36].

2.8. Model Evaluation

The evaluation methods used in this study include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). MAE measures the average magnitude of errors between predicted and actual values without considering their direction, making it easy to interpret since it has the same unit as the original data [37]. RMSE is the square root of the average squared differences between predicted and observed values, making it more sensitive to large errors and useful for detecting high deviations. Meanwhile, MAPE expresses the error in percentage terms, which

facilitates interpretation as it reflects the average relative error compared to actual values [38]. The smaller the values of these three metrics, the better the model performance.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \quad (8)$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \quad (9)$$

In these equations, y_t represents the actual value at time t, $\hat{y}_t$ is the predicted value at time t, and n denotes the total number of observations. Lewis's benchmark is commonly used to interpret MAPE values, where values below 10% indicate highly accurate forecasts, values between 10% and 20% indicate good forecasting performance, and values between 20% and 50% suggest reasonable predictive accuracy [39].

It should be noted that the forecasting models developed in this study rely exclusively on significant lag variables derived from historical rice price data. Consequently, the models do not account for exogenous factors that may influence price movements. While this univariate approach enables a focused assessment of the predictive capability of lag-based information, it may limit the models' ability to capture influences beyond the observed historical patterns.

3. Results and Discussion

3.1. Data Exploration

After data pre-processing, including the handling of missing values using linear interpolation, an exploratory data analysis was conducted to examine the characteristics of rice prices in Indonesia, including trends, variability, and price dynamics overtime.

Table 1. Descriptive statistics of rice prices (IDR/kg)

Mean	Median	Standard Deviation	Minimum	Maximum
12,308.71	11,985	1,383.67	7,850	14,595

Table 1 shows that the average rice price is IDR 12,308.71/kg, while the median value is IDR 11,985/kg. The relatively small difference between meand and media suggest that the price distribution is approximately symmetric. The standard deviation of IDR 1,383.67 indicates a moderate level of variability around the mean, reflecting price fluctuations over time. The minimum rice value of IDR 7,850/kg was recorded on March 12, 2021, which is associated with incomplete interprovincial data reporting during that period. In contrast, the maximum price of IDR 14,595/kg occurred on April 11, 2024, two days after Eid al-Fitr 1445 H, likely driven by increased demand around religious holiday. These statistics highlight the sensitivity of rice price to both data condition and seasonal demand factors.

Figure 5 illustrates the movement of rice prices in Indonesia over the study period. The prices series exhibits distinct phases, beginning with pronounced short-term fluctuation in early 2021, which are likely associated with incomplete regional data reporting during that period. From mid-2021 to mid-2022, rice prices display a relatively stable pattern, indicating balanced market conditions with limited votality. However, starting in late 2022, a persistent upward trend becomes evident, reflecting increasing price pressure over time. A notable price peak is observed in early 2024, influenced by production disruptions caused by extreme weather conditions and the El Niño phenomenon, which suppressed supply amid high demand. This increase was also driven by rising non-subsidized fertilizer prices, reduced subsidies, land conversion, and global factors such as increased rice imports by Indonesia and export restrictions by India [40]. After a temporary decline, prices rose again and reached another peak in mid-2025, before correcting in September 2025, which coincided with government intervention aimed at stabilizing rice prices [41].

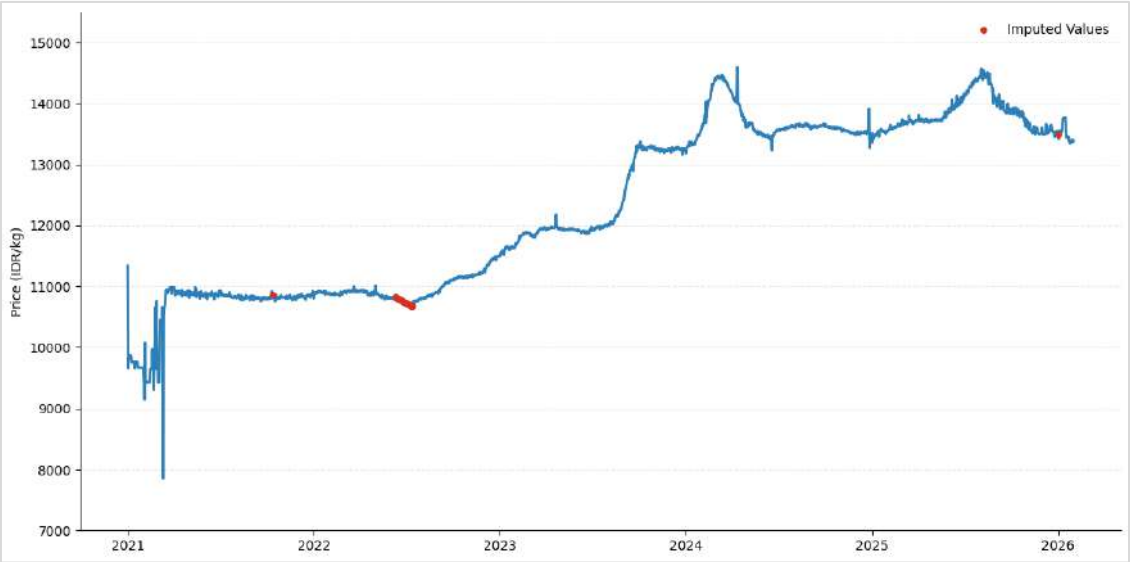


Figure 5. Daily rice price movement

3.2. Input Variable Selection

Variable selection in this study was conducted using Partial Autocorrelation Function (PACF) analysis to identify lags that have a direct influence on rice price movements. The analysis was carried out up to lag 30, corresponding to a monthly cycle, to capture short-term temporal relationships without being affected by indirect dependencies across lags. Significant lags are those whose PACF values fall outside the confidence interval, indicating a meaningful contribution in explaining rice price variation.

