EISSN : 2654-7511
PISSN : 2089-0028

J STATISTIKA

JURNAL ILMIAH TEORI DAN APLIKASI STATISTIKA

Volume 18 | Nomor 1 | 2025

EDITORIAL TEAM

INTRODUCTION

We are delighted to announce the current publication of Volume 18, Number 1 of JStatistika, affiliated with the Statistics Department at PGRI Adi Buana University Surabaya, has been released in July 2025. This particular issue of the JStatistika Scientific Journal features a diverse array of articles addressing a wide spectrum of topics. One of the highlighted articles delves into "Application of Dummy Regression to Estimate the Income of the Working Population in East Lombok; The Modeling of The Poverty Rate In Indonesia From 2018 to 2023 Using A Panel Data Regression Approach; Forecasting Stock Prices Using a Nonlinear Approach with the Exponential Smooth Transition Autoregressive (ESTAR) Model; Geographically Weighted Negative Binomial Regression (GWNBR) Modeling In Infant Mortality Rate Cases In South Sulawesi; Cluster Analysis Using the Ward Algorithm for Grouping Regency / City in Central Java Province Based on Poverty Indicators 2023; Implementation of Geographically and Temporally Weighted Regression with Cross Validation and Generalized Cross Validation Methods for Deforestation Modeling in Kalimantan; Comparison of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) Methods in Determining Factors Affecting Tuberculosis Cases in Indonesia; Exploring Association of Household Conditions and Community Behavior in Flood Events in Banjarbaru Using Apriori Method; Implementation of K-Means Cluster for Districts or Cities in West Java Province Based on Unemployment Indicators; Bonus-Malus Premium for Third Party Liability Insurance with Poisson-Lindley Distribution Claim Frequency and Exponential-Inverse Gamma Distribution Claim Severity; Evaluating Patient Satisfaction in Surabaya Public Health Centers Using an Integrated IPA–Kano Framework; Application of ARIMAX-LSTM Model in Forecasting the Price of Broiler Chicken in Central Java; Rehabilitation and Law Enforcement as Optimal Controls in a Mathematical Model of Social Behavior; Rejecting Reduction: Clarifying the Concept of Deep Learning in Mathematics Teaching in the Era of Artificial Intelligence; Classifying Disadvantaged Districts/Cities in Indonesia: A Support Vector Machine Approach; Logistic Regression for Sentiment Analysis of Insecurity Phenomena on Platform X"

The JStatistika Scientific Journal enthusiastically welcomes and invites contributions in a diverse range of formats, including but not limited to scholarly scientific articles that encompass various facets of statistical science. We eagerly seek research findings, comprehensive reports, insightful case studies, thorough literature reviews, and updates that pertain to the dynamic landscape of statistical science. Our overarching objective is to cultivate a repository of knowledge that is not only current but also invaluable in tackling the ever-evolving and intricate challenges confronting our field. We actively encourage authors to submit their work if it resonates with the most recent advancements and frontiers in statistical science. Our aspiration is to foster an environment where these contributions can flourish, ultimately serving as a

wellspring of cutting-edge insights and understanding. We believe that these insights are instrumental in addressing the multifaceted issues that confront us in today's complex world.

Our editorial team extends a warm and inclusive invitation to scientists and scholars from diverse backgrounds and affiliations, including institutions of higher learning and esteemed research organizations. We seek your valuable contributions, whether they be grounded in empirical research results or rooted in rigorous scholarly studies within the expansive domain of statistics and its myriad practical applications. We hold a deep appreciation for the feedback and perspectives of our esteemed readership. Your input not only enriches the discourse but also plays a pivotal role in our continuous efforts to elevate the quality and relevance of the journal. We earnestly value your insights and ideas, recognizing that they are integral to our ongoing pursuit of excellence. Our ultimate vision is for the articles featured in the JStatistika Scientific Journal to transcend the confines of academia and serve as a wellspring of knowledge that benefits not only scholars and researchers but also professionals actively engaged in the diverse realms of statistical science and its multifaceted real-world applications. Through collaborative efforts and a shared commitment to advancing our understanding of statistics, we aim to make a meaningful impact in the broader scientific community and beyond.

Jstatistika has been indexed by Sinta 4 Kemendikbud, Garuda, Google Scholar, Crossref, Worldcat, Scilit, ROAD, Onesearch, Journal Stories, Dimensions, Base, Open Alex, Wikidata, Internet Archieve, Root Indexing, Core, Harvard Library, Universiteit Leiden Library, Semantic Scholar, Open Air Explore, ASCI, Cite Factor, University of Saskatchewan Library, The University of Queensland Library, George University Library and Boston University Library.