As illustrated in Figure 6, the presence of significant positive autocorrelation at early lags, particularly lags 1 to 4, indicates that current prices are strongly influenced by prices from recent periods. Autocorrelation values tend to decrease with increasing lag, indicating a weakening temporal dependence. At higher lags, the pattern becomes unstable, alternating between positive and negative values. Although some lags up to lag 30 remain significant, their influence is relatively small compared to early lags, thus contributing less to the model.

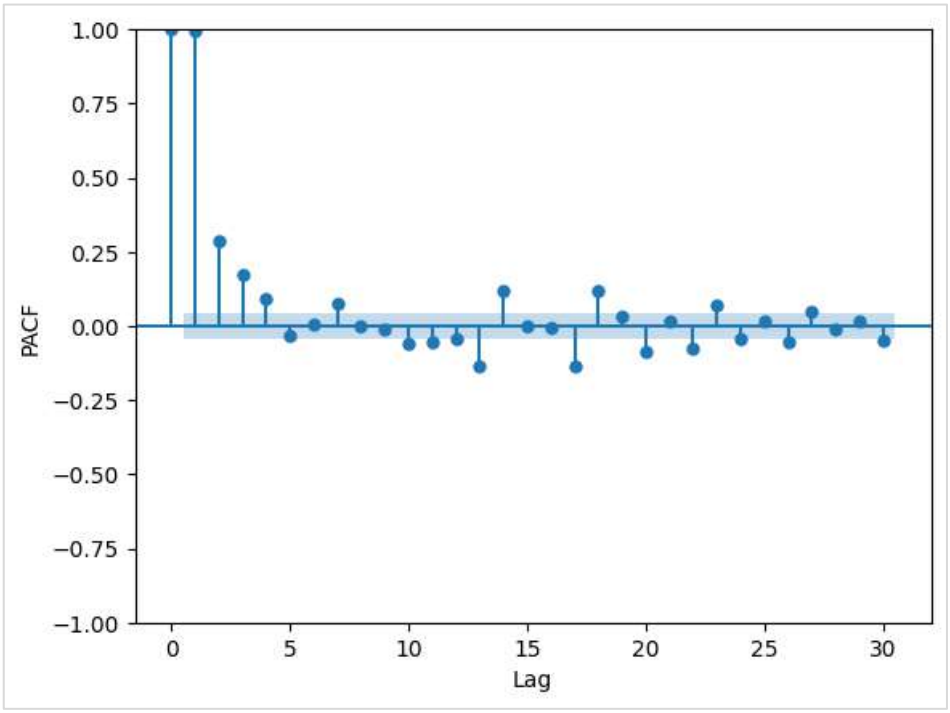


Figure 6. PACF plot of rice price data

Based on PACF analysis, 18 significant lags were identified. However, to maintain modeling efficiency, further selection was performed by choosing the 10 lags with the highest PACF values, namely lags 1, 2, 3, 4, 7, 13, 14, 17, 18, and 20. This selection aims to retain important information while reducing model complexity.

Table 2. Input variable schemes based on lag combinations

Scheme	Lag Variables
Scheme 1	1
Scheme 2	2
Scheme 3	3
…	…
Scheme 1021	1, 3, 4, 7, 13, 14, 17, 18, 20
Scheme 1022	2, 3, 4, 7, 13, 14, 17, 18, 20
Scheme 1023	1, 2, 3, 4, 7, 13, 14, 17, 18, 20

The selected lag variables were then used to form various combinations of input variables in the modeling process. With 10 lag variables, the total number of possible combinations is 1,023 schemes, derived from all possible subsets excluding the empty set ($2^{10} - 1$). Each scheme represents a different combination of lag variables used as model input. This approach aims to evaluate various input structures to obtain the most optimal variable combination without assuming specific relationships between lags.

3.3. Model Development

Model development was carried out using two ensemble boosting approaches, namely Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (LightGBM), by utilizing combinations of selected lag variables. Model performance was evaluated using Expanding Window Cross-Validation with five folds and the Root Mean Square Error (RMSE) metric to maintain the temporal structure of the data. The evaluation results indicate that both models achieve their best performance with lag combinations dominated by short-term lags. A summary of the ten best schemes for each model is presented in Table 3.

Table 3. Top ten schemes based on RMSE CV for GBM and LightGBM

GBM			LightGBM		
Scheme	Lag Variables	RMSE CV	Scheme	Lag Variables	RMSE CV
392	1,2,3,7,13	291.7	177	1,2,3,7	333.619
183	1,2,4,7	293.379	386	1,2,3,4,7	333.74
638	1,2,3,4,7,13	293.676	204	1,3,4,7	334.214
58	1,2,7	293.775	58	1,2,7	334.368
393	1,2,3,7,14	294.092	183	1,2,4,7	334.449
407	1,2,4,7,13	294.162	71	1,4,7	334.656
215	1,3,13,14	294.169	393	1,2,3,7,14	334.689
60	1,2,14	294.214	65	1,3,7	334.971
397	1,2,3,13,14	294.499	179	1,2,3,14	335.286
653	1,2,3,7,13,14	294.517	176	1,2,3,4	335.725

Table 3 shows that the GBM model produces lower RMSE CV values compared to LightGBM. The best-performing GBM scheme is obtained using a combination of lags 1, 2, 3, 7, and 13, achieving an RMSE CV of 291.70, while the optimal LightGBM scheme uses lags 1, 2, 3, and 7 with an RMSE CV of 333.62. The dominance of lags 1 and 2 in both models indicates that the most recent historical information plays a key role in modeling rice prices. In addition, the inclusion of lag 7 suggests the presence of a weekly periodic pattern in rice price movement.

Table 4. Hyperparameter search space for GBM and LightGBM

Model	Hyperparameter	Default Value	Candidate Range
GBM	n estimators	100	100–500 (step of 50)
	learning rate	0.1	0.005–0.2
	max depth	3	3–10
LightGBM	n estimators	100	100–500 (step of 50)
	learning rate	0.1	0.005–0.2
	max depth	-1	3–10

To improve model accuracy, hyperparameter optimization was performed using Optuna with the search space presented in Table 4. The parameter n_estimators controls the number of trees in the boosting process, learning_rate controls the contribution of each tree to the final model, and max_depth controls model complexity. The optimization process was conducted over 100 trials to obtain the parameter combination with the lowest RMSE CV.

Table 5. Optimal hyperparameter configuration and RMSE CV Results for GBM and LightGBM

Model	n estimators	learning rate	max depth	RMSE CV
GBM	350	0.1351	4	285.73
LightGBM	100	0.0953	10	334.05

These results show that hyperparameter optimization in GBM significantly improves model performance by reducing the RMSE CV compared to the initial configuration. On the other hand, optimization in LightGBM does not lead to performance improvement, indicating that its default configuration is already sufficient in capturing the main data patterns. Results presented in Table 5 indicated that the GBM model outperforms LightGBM in modeling rice prices.

3.4. Evaluation and Forecasting

Model performance evaluation was conducted to compare the performance of GBM and LightGBM under both default settings and optimized hyperparameter. The evaluation employed RMSE, MAE, and MAPE metrics on training and testing datasets, while also accounting for computational time efficiency.

Table 6. Model performance evaluation on training and testing data

Model	Training Data			Testing Data			Runtime (sec)
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	
GBM Default	42.750	22.004	0.190%	70.404	51.276	0.369%	0.393
GBM Optuna	**12.460**	**9.136**	**0.077%**	73.651	53.870	0.388%	1.751
LightGBM Default	83.524	25.845	0.233%	67.161	50.774	0.366%	0.056
LightGBM Optuna	83.978	26.492	0.239%	**66.389**	**50.213**	**0.362%**	**0.045**

As demonstrated in Table 6, the GBM Optuna model achieves the best performance on the training data, with substantially lower RMSE, MAE, and MAPE values than all other models. This significant reduction in error demonstrates that hyperparameter optimization effectively enhances the model's ability to learn and fit the underlying patterns in the training data. The improvement is particularly notable compared to the GBM default configuration, suggesting that parameter tuning plays a crucial role in maximizing the potential of the GBM algorithm. On the other hand, LightGBM shows relatively stable performance before and after optimization, with only marginal changes in evaluation metrics. This suggests that the default configuration of LightGBM is already well-suited to the data structure and does not benefit significantly from further tuning in this case. Based on Lewis's benchmark, all models can be categorized as having highly accurate forecasting performance, as all testing MAPE values are below the 10% threshold. From a computational efficiency perspective, LightGBM clearly outperforms GBM. The training time required by LightGBM is considerably shorter, even after optimization, highlighting its advantage for efficiently handling large-scale data. In contrast, optimized GBM model

exhibits longer training time, reflecting the computational cost of their iterative parameter refinement and exhaustive learning strategy.

However, a different pattern emerges when evaluating the models on the test data. The LightGBM model, especially with Optuna optimization, produces the lowest RMSE, MAE, and MAPE, indicating superior generalization performance. This means that LightGBM is better at capturing the true underlying patterns in the data and applying them effectively to unseen observations. In contrast, the GBM model experiences a noticeable decline in performance when moving from training to testing data. Despite achieving very low errors during training, its higher error values on the test data suggest that the model may have overfitted the training data, learned noise or overly specific patterns that do not generalize well. This interpretation is further supported by the Time Series Cross-Validation results, which show that validation RMSE values are higher than those in training. This discrepancy highlights the importance of generalization in a univariate forecasting framework based solely on significant lag variables. Models that perform exceptionally well during training may not necessarily maintain the same level of accuracy on unseen data. The stronger performance of LightGBM with Optuna optimization on the test data suggests that it was better able to produce consistent predictions beyond the training period.