Surabaya, July 2025

Editor in Chief

LIST OF CONTENTS

COVER
EDITORIAL TEAM
INTRODUCTION
LIST OF CONTENTS

- **Application of Dummy Regression to Estimate the Income of the Working Population in East Lombok**
  Umam Hidayaturrohman, Basirun, Dita Septiana Ayundasari, and Muh. Zulkarnain Alayyubi
  Universitas Hamzanwadi 789 - 797

- **The Modeling of The Poverty Rate In Indonesia From 2018 to 2023 Using A Panel Data Regression Approach**
  Dhyana Venosia, Toha Saifudin, and Nur Chamidah
  Airlangga University 798 - 807

- **Forecasting Stock Prices Using a Nonlinear Approach with the Exponential Smooth Transition Autoregressive (ESTAR) Model**
  Allan Ruhui Fatmah Sari, Dwi Arman Prasetya, and Trimono
  Universitas Pembangunan Nasional Veteran Jawa Timur 808 - 818

- **Geographically Weighted Negative Binomial Regression (GWNBR) Modeling In Infant Mortality Rate Cases In South Sulawesi**
  Ilmah Nurul Fajril, Bobby Poerwanto, and Hafid Hardianti
  Universitas Negeri Makassar 819 – 829

- **Cluster Analysis Using the Ward Algorithm for Grouping Regency / City in Central Java Province Based on Poverty Indicators 2023**
  Agung Supriyono and Atika Nurani Ambarwati
  Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang 830 – 839

- **Implementation of Geographically and Temporally Weighted Regression with Cross Validation and Generalized Cross Validation Methods for Deforestation Modeling in Kalimantan**
  Gema Khusnul Ma'rifah, Mohammad Idhom, and Trimono
  Universitas Pembangunan Nasional Veteran Jawa Timur 840 - 850

- **Comparison of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) Methods in Determining Factors Affecting Tuberculosis Cases in Indonesia**
  Muhammad Arib Alwansyah, Awaliyatul Uswah Awaliyatul, and Yulian Fauzi
  Universitas Bengkulu 851 - 865

- **Exploring Association of Household Conditions and Community Behavior in Flood Events in Banjarbaru Using Apriori Method**
  Rifqi Aulya Rahman, Yuana Sukmawaty, and Selvi Annisa
  Lambung Mangkurat University 866 - 876

- **Implementation of K-Means Cluster for Districts or Cities in West Java Province Based on Unemployment Indicators**
  Alifa Nur Oktaviani and Atika Nurani Ambarwati
  Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang 877 - 886

- **Bonus-Malus Premium for Third Party Liability Insurance with Poisson-Lindley Distribution Claim Frequency and Exponential-Inverse Gamma Distribution Claim Severity**
  Bilqis Nur Rizkia, and Aceng Komarudin Mutaqin
  Universitas Islam Bandung 887 - 896

- **Evaluating Patient Satisfaction in Surabaya Public Health Centers Using an Integrated IPA–Kano Framework**
  Salman Alfarizi Pradana Andikaputra, Sumartono, and Nuril Huda
  Universitas Dr. Soetomo Surabaya 897 – 908

- **Application of ARIMAX-LSTM Model in Forecasting the Price of Broiler Chicken in Central Java**
  Divayanti Febri Sakina, Trimono, Amri Muhaimin
  Universitas Pembangunan Nasional Veteran Jawa Timur 909 – 919

- **Rehabilitation and Law Enforcement as Optimal Controls in a Mathematical Model of Social Behavior**
  Nailul Izzati and Wahyuni Ningsih
  Universitas Hasyim Asy'ari and Politeknik Negeri Malang 920 - 929

- **Rejecting Reduction: Clarifying the Concept of Deep Learning in Mathematics Teaching in the Era of Artificial Intelligence**
  Anis Munfarikhatin, and Irmawaty Natsir
  Universitas Musamus 930 - 936

- **Classifying Disadvantaged Districts/Cities in Indonesia: A Support Vector Machine Approach**
  Wanda Surianto, Lia Mauliani, Ridha Ferdhiana, and Nurhasanah
  Universitas Syiah Kuala 937 - 947

- **Logistic Regression for Sentiment Analysis of Insecurity Phenomena on Platform X**
  Emeylia Safitri, Wara Alfa Syukrilla, and Ika Nur Laily Fitriana
  Universitas Terbuka and UIN Syarif Hidayatullah Jakarta 948 - 956

Application of Dummy Regression to Estimate the Income of the Working Population in East Lombok

Umam Hidayaturrahman(1), Basirun(2), Dita Septiana Ayundasari (3), Muh. Zulkarnain Alayyubi (4)

Program Studi Statistika, Fakultas MIPA, Universitas Hamzanwadi

Jl. TGKH M. Zainuddin Abdul Madjid No. 132 Pancor (83611) Selong-Lombok Timur-NTB
e-mail: hidayaturrohman0809@gmail.com(1) basirun@hamzanwadi.ac.id(2)
ertinasa.210304007@student.hamzanwadi.ac.id(3) efidaa80@gmail.com(4)