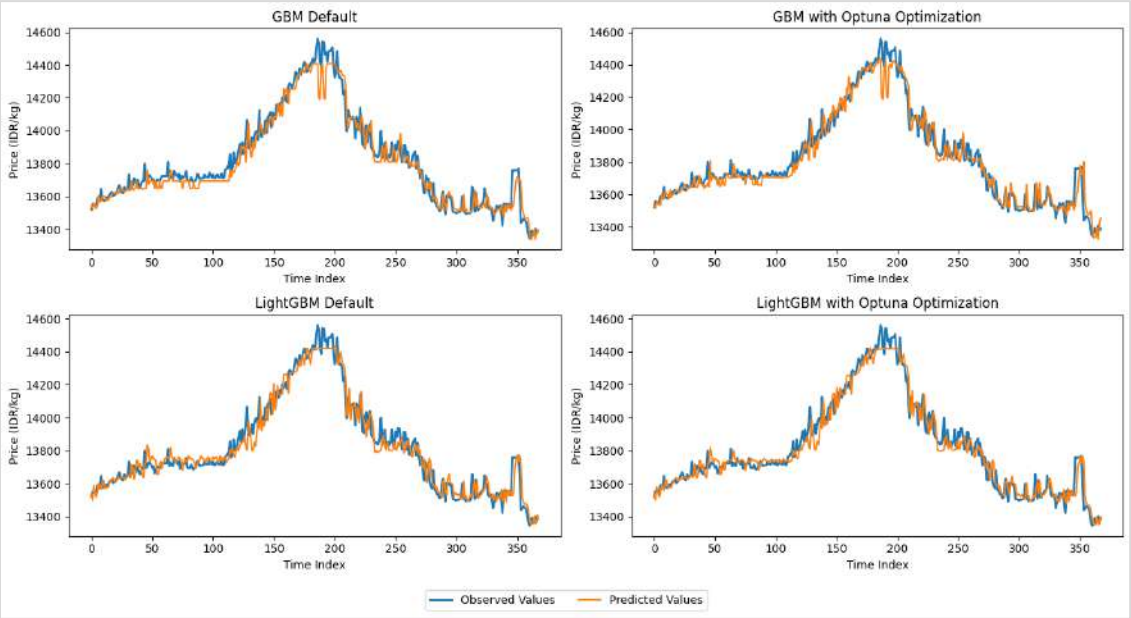


Figure 7. Comparison of GBM and LightGBM performance on testing data

The visualization in Figure 7 provides a clearer picture representation of how each model follows the movement of rice prices over time. Overall, both models are able to capture the general direction of price changes, particularly during periods characterized by gradual trend. However, differences between models become apparent during episodes of sharper price movements, such as sudden increase decreases. LightGBM tends to generate smoother and more stable over time. Its estimated values remain close to the actual data not only under normal conditions but also around turning points, including periods when prices peak and subsequently decline. This shows that LightGBM is able to adjust its predictions more consistently when the underlying price of the data shifts. Conversely, GBM shows larger gaps between predicted and actual values during certain periods. This is most visible after the price reaches a peak, where the model reacts more slowly to the downward trend. As a result, the predictions tend to lag behind actual price movements, creating wider prediction errors. This indicates that GBM faces greater difficulty in adapting to rapid changes in trend direction.

Based on the overall evaluation results, the LightGBM model with Optuna optimization is selected as the best-performing model. While GBM achieves very low error values on the training data, the decline in its performance on the testing data highlights the importance of generalization when evaluating forecasting models. LightGBM, in contrast, maintains more consistent performance between training and testing data. Moreover, it requires less computation time, making it more efficient. These factors indicate that LightGBM is more reliable for practical forecasting of rice prices.

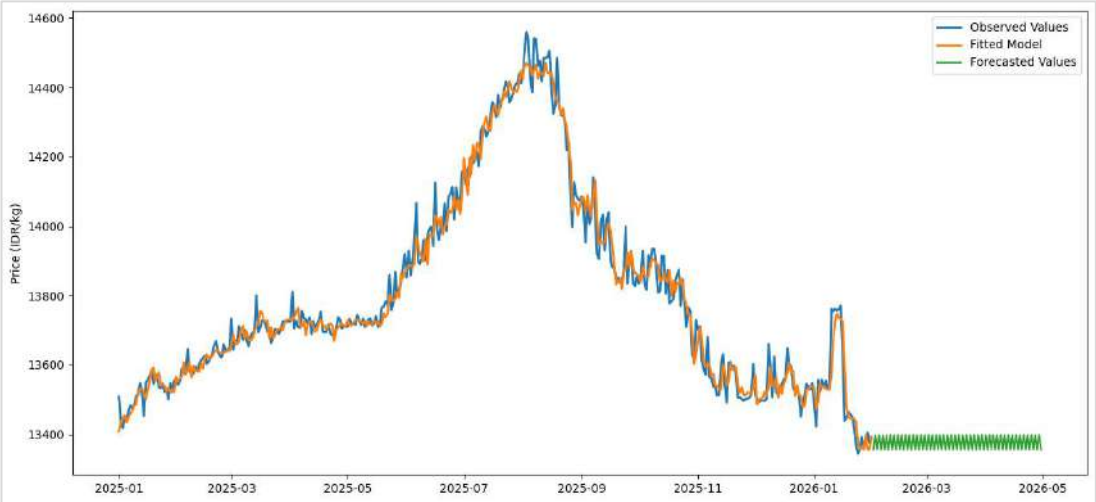


Figure 8. Model fitting and rice price forecasting results

Figure 8 illustrates the comparison between actual values, fitted values, and forecast results. The model demonstrates a strong ability to replicate historical patterns, as indicated by the close alignment between fitted values and actual observations. In the forecasting period, the predicted values show a relatively stable trend without extreme fluctuations. This suggests that, in the short term, rice prices are expected to remain relatively stable with only minor variations. The absence of sharp increases or decreases indicates that no strong signals of significant market disruption are detected based on historical patterns learned by the model.

4. Conclusion

The evaluation results show that the Gradient Boosting Machine (GBM) model achieves very low error values on the training data, particularly after hyperparameter optimization using Optuna. However, this strong in sample performance does not translate well to the testing data, indicating a tendency toward overfitting. In contrast, the Light Gradient Boosting Machine (LightGBM) model demonstrates more stable and consistent performance on out-of-sample data, reflecting superior generalization capability.

Based on the evaluation results on the testing data, the LightGBM model with hyperparameter optimization achieves the best performance, with an RMSE of 66.389, MAE of 50.213, and MAPE of 0.362%. In addition to higher accuracy, LightGBM also exhibits greater computational efficiency, with a significantly faster training time compared to the GBM model. Therefore, the LightGBM Optuna model is selected as the best model for forecasting rice commodity prices in Indonesia. Rice price forecasting using the LightGBM Optuna model for the 89-day period (February to April 2026) indicates that prices are expected to remain relatively stable in the short term, ranging from IDR 13,360/kg to IDR 13,395/kg. This suggests that over the next three months, rice prices are predicted to move within a narrow and controlled range, with no indication of extreme price changes based on the historical patterns learned by the model.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data from National Food Agency.

Credit Authorship

Muhammad Jimmy Saputra: Conceptualization, Methodology, Software, Writing, Editing, Visualization. **Yeni Rahkmawati:** Conceptualization, Reviewing, Validation, Supervision. **Selvi Annisa:** Methodology, Reviewing, Editing, Validation. **Anne Mudya Yolanda:** Reviewing, Validation, Editing.

References

[1] U.S. Department of Agriculture, "Rice - Rice Sector at a Glance," Economic Research Service. Accessed: Aug. 05, 2025. [Online]. Available: https://www.ers.usda.gov/topics/crops/rice/rice-sector-at-a-glance#Global

[2] Badan Pusat Statistik, *Ringkasan Eksekutif Pengeluaran dan Konsumsi Penduduk Indonesia Maret 2025*. Jakarta, 2025.

[3] A. Fauzi, B. Firmansyah, and D. P. Sari, "Analisis Volatilitas Harga Beras dan Segmentasi Wilayah Berbasis Data Digital Pemanfaatan data crawling dari Panel Harga Pangan Nasional," in *Seminar Nasional Official Statistics*, 2025, pp. 277–286. doi: 10.34123/semnasoffstat.v2025i1.2420.

[4] A. M. Ginting, "Strategi Kebijakan Mencegah Kenaikan Harga Beras di Indonesia," 2024.

[5] P. Subianto and G. R. Raka, "Visi, Misi dan Program Prabowo - Gibran," 2023.

[6] H. Santoso, L. Hakim, and H. Magdalena, "Sosialisasi Dampak Kenaikan Beras dengan Prediksi Kebutuhan Beras Masyarakat di Pasar Induk Cipinang dengan Kerjasama Badan Pangan Nasional," *J. Abdidas*, vol. 5, no. 2, pp. 90–96, 2024, doi: 10.31004/abdidas.v5i2.901.

[7] N. A. Muzakir and M. Z. Yahya, "Analisis Perbandingan Model Double Exponential Smoothing dan ARIMA untuk Prediksi Harga Beras di Indonesia," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 7, no. 1, pp. 7–20, 2025, doi: 10.35580/variansiunm349.

[8] V. Arinal and M. Azhari, "Penerapan Regresi Linear untuk Prediksi Harga Beras di Indonesia," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 341–346, 2023, doi: 10.55338/saintek.v5i1.1417.

[9] E. D. B. Tarigan, M. F. Balqis, T. A. Hutapea, and D. I. Sihombing, "Peramalan Harga Beras di Indonesia Dengan ARIMA," *Sepren*, vol. 5, no. 02, pp. 117–126, 2024, doi: 10.36655/sepren.v5i02.1508.

[10] D. A. Fajari, M. F. Abyantara, and H. A. Lingga, "Peramalan Rata-Rata Harga Beras Pada Tingkat Perdagangan Besar Atau Grosir Indonesia Dengan Metode Sarima (Seasonal Arima)," *J. Agribisnis Terpadu*, vol. 14, no. 1, p. 88, 2021, doi: 10.33512/jat.v14i1.11460.

[11] B. Agan, S. Celik, O. I. Damak, and B. Miba'am, *Evaluating the machine learning-based models for predicting carbon neutrality in Sub-Saharan African Nations*, no. 0123456789. Springer Netherlands, 2025. doi: 10.1007/s10668-025-06289-y.

[12] D. Saputra, D. R. Trinadi, and D. Agustina, "Perbandingan Metode Random Forest , Linier Regression , SVM Untuk Memprediksi Harga Beras Premium," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2025, pp. 154–160. doi: 10.47701/yqx4ss53.

[13] M. Andriyani, S. Nurwilda, D. Zatusiva Haq, and D. Candra Rini Novitasari, "Prediksi Harga Beras Premium Tahun 2024 Menggunakan Metode Gradient Boosted Trees Regression," *J.*

Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform., vol. 18, no. 2, pp. 75–84, 2024, doi: doi.org/10.47111/JTI.

[14] I. Sabilirrasyad, N. A. Prasetyo, and M. Hermansyad, "Modeling and Predicting Indonesia Rice Prices Using Hyperparameter Optimization XGBoost," in *International Conference On Economics , Business and Information Technology*, Institut Teknologi dan Sains Mandala, 2024, pp. 459–470. doi: 10.31967/prmandala.v5i0.1226.

[15] R. De Bin and V. G. Stikbakke, "A Boosting First-Hitting-Time Model for Survival Analysis in High-Dimensional Settings," *Lifetime Data Anal.*, vol. 29, no. 2, pp. 420–440, 2023, doi: 10.1007/s10985-022-09553-9.

[16] I. B. Mustapha *et al.*, "Comparative Analysis of Gradient - Boosting Ensembles for Estimation of Compressive Strength of Quaternary Blend Concrete," *Int. J. Concr. Struct. Mater.*, 2024, doi: 10.1186/s40069-023-00653-w.

[17] G. Ke *et al.*, "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," in *31st Conference on Neural Information Processing Systems*, California, 2017, pp. 1–9. doi: 10.5555/3294996.3295074.

[18] M. Kopyt, P. Piotrowski, and D. Baczyński, "Short-Term Energy Generation Forecasts at a Wind Farm—A Multi-Variant Comparison of the Effectiveness and Performance of Various Gradient-Boosted Decision Tree Models," *Energies*, vol. 17, no. 23, 2024, doi: 10.3390/en17236194.

[19] N. Zhang, Q. An, S. Zhang, and H. Ma, "Price Prediction for Fresh Agricultural Products Based on a Boosting Ensemble Algorithm," *Mathematics*, vol. 13, no. 1, 2025, doi: 10.3390/math13010071.

[20] Badan Pangan Nasional, "Panduan Implementasi Persyaratan Mutu Dan Label Beras," 2023. [Online]. Available: https://badanpangan.go.id/storage/app/media/Panduan Implementasi Persyaratan Mutu dan Label Beras_compressed.pdf

[21] M. Alam *et al.*, "Improving Rice Grain Quality Through Ecotype Breeding for Enhancing Food and Nutritional Security in Asia–Pacific Region," *Rice*, vol. 17, no. 1, 2024, doi: 10.1186/s12284-024-00725-9.

[22] H. F. Fiqa, A. R. Dewi, and R. Pandiya, "Perbandingan Metode ARIMA dan Prophet dalam Prediksi Harga Cabai Rawit di Provinsi Jawa Timur," *Pros. Semin. Nas. Sains Data*, vol. 4, no. 1, pp. 850–862, 2024, doi: 10.33005/senada.v4i1.350.

[23] M. S. Buntara, H. Napitupulu, and N. Gusriani, "Pemrograman Python Untuk Peramalan Data Deret Waktu Menggunakan Metode Seasonal Autoregressive Integrated Moving Average (Sarima)," *In Search*, vol. 22, no. 2, pp. 354–362, 2023, doi: 10.37278/insearch.v22i2.774.

[24] W. W. S. Wei, *Time Series Analysis Univariate and Multivariate Methods*. Boston: Pearson-Addison Wesley, 2006.