ABSTRAK

Penelitian ini dilaksanakan di Badan Pusat Statistik Kabupaten Lombok Timur pada periode 7 Oktober hingga 20 November 2024. Tujuan utama dari penelitian ini adalah untuk mengevaluasi berbagai faktor yang berkontribusi terhadap tingkat pendapatan penduduk di wilayah tersebut. Data yang dianalisis berasal dari data sekunder, yakni hasil Survei Angkatan Kerja Nasional (Sakernas). Fokus penelitian ini adalah untuk menilai sejauh mana variabel-variabel seperti jenis kelamin, usia, status pernikahan, tingkat pendidikan terakhir, total jam kerja per minggu, status dalam pekerjaan, klasifikasi bidang kerja, serta jenis pekerjaan memengaruhi penghasilan individu yang bekerja di Lombok Timur. Dari hasil analisis, diperoleh sebuah model regresi yang mampu memprediksi pendapatan masyarakat berdasarkan faktor-faktor tersebut. Model ini memiliki koefisien determinasi sebesar 0,747 atau 74,7%, yang mengindikasikan adanya hubungan yang cukup kuat antara variabel-variabel bebas dan pendapatan sebagai variabel terikat. Selain itu, ditemukan bahwa 52,1% variasi dalam tingkat pendapatan dapat dijelaskan oleh variabel-variabel dalam model, sementara sisanya sebesar 47,9% dipengaruhi oleh faktor lain yang tidak tercakup dalam penelitian ini. Temuan dari penelitian ini memberikan pemahaman lebih mendalam mengenai faktor-faktor penentu pendapatan di Lombok Timur, dan hasilnya berpotensi dijadikan dasar dalam perumusan kebijakan untuk mendorong peningkatan kesejahteraan serta menurunkan tingkat kemiskinan di wilayah tersebut.
Kata kunci: Sakernas, Model Regresi, Pendapatan, Kemiskinan

ABSTRACT

This research was carried out at the Central Bureau of Statistics in East Lombok between October 7 and November 20, 2024, aiming to explore the determinants of individual income within the region. The study utilized secondary data, specifically drawn from the National Labor Force Survey (Sakernas). It investigates how variables such as sex, age, marital status, level of education, working hours, job status, industry sector, and job type contribute to the income levels of workers in East Lombok. The analysis produced a regression model capable of estimating income based on the aforementioned factors. The model achieved a coefficient of determination of 0.747, suggesting a moderately strong relationship between the predictor variables and income as the outcome variable. Furthermore, the results indicate that 52.1% of income variation is explained by the variables included in the model, while the remaining 47.9% is attributable to external influences not captured in this study. Overall, this study offers valuable insight into the key factors shaping income in East Lombok and may serve as a useful reference for policymakers aiming to enhance community welfare and address poverty reduction in the area.
Keywords: *Sakernas, Regression Model, Income, Unemployment*

INTRODUCTION

Indonesia, as a vast archipelagic nation endowed with rich natural resources and a substantial population, faces ongoing social challenges. As recorded in the 2020 Population Census, the population reached 270.20 million. This continual demographic growth is intrinsically tied to a range of socio-economic problems, particularly poverty. Poverty refers to a condition in which individuals are unable to fulfill essential needs—such as food, clothing, and housing—due to a combination of internal and external contributing factors [1]

A key driver of poverty is the high rate of unemployment. Data from the Central Statistics Agency (BPS) indicates that 7.39 million people were unemployed out of a total labor force of 118.19 million. This disparity between workforce expansion and limited job availability intensifies socio-economic inequality. The issue is further compounded by the presence of approximately 360,000 unemployed university graduates [2].

Among the provinces experiencing persistent poverty is West Nusa Tenggara (NTB), despite its wealth of natural assets. In particular, East Lombok District continues to report a high poverty incidence. According to BPS East Lombok data from 2013, around 20.7% of the district's residents were living below the poverty line. A key contributor to this is the relatively low regional minimum wage (UMR) compared to other provinces, alongside factors such as gender, age, educational attainment, employment type, and economic sector, all of which significantly affect income levels in the area.

To examine the determinants of income, this study employs the dummy regression technique, which enables the incorporation of qualitative variables into quantitative analysis. Dummy variables act as proxies for categorical factors assumed to influence continuous outcomes—in this case, individual income [3]. The strength of this approach lies in its enhanced predictive accuracy and its ability to support more informed decision-making compared to conventional multiple regression models.

In light of these considerations, this research aims to thoroughly investigate the factors affecting income among East Lombok residents. By applying dummy regression, the study will identify key qualitative attributes—such as gender, age, educational background, employment status, and industry classification—that exert a significant impact on earnings. The findings are intended to inform policymakers in designing targeted interventions to alleviate poverty in the region

METHODS

Regression Analysis

Regression analysis is a widely utilized statistical method for examining the association between a dependent variable and one or more explanatory variables [4]. As noted by [5], this technique provides a mathematical framework for detecting and interpreting patterns in variable relationships, with the principal aim of forecasting or estimating the outcome variable based on known values of the predictors. Through this approach, researchers are able to assess the strength and structure of the relationship between variables of interest.

Essentially, regression analysis serves as a tool for modeling how one or more independent variables relate to a dependent variable. It facilitates insights into potential causal links among the

variables. Nevertheless, the presence of a statistical relationship does not inherently imply a causal effect, and further investigation is needed to substantiate causality claims [6]. Muhyi et al. [7] suggest that a model represents a simplified abstraction of a complex real-world system. Given the intricacy of natural phenomena and the limitations of human perception and analytical tools, modeling enables the reduction of complexity into manageable forms—often based on past knowledge or assumptions about inter-variable relationships within a system.

The general form of a simple linear regression model, involving only one independent variable, is expressed as follows [8]:

$$Y_i = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

In this case :

Y_i : response variable
β_0 : arameter, which is the y-intercept
β_1 : arameter, which is the slope of the line
ε : random error component
x : independent variable (predictor of y)

To ensure accurate and reliable estimation of regression parameters, several key assumptions must be satisfied in simple regression analysis [9]. These include: linearity of the relationship, absence of autocorrelation, homoscedasticity (constant variance of errors), normal distribution of residuals, and no multicollinearity among variables.

The Kolmogorov-Smirnov (K-S) test is a non-parametric tool used to compare the cumulative distributions of two independent datasets, in order to assess whether they originate from the same underlying distribution. It also serves to evaluate the conformity of a sample distribution to a specified theoretical distribution, such as the normal distribution. One advantage of the K-S test lies in its simplicity and objectivity, especially when compared to graphical normality assessments [10]. This test works by identifying the maximum difference between the empirical cumulative distribution function of the observed data and the cumulative distribution function of the reference distribution. This difference, referred to as the Maximum Deviation (D), is evaluated in relation to the sample size (N). By comparing the computed D value to critical values in the K-S distribution table, researchers can determine whether the deviation is statistically significant [11].

Dummy Variabel

In the context of multiple regression analysis, independent variables are generally numerical in nature. However, in practical research settings, not all predictors are quantitative. Researchers frequently encounter explanatory variables that are categorical. To address this, a technique known as dummy regression—utilizing dummy variables—has been developed. These variables are employed to convert qualitative attributes (such as gender, ethnicity, religion, policy interventions, or contextual variations) into a numerical format. Essentially, dummy variables are used to represent categorical data that are believed to influence a continuous outcome variable [3].

A crucial aspect of regression involving categorical predictors is the coding system used for these variables. In such cases, mutually exclusive classification is fundamental—each

observation must belong to only one category. For instance, an individual cannot simultaneously be classified as both male and female, nor can someone be coded as both a civil servant and an entrepreneur, even if they engage in both roles. This exclusivity underpins the structure of dummy coding.

Using binary coding (0 and 1), dummy variables are always dichotomous. A value of 1 indicates that the respondent belongs to a specific category, while a value of 0 indicates otherwise. This approach ensures that for each category, an individual is either classified as included (1) or excluded (0). The binary coding system can be conceptually likened to an on/off switch: a code of 1 turns a category "on" (indicating the presence of a trait or group membership), whereas a code of 0 switches it "off" (indicating its absence).

Dummy variables are also known as binary, proxy, dichotomous, or categorical variables. By design, a dummy variable assumes the value D = 1 for one group and D = 0 for the comparison group. Several general model forms involving dummy variables are as follows:

I. $Y = a + bX + c\ D1$ (**Intercept Dummy Model**) (2)

II. $Y = a + bX + c\ (D1X)$ (**Slope Dummy Model)** (3)

III. $Y = a + bX + c\ (D1X) + d\ D1$ (**Combination**)` (4)

Each model reflects a different way dummy variables can modify either the intercept, the slope, or both, depending on category membership. These models enhance the flexibility of regression analysis when incorporating qualitative information into quantitative frameworks.

RESULT AND DISCUSSION

Descriptive Statistic

Table 1. Summary of Descriptive Statistics of Wage Variables in East Lombok

Wage	Range	Minimum	Maximum	Mean
Lombok Timur	4.380.000	120.000	4.500.000	1.084.252

Based on the data presented in the output above, the maximum monthly income of the people in East Lombok reaches Rp 4,500,000.00, while the minimum recorded income is Rp 120,000.00 per month. The average income of the East Lombok community is Rp 1,084,252.00 per month, which is lower than the Regional Minimum Wage (UMR) of West Nusa Tenggara Province, which stands at Rp 1,120,000.00 per month. Furthermore, when compared to the average income of Indonesians in general, which is Rp 2,236,045.00 per month, the income of the East Lombok community is also below the national average. This indicates an economic disparity in the region.

Dummy Regression Analysis for Estimating Wage Variables

Before conducting the dummy regression analysis, the first step was to assign levels to the categories of each independent variable with categorical data.

1. Gender

The dummy coding for the Gender variable was as follows:

Table 2. Dummy coding for Gender

Gender	DJK1
Male (1)	0

Female (2)	1

In each gender variable presented in the table above, there are two levels, consisting of one dummy variable (DJK1) and one level designated as the reference category (male). Each level is assigned a code (0,1), as explained below:

$$\text{DJK1} = \begin{cases} 1, \text{if the level was Female} \\ 0, \text{if the level was not Female} \end{cases} \quad (5)$$

This means that when the category is "Female," the dummy variable (DJK1) is coded as 1, and when the category is not "Female" (i.e., Male), it is coded as 0.