[25] S. Annisa, Y. Rahkmawati, and H. Hafid, "Peramalan Harga Minyak Mentah Indonesia Menggunakan Algoritma Random Forest," *J. Gaussian*, vol. 13, no. 2, pp. 472–478, 2025, doi: 10.14710/j.gauss.13.2.472-478.

[26] W. J. Niu, Z. K. Feng, S. S. Li, H. J. Wu, and J. Y. Wang, "Short-Term Electricity Load Time Series Prediction by Machine Learning Model via Feature Selection and Parameter Optimization Using Hybrid Cooperation Search Algorithm," *Environ. Res. Lett.*, vol. 16, no. 5, 2021, doi: 10.1088/1748-9326/abeeb1.

[27] M. M. Høgsgaard, K. G. Larsen, and M. E. Mathiasen, "The Many Faces of Optimal Weak-to-Strong Learning," in *38th Conference on Neural Information Processing Systems*, 2024, pp. 51885–51904. doi: 10.5555/3737916.3739560.

[28] G. Airlangga and A. Liu, "A Hybrid Gradient Boosting and Neural Network Model for Predicting Urban Happiness: Integrating Ensemble Learning with Deep Representation for Enhanced Accuracy," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 1, pp. 1–23, 2025, doi: 10.3390/make7010004.

[29] E. Q. Shehab, F. F. Taha, S. H. Muhodir, H. Imran, K. A. Ostrowski, and M. Piechaczek, "Gradient Boosting Regression Tree Optimized with Slime Mould Algorithm to Predict the Higher Heating Value of Municipal Solid Waste," *Energies*, vol. 17, no. 17, 2024, doi: 10.3390/en17174213.

[30] M. Saied, S. Guirguis, and M. Madbouly, "A Comparative Study of Using Boosting-Based Machine Learning Algorithms for IoT Network Intrusion Detection," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 177, 2023, doi: 10.1007/s44196-023-00355-x.

[31] M. Gan, S. Pan, Y. Chen, C. Cheng, H. Pan, and X. Zhu, "Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River," *J. Mar. Sci. Eng.*, vol. 9, no. 5, 2021, doi: 10.3390/jmse9050496.

[32] D. Zhao, Z. Hu, Y. Yang, and Q. Chen, "Energy Conservation for Indoor Attractions Based on

NRBO-LightGBM," *Sustain.*, vol. 14, no. 19, 2022, doi: 10.3390/su141911997.
[33] J. Zhao, J. He, J. Wan, and K. Liu, "Energy Consumption Prediction for Electric Buses Based on Traction Modeling and LightGBM," *World Electr. Veh. J.*, vol. 16, no. 3, p. 159, 2025, doi: 10.3390/wevj16030159.
[34] A. K. Karakutuk, O. Ozdemir, and S. Senturk, "Optuna-Optimized Pythagorean Fuzzy Deep Neural Network : A Novel Framework for Uncertainty-Aware Image Classification," *Appl. Sci.*, vol. 15, no. 20, p. 11097, 2025, doi: 10.3390/app152011097.
[35] L. Wang, C. Zhang, W. Wang, T. Deng, T. Ma, and P. Shuai, "Slope Stability Assessment Using an Optuna-TPE-Optimized CatBoost Model," *Eng*, vol. 6, no. 8, p. 185, 2025, doi: 10.3390/eng6080185.
[36] Q. Qin and L. Li, "A VMD-Based Four-Stage Hybrid Forecasting Model with Error Correction for Complex Coal Price Series," *Mathematics*, vol. 13, no. 18, p. 2912, 2025, doi: 10.3390/math13182912.
[37] M. Alharithi, E. M. Almetwally, O. Alotaibi, M. M. Eid, E. S. M. El-kenawy, and A. A. Elnazer, "A Comparative Study of Statistical and Intelligent Classification Models for Predicting Airlines Passenger Management Satisfaction," *Alexandria Eng. J.*, vol. 119, pp. 99–110, 2025, doi: 10.1016/j.aej.2025.01.109.
[38] X. Wen, M. Jaxa-Rozen, and E. Trutnevyte, "Accuracy Indicators for Evaluating Retrospective Performance of Energy System Models," *Appl. Energy*, vol. 325, p. 119906, 2022, doi: 10.1016/j.apenergy.2022.119906.
[39] A. Fayyazbakhsh, T. Kienberger, and J. Vopava-wrienz, "Comparative Analysis of Load Profile Forecasting : LSTM , SVR , and Ensemble Approaches for Singular and Cumulative Load Categories," *Smart Cities*, vol. 8, no. 2, p. 65, 2025, doi: 10.3390/smartcities8020065.
[40] Kementerian Pertanian, "Analisis Kinerja Perdagangan Beras," 2024, [Online]. Available: https://satudata.pertanian.go.id/assets/docs/publikasi/1A_Analisis_Kinerja_Perdagangan_Beras_2024_-_publish.pdf
[41] Institute for Development of Economics and Finance, "Dinamika Pangan dan Energi: Dari Gejolak Harga Beras hingga Perbaikan Tata Kelola Impor BBM," 2025.

DOI: https://doi.org/10.34123/jurnalasks.v17i2.856

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 2, December 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

TEMPLATE

Click Here, Type the Title of Your Paper, Capitalize First Letter of Each Word (Times New Romans (TNR), Size 17 pt, exactly spacing at 20 pt, 12 pt spacing for next heading, justify))

First Author Name[1*], Second Author Name[2] , Third Author Name[3] (TNR font, size 13 pt, exactly spacing at at 15 pt and 8 pt spacing for the next heading.)

[1]First Affiliation, City, Country, [2,3]Author Affiliation, City, Country [2]Second Affiliation, City, Country, [2,3]Author Affiliation, City, Country
**Corresponding Author: E-mail address: author@institute.xxx*
(TNR font, size 10 pt, with single spacing and 0 pt spacing for the next heading. And for the corresponding author use 10 pt Times New Roman font with single spacing and 8 pt spacing for the next heading.)