2. Marital Status

Below is the dummy variable for marital status. The dummy coding for the marital status variable will be displayed in the table below:

Tabel 3. Dummy Coding for Marital Status

Marital Status	**Dstat_Kawin1**	**Dstat_kawin2**	**Dstat_kawin3**
Unmarried (1)	0	0	0
Married (2)	1	0	0
Divorced (3)	0	1	0
Widowed (4)	0	0	1

For marital status, as shown in the table above, there were four categories: unmarried, married, divorced, and widowed. Since there were four categories, three dummy variables were created: Dstat_kawin1, Dstat_kawin2, and Dstat_kawin3, with one category serving as the reference category, which in this case was unmarried. Each category was assigned a code (0,1) as follows:

$$\text{Dstat\_kawin1} = \begin{cases} 1, \text{if the level was Married} \\ 0, \text{if the level was not Married} \end{cases} \quad (6)$$

In dummy Dstat_kawin1, the code 1 was assigned to all samples that were married, and code 0 was assigned to all samples that were not married.

$$\text{Dstat\_kawin2} = \begin{cases} 1, \text{if the level was Divorced} \\ 0, \text{if the level was not Divorced} \end{cases} \quad (7)$$

In dummy Dstat_kawin2, the code 1 was assigned to all samples that were divorced, and code 0 was assigned to all samples that were not divorced.

$$\text{Dstatus\_kawin3} = \begin{cases} 1, \text{if the level was Widowed} \\ 0, \text{if the level was not Widowed} \end{cases} \quad (8)$$

In dummy Dstat_kawin3, the code 1 was assigned to all samples that were widowed, and code 0 was assigned to all samples that were not widowed.

Additionally, for other variables such as Higher Education, Employment Status, Job Classification, and Type of Occupation, the same coding process was applied.

Normality Test

One-Sample Kolmogorov-Smirnov Test

		Unstandardized Residual
N		105
Normal Parameters[a]	Mean	.0000000
	Std. Deviation	2.93021631E2
Most Extreme Differences	Absolute	.056
	Positive	.056
	Negative	-.039
Kolmogorov-Smirnov Z		.572
Asymp. Sig. (2-tailed)		.899

a. Test distribution is Normal.

Figure 1. Kolmogorov-Smirnov normality test results for the data distribution

Referring to Figure 1, the Kolmogorov-Smirnov test for normality produced a p-value of 0.899, which exceeds the significance threshold of 0.05. Consequently, it can be inferred that the assumption of normality is satisfied, indicating that the data are normally distributed.

Heteroscedasticity Test

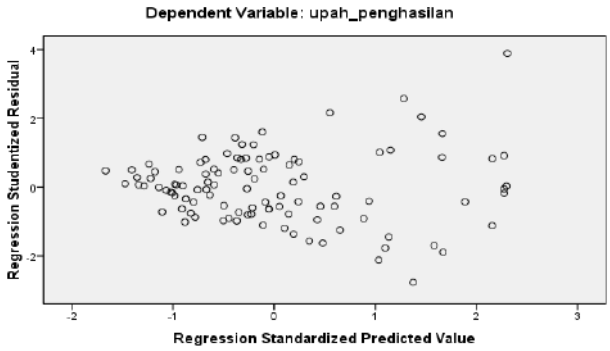


Figure 2. Results of the heteroscedasticity testn the regression residuals

Based on the scatter plot presented above, it can be observed that the data points are randomly and evenly dispersed around the zero line on the Y-axis, both above and below it. This random distribution of points suggests the absence of any discernible pattern or systematic variance in the residuals, indicating that heteroscedasticity is not present in the regression model. Consequently, this supports the validity of the model for prediction purposes, as the constant variance assumption is met.

Autocorelation Test

A robust regression model should be free from autocorrelation among residuals, meaning that the residuals must be independent of each other. One common method to detect the presence or absence of autocorrelation is the Run test, a non-parametric statistical procedure. The Run test evaluates whether the sequence of residuals exhibits randomness or shows systematic patterns indicating correlation. When residuals are uncorrelated, they are considered random, satisfying a

key assumption of regression analysis. The Run test helps to statistically verify this randomness by analyzing the sequence of residuals for runs, or consecutive observations of similar signs. Below are the results of the Run test applied to the residual data, which provide insight into the presence or absence of autocorrelation in the model [9].