ARTICLE INFO

Article history: (TNR, 10pt)
Received dd month, yyyy
Revised dd month, yyyy
Accepted dd month, yyyy
Published dd month, yyyy

Keywords: (TNR, 10 pt)
Type your keywords here, separated by semicolon (;) Capitalize first letter of each word, times new roman, use 10 pt and write alphabetically in 5-10 words

Abstract (Times New Roman, size font 12)

Introduction/Main Objectives: Describe the topic your paper examines. Provide a background to your paper and why is this topic interesting. Avoid unnecessary content. **Background Problems:** State the problem or statistical applied/statistic computing phenomena studied in this paper and specify the research question(s) in one sentence. **Novelty:** Summarize the novelty of this paper. Briefly explain why no one else has adequately researched the question yet. **Research Methods:** Provide an outline of the research method(s) and data used in this paper. Explain how did you go about doing this research. Again, avoid unnecessary content and do not make any speculation(s). **Finding/Results:** List the empirical finding(s) and write a discussion in one or two sentences. Abstract written in English, with a length of 150 - 200 words. Use 10 pt Times New Roman font with justified alignment, single spacing, and 1 pt spacing for the next heading.

1. Main Text (bold, TNR, 14 pt, spacing before- after 12 pt, line spacing 12 pt)

These instructions give you guidelines for preparing papers for Jurnal Aplikasi Statistika & Komputasi Statistik which is published by Politeknik Statistika STIS, effective from the June 2024 edition. Starting from June 2024 Volume 16 No. 1, please use the template available at the following link https://s.stis.ac.id/TemplateJurnalASKS. The paragraphs continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers, bold and 14 pt. Here follow further instructions for authors.

The manuscript was created using Microsoft Office Word only and should be formatted for direct printing. As indicated in the template, manuscript should be prepared in single column format that suitable for direct printing onto paper with A4 paper size (21 x 29.7 cm). All parts of the manuscript are typed in Times New Roman font, size 11, line spacing exactly at 12 pt, with 0.2 line spacing for the next heading and margins of 3 cm of left and 2 cm for top, bottom, and right, the length of header from the top is 1.5 cm and the length of footer from the bottom is 1 cm. For the main text, use justify alignment and special indent for the first line in 0.76 cm. For the purpose of editing the manuscript, all parts of the manuscript (including tables, figures and mathematical equations) are made in a format that can be edited by the editor [1, 2].

The writing style for the Jurnal Aplikasi Statistika & Komputasi Statistik is written in English with a narrative style. Tracing is kept simple and as far as possible avoiding multilevel chronology.

1.1. Structure

Please make sure that you use as much as possible normal fonts in your documents. Special fonts, such as fonts used in the Far East (Japanese, Chinese, Korean, etc.) may cause problems during processing. To avoid unnecessary errors, you are strongly advised to use the 'spellchecker' function of MS Word. Follow this order when typing manuscripts: **Title, Authors, Affiliations, Abstract, Keywords, Main Text** (**Introduction, Material and Methods, Result and Discussion, Conclusion**, including figures and tables), **Acknowledgements, and References**.

Introduction coverage What is the purpose of the study? Why are you conducting the study? The main section of the article should start with an introductory section, which provides more details about the paper's purpose, motivation, research methods and findings. The introduction should be relatively nontechnical, yet clear enough for an informed reader to understand the manuscript's contribution. The Introduction is not an extended version of the abstract; never use the same sentences in both sections

The "introduction" in the manuscript is important to demonstrate the motives of the research. It analyzes the empirical, theoretical and methodological issues in order to contribute to the extant literature. This introduction will be linked with the following parts, most noticeably the literature review. Explaining the problem's formulation should cover the following points: (1) Problem recognition and its significance; (2) clear identification of the problem and the appropriate research questions; (3) coverage of problem's complexity; and (4) well-defined objectives.

The second part of the manuscript, "Method, Data, and Analysis" is designed to describe the nature of the data. The method should be well elaborated and enhance the model, the approach to the analysis and the step taken. Equations should be numbered as we illustrate.

This section typically has the following sub-sections: Sampling (a description of the target population, the research context, and units of analysis; the sample; and respondents' profiles); data collection; and measures (or alternatively, measurements).

The research methodology should cover the following points: Concise explanation of the research's methodology is prevalent; reasons for choosing the particular methods are well described; the research's design is accurate; the sample's design is appropriate; the data collection processes are properly conducted; the data analysis methods are relevant and state-of-the-art.

The second part of manuscript, "Result and Discussion" The author needs to report the results in sufficient detail so that the reader can see which statistical analysis was conducted and why, and later to justify their conclusions.

The "Discussion and Analysis" part, highlights the rationale behind the result answering the question "why the result is so?" It shows the theories and the evidence from the results. The part does not just explain the figures but also deals with this deep analysis to cope with the gap that it is trying to solve.

The "Conclusion and Suggestion", in this section, the author presents brief conclusions from the results of the research with suggestions for advanced researchers or general readers. A conclusion may cover the main points of the paper, but do not replicate the abstract in the conclusion. Authors should explain the empirical and theoretical benefits, and the existence of any new findings. The author may present any major flaws and limitations of the study, which could reduce the validity of the writing, thus raising questions from the readers (whether, or in what way), the limits in the study may have affected the results and conclusions. Limitations require a critical judgment and interpretation of the impact of their research. The author should provide the answer to the question: Is this a problem caused by an error, or in the method selected, or the validity, or something else?

The manuscript including the graphic contents and tables should be around 15-20 pages. The manuscript is written in English. The Standard English grammar must be observed. The title of the article should be brief and informative and it is recommended not to exceed 12 words. When writing numbers, use a period to separate decimal points and a comma to separate thousands.

The use of abbreviation is permitted, but the abbreviation must be written in full and complete when it is mentioned for the first time and it should be written between parentheses. Terms/foreign words or regional words should be written in italics. Notations should be brief and clear and written according to the standardized writing style. Symbols/signs should be clear and distinguishable, such as the use of number 1 and letter l (also number 0 and letter O).