Runs Test

	Unstandardized Residual
Test Value[a]	-3410.28425
Cases < Test Value	52
Cases >= Test Value	53
Total Cases	105
Number of Runs	60
Z	1.276
Asymp. Sig. (2-tailed)	.202

a. Median

Figure 3. Results of the runs test for autocorrelation detection

Based on the results presented in the table above, the runs test produced a significance value of 0.202, which is higher than the conventional threshold of 0.05. This indicates that there is insufficient evidence to reject the null hypothesis of randomness in the residuals. Consequently, it can be concluded that no autocorrelation exists within the residual data, implying that the residuals are independent and the regression model meets this important assumption

Multicollinearity Test

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1.602E6	530401.500		3.021	.003		
	Umur	19356.508	8672.345	.156	2.232	.028	.939	1.065
	Jam_kerja	26910.320	7750.981	.275	3.472	.001	.732	1.366
	DKBJI1	-1.610E6	588703.764	-.203	-2.734	.007	.835	1.197
	DKBJI2	-1.179E6	398498.223	-.250	-2.957	.004	.646	1.549
	DKBJI3	-2.468E6	347564.729	-.689	-7.102	.000	.490	2.040
	DKBJI4	-2.506E6	528842.399	-.363	-4.738	.000	.784	1.276
	DKBJI5	-2.593E6	304500.724	-.932	-8.516	.000	.385	2.600
	DKBJI6	-2.607E6	322211.982	-.817	-8.092	.000	.452	2.210

a. Dependent Variable: Upah

Figure 4. Results of the multicollinearity test based on tolerance and variance inflation factor (VIF) values

The output from the SPSS Coefficients table provides key indicators to detect multicollinearity among the independent variables, specifically through the Tolerance and

Variance Inflation Factor (VIF) statistics. According to established criteria, a Tolerance value below 0.10 suggests the presence of multicollinearity, while a VIF value exceeding 10 similarly indicates multicollinearity concerns [9]. Reviewing the results presented above, it is evident that all independent variables have Tolerance values well above the 0.10 threshold and none of the VIF values surpass the critical value of 10. These findings confirm that multicollinearity is not present in the regression model, ensuring that the estimates of the independent variables are reliable and not distorted by intercorrelations.

Multiple Regression Analysis

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.747[a]	.557	.521	918500.630

a. Predictors: (Constant), DKBJI6, DKBJI1, DKBJI4, DKBJI2, Jam_kerja, Umur, DKBJI3, DKBJI5

Figure 5. Results of the multiple regression analysis

Based on the output presented above, the regression model demonstrated an accuracy reflected by an R-value of 0.747, or 74.7%. This value represents the strength of the association between the independent variables and the dependent variable. Additionally, the Adjusted R Square value was reported as 0.521, indicating that the independent variables included in the model collectively explain 52.1% of the variance in the dependent variable. The remaining 47.9% of the variation is attributed to other factors not captured within this regression model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.602E6	530401.500		3.021	.003
	Umur	19356.508	8672.345	.156	2.232	.028
	Jam_kerja	26910.320	7750.981	.275	3.472	.001
	DKBJI1	-1.610E6	588703.764	-.203	-2.734	.007
	DKBJI2	-1.179E6	398498.223	-.250	-2.957	.004
	DKBJI3	-2.468E6	347564.729	-.689	-7.102	.000
	DKBJI4	-2.506E6	528842.399	-.363	-4.738	.000
	DKBJI5	-2.593E6	304500.724	-.932	-8.516	.000
	DKBJI6	-2.607E6	322211.982	-.817	-8.092	.000

a. Dependent Variable: Upah

Figure 6. Results of the dummy regression model

Based on the output above, it was obtained that the formed dummy regression model was:

Y = 1.602.488 + 19.356,51X1 + 26.910,32X2 - 1.609.782DKBJI1 - 1.178.512DKBJI2 – 2.468.488DKBJI3 – 2.505.716DKBJI4 – 2.593.132 DKBJI5 – 2.607.300DKBJI6

CONCLUSION

Based on the results of the analysis, West Nusa Tenggara Province ranks as the province with the second-lowest average minimum wage in Indonesia. In East Lombok specifically, the average monthly income of residents is approximately Rp 1,084,252.00, which falls below the

provincial Regional Minimum Wage (UMR) of Rp 1,120,000.00 per month. The study identified several key variables influencing the income of employed individuals in East Lombok, notably age, working hours, and job classification. The constructed dummy regression model revealed a statistically significant relationship between these independent variables and income, as expressed in the following regression equation: $Y = 1{,}602{,}488 + 19{,}356.51X_1 + 26{,}910.32X_2 - 1{,}609{,}782DKBJI_1 - 1{,}178{,}512DKBJI_2 - 2{,}468{,}488DKBJI_3 - 2{,}505{,}716DKBJI_4 - 2{,}593{,}132DKBJI_5 - 2{,}607{,}300DKBJI_6$. The coefficient of determination (R^2) of 0.747 (74.7%) confirms that the model has a strong explanatory power, indicating a substantial relationship between the independent variables and income. Additionally, the analysis shows that 52.1% of the income variation is accounted for by the variables included in the model, while 47.9% of the variation is likely influenced by other external factors not examined within the scope of this study.

REFRENCES

[1] Binti, M. T. (2017). Analisa Pengaruh Pertumbuhan Ekonomi Terhadap Penurunan Tingkat Kemiskinan di Kalimantan Tengah. *AlKALAM: JURNAL KOMUNIKASI, BISNIS DAN MANAJEMEN*, *3*(2), 69-78.

[2] Badan Pusat Statistik.____.*Survei Angkatan Kerja Nasional*. http://sirusa.bps.go.id/index.php?r=istilah/view&id=699. Diunduh Tanggal 10 November 2024 pukul 9.44

[3] Syahbania, M. 2017. *Variabel Dummy*. http://ethasyahbania.blogspot.com /2011/01/variabel-dummy.html. Diunduh Tanggal 10 Oktober 2024 pukul 19.21.

[4] Basri, H. (2019). Pemodelan Regresi Berganda Untuk Data Dalam Studi Kecerdasan Emosional. *DIDAKTIKA: Jurnal Kependidikan*, *12*(2), 103-116.

[5] Padilah, T. N., & Adam, R. I. (2019). Analisis regresi linier berganda dalam estimasi produktivitas tanaman padi di Kabupaten Karawang. *FIBONACCI: Jurnal Pendidikan Matematika Dan Matematika*, *5*(2), 117-128.

[6] Widodo, E. (2016). Analisis Arah Kausalitas (Causal Ordering). *Journal of Indonesian Economy and Business (JIEB)*, *31*(1), 75-82.

[7] Muhyi, M., et al. (2018). *Qualitative Research Methods, Quantitative Research Methods, and Mixed Methods in Education*.

[8] Padilah, D., & Wulandari, S. (2018). *Multiple Linear Regression Analysis for Estimating Rice Crop Productivity in Karawang Regency*.

[9] Nurdin, I., Sugiman, S., & Sunarmi, S. (2018). Penerapan Kombinasi Metode Ridge Regression (RR) dan Metode Generalized Least Square (GLS) untuk Mengatasi Masalah Multikolinearitas dan Autokorelasi. *Indonesian Journal of Mathematics and Natural Sciences*, *41*(1), 58-68.

[10] Quraisy, A. (2020). Normalitas data menggunakan uji kolmogorov-smirnov dan saphiro-wilk: studi kasus penghasilan orang tua mahasiswa prodi pendidikan matematika unismuh makassar. *J-HEST Journal of Health Education Economics Science and Technology*, *3*(1), 7-

[11] Ahadi, G. D., & Zain, N. N. L. E. (2023). *Pemeriksaan Uji Kenormalan dengan Kolmogorov-Smirnov, Anderson-Darling dan Shapiro-Wilk. Eigen Mathematics Journal, 6 (1), 11–19*.

The Modeling of The Poverty Rate In Indonesia From 2018 to 2023 Using A Panel Data Regression Approach

Dhyana Venosia[(1)], Toha Saifudin [(2)], Nur Chamidah[(3)]

[1,2,3]Departement of Mathematics, Faculty of Science and Technology, Airlangga University

Jl. Dr. Ir. H. Soekarno, Mulyorejo, Surabaya, Jawa Timur

e-mail: dhyana.venosia-2023@fst.unair.ac.id [(1)], tohasaifudin@fst.unair.ac.id [(2)], nur-c@fst.unair.ac.id[(3)]

ABSTRAK

Indonesia menduduki peringkat ke 6 dari 11 negara di Asia Tenggara dengan persentase kemiskinan tertinggi, sehingga diperlukan strategi dalam mengurangi masalah kemiskinan. Strategi yang diterapkan oleh pemerintah selaras dengan *Sustainable Development Goals* (SDGs) dengan tujuan utama *zero poverty*. Dalam penelitian ini, persentase kemiskinan di Indonesia tahun 2018-2023 mengalami peningkatan dan penurunan yang bervariatif. Fenomena tersebut membuktikan terjadinya dinamika persentase kemiskinan. Persentase kemiskinan diduga berkaitan erat dengan aspek pendidikan dan perekonomian. Merujuk pernyataan tersebut, penelitian ini melibatkan persentase kemiskinan sebagai variabel dependen serta akses air besih, gini ratio, tingkat pengangguran terbuka, dan angka melek huruf sebagai variabel independen. Bedasarkan struktur data penelitian, terjadinya dinamika persentase kemiskinan tahun 2018-2023 melibatkan struktur data *cross section* dan *time series*. Metode regresi data panel berbasis Random Effect Model dalam hal ini sangatlah tepat dan dapat mengakomodir proses identifikasi hingga perolehan kesimpulan. Penelitian ini bertujuan untuk mengidentifikasi faktor-faktor yang mempengaruhi persentase kemiskinan. Selanjutnya, penelitian ini menunjukkan bahwa seluruh variabel independen berpengaruh secara serempak maupun parsial terhadap persentase kemiskinan.

Kata kunci: Persentase Kemiskinan, Regresi Data Panel, Random Effect Model

ABSTRACT

Indonesia was ranked sixth out of eleven Southeast Asian countries with the highest poverty rate, highlighting the need for effective strategies to address poverty issues. The governmental strategy was aligned with the Sustainable Development Goals (SDGs), with the primary objective of achieving zero poverty. In this study, the poverty rate in Indonesia from 2018-2023 exhibited fluctuating trends, marked by both increases and decreases over the years. This phenomenon reflects the dynamic nature of poverty levels in the country. Poverty rates are assumed to be related to education and the economy. Referring to the statement, this study involves the percentage of poverty as the dependent variable and access to clean water access, gini ratio, open unemployment rate, and literacy rate as independent variables. Based on the structure of the research data, the dynamics of the poverty rate from 2018-2023 involve both cross-sectional and time series data structures. In this case, the panel data regression method based on the Random Effects Model is appropriate and can accommodate the identification process to the conclusion. This study aims to identify the factors that influence the poverty rate. Furthermore, the findings indicate that all independent variables have simultaneous and partial effects on the poverty rate.