Bulleted lists may be included and should look like this:

- First point
- Second point
- And so on

Ensure that you return to the 'body-text' style, the style that you will mainly be using for large blocks of text, when you have completed your bulleted list.

Please do not alter the formatting and style layouts which have been set up in this template document.

1.2. Tables

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables with left justified alignment. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Below is an example which the authors may find useful.

Table 1. Rice coefficient for various climatic conditions

Humidity	Wind Speed		
	Low	Medium	High
Dry	1.10	1.15	1.20
Medium	1.05	1.10	1.15
High	1.00	1.05	1.10

1.3. Construction of references

References must be listed at the end of the paper. Do not begin them on a new page unless this is absolutely necessary. Authors should ensure that every reference in the text appears in the list of references and vice versa. Indicate references by [1] or [2] or [3] in the text.

Some examples of how your references should be listed are given at the end of this template in the 'References' section, which will allow you to assemble your reference list according to the correct format and font size. The paper must include a reference list containing only the quoted work and using the Mendeley tool. Each entry should contain all the data needed for unambiguous identification. With the author-date system, use the following format recommended by IEEE Citation Style. The first line of each citation is left adjusted. Every subsequent line is indented 5-7 spaces. The references are arranged in alphabetical order, written in 11pt Times New Roman font with 0 pt spacing for the next heading.

The references shall contain at least 20 (twenty) references. For whole references, at least 16 references or 80% of them must be refer to primary sources (scientific journals, conference proceedings, research reference books) which are published within 5 (five) year. The IEEE citation guide can be access here: https://ieee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf

1.4. Section headings

Section headings should be left justified, bold, with the first letter capitalized and numbered consecutively, starting with the Introduction. Section headings use 14 pt Times New Roman and exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment and special hanging indentation at 0.63 cm. Sub-section headings should be in capital and lower-case italic letters, numbered 1.1, 1.2, etc, exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment with 0.12 cm left indentation and special hanging indentation at 0.63 cm, with second and subsequent lines indented. All headings should have a minimum of three text lines after them before a page or column break. Ensure the text area is not blank except for the last page. Both section heading and sub-section headings are in dark blue color with the code #034F84 (R: 3 G: 79 B: 132).

1.5. General guidelines for the preparation of your text

Avoid hyphenation at the end of a line. Symbols denoting vectors and matrices should be indicated in bold type. Scalar variable names should normally be expressed using italics. Weights and measures should be expressed in SI units. All non-standard abbreviations or symbols must be defined when first mentioned, or a glossary provided.

1.6. Footnotes

Footnotes should be avoided if possible.

2. Illustrations

All figures should be numbered with Arabic numerals (1,2,3,….). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Preferred format of figures are PNG, JPEG, GIF etc. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper. Please ensure that all the figures are of 300 DPI resolutions as this will facilitate good output. Figures should be embedded and not supplied separately.

The figure number and caption should be typed below the illustration in 11 pt and left justified [***Note:*** one-line captions of length less than column width (or full typesetting width or oblong) centered].

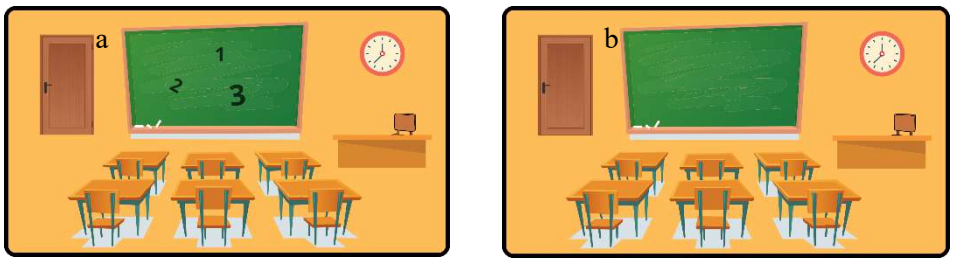


Figure 1. (a) first picture; (b) second picture

3. Equations

Equations and formulae should be typed in MathType or Microsoft Equation, and numbered consecutively with Arabic numerals in parentheses on the right hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space.

$$\rho = \frac{\vec{E}}{J_C\left(T = \text{const.}\right)\cdot\left(P\cdot\left(\frac{\vec{E}}{E_C}\right)^m + \left(1-P\right)\right)} \tag{1}$$

Ethics approval

The Ethical approval statement should be provided including the consent. If not appropriate, authors should state: “Not required.”

Acknowledgments

This section contains a form of thanks to individuals or institutions who have provided assistance in carrying out research, preparing the article, providing language help, writing assistance or proof reading the article and others.

Competing interests

A competing interest statement should be provided, even if the authors have no competing interests to declare. If no conflict exists, authors should state: “All the authors declare that there are no conflicts of interest.”

Funding

List funding sources in this standard way to facilitate compliance to funder's requirements. It is not necessary to include detailed descriptions on the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding. If no funding has been provided for the research, please include the following sentence: “This study received no external funding.”

Underlying data

This can be written as: “Derived data supporting the findings of this study are available from the corresponding author on request.”

Credit Authorship

Please outline the contributions of each co-author, using the following categories: conceptualization, methodology, software, validation, formal analysis, investigation, data collection, data curation, writing—original draft, writing—review & editing, visualization, supervision, and funding acquisition.

References

[1] W.K. Chen, *Linear Networks and Systems*. Belmont, CA: Wadsworth Press, 2003.
[2] R. Hayes, G. Pisano, and S. Wheelwright, *Operations, Strategy, and Technical Knowledge*. Hoboken, NJ: Wiley, 2007.
[3] K. A. Nelson, R. J. Davis, D. R. Lutz, and W. Smith, “Optical generation of tunable ultrasonic waves,” *J Appl Phys*, vol. 53, no. 2, pp. 1144–1149, Feb. 2002.