Keywords: *Poverty Rate, Panel Data Regression, Random Effects Model*

INTRODUCTION

Indonesia is one of the developing countries that have been constantly working to eradicate poverty. Indonesia is ranked 6th in Southeast Asia for having the highest percentage of people living in poverty at 9.5% [1]. Poverty can be defined as a condition in which basic needs are not adequately met, particularly in the economic aspects. Alleviating poverty is explicitly stated as one of the goals of sustainable development, aligning with the Sustainable Development Goals (SDGs), with the primary objective of achieving zero poverty. Government steps to alleviate poverty are formulated and implemented through various subsidy programs [2]. Referring to the average poverty rate in Indonesia from 2018 to 2023, the governmental steps appear to be suboptimal, as the poverty rate consecutively stood at 10.61%, 10.24%, 10.81%, 10.43%, 10.30%, and 10.09% [3]. This situation has raised concerns within the community regarding the effectiveness of government efforts to alleviate poverty. However, poverty is not an easy problem, but there are approaches to overcome it. Government attention is considered crucial in overcoming poverty. Several efforts to eradicate poverty include the implementation of the *Indonesia Pintar* program, the KIP *Kuliah* program, the *Merdeka Belajar Kampus Merdeka* (MBKM) program, and village funds. These conditions show that the government is trying to eradicate poverty through the education and economic sectors [4].

Based on the report from the Central Bureau of Statistics, it has been identified that the percentage of poverty in Indonesia from 2018 to 2023 experienced fluctuations. Specifically, in 2019, it decreased by 0.37%; in 2020, it increased by 0.57%; in 2021, it decreased by 0.38%; in 2022, it decreased by 0.13%; and in 2023, it decreased by 0.21%. Fluctuations in the poverty percentage in Indonesia are naturally due to variations in each province. For example, Central Sulawesi Province experienced a decrease of 0.51%, 0.12%, and 0.88%, and an increase of 0.12% and 0.11%, respectively. Furthermore, East Kalimantan Province experienced a decrease of 0.15%, an increase of 0.73%, a decrease of 0.37%, an increase of 0.17%, and a decrease of 0.33%. The province of South Sumatra experienced a decrease of 0.31%, an increase of 0.51%, a decrease of 0.65%, 0.16%, and 0.18%, respectively [3]. Based on the programs implemented by the government in the education and economic sectors, poverty in Indonesia fluctuated from 2018 to 2023, representing a crucial issue due to its significant impact on societal welfare as a consequence of persistent poverty. Grounded in this information, this study holds significant relevance, as it utilizes the most recent data, which are assumed to deliver accurate assessments in identifying the factors that influence the poverty rate in Indonesia. Specifically, this study aims to provide valuable insights for the academic community in developing future research and to serve as a foundation for formulating poverty-related policies by the government [5]. One of the assessment approaches related to the factors influencing poverty is considered appropriate using the panel data regression method. It is due to the method's ability to accommodate the combination of cross-sectional and time series data, making it technically well-suited to the structure and framework required by the research.

Previous research in this study has served as the foundation for determining the independent variables, including the provision of access to clean water, which aims to meet basic needs and has the potential to alleviate poverty through improved health outcomes, such as a decrease in morbidity rates and an increase in the Human Development Index [6]. The implementation results of the Ordinary Least Squares (OLS) method stated that the access provision to clean water

contributes to alleviating poverty-related issues [7]. Specifically, poverty is closely related to the gini ratio, where positive and negative correlations may exist. The gini ratio itself essentially indicates income inequality, which impacts poverty. This finding was explored in previous research using the panel data regression method with a Fixed Effects Model [8]. The gini ratio has an indirect relationship with the open unemployment rate, where an increase in the open unemployment rate accompanies an increase in poverty, as evidenced by previous research through the implementation of the OLS method [9]. Economic factors contributing to poverty are highly correlated with educational factors. One of the educational factors examined is the literacy rate, where a previous study using the panel data regression method based on the Random Effects Model concluded that the literacy rate has a negative correlation with poverty.

Building on the previous research outlined above, the poverty rate is specifically hypothesized to be influenced by economic and educational factors, including access to clean water, the gini ratio, the open unemployment rate, and the literacy rate. These factors are examined across provinces in Indonesia as cross-sectional units, with the period from 2018 to 2023 as the time series unit. According to the statement, panel data regression is highly relevant for this study. Panel data can be defined as a dataset that tracks specific individual samples over time, thereby providing multiple observations for each individual in the sample [5]. In this research, the application of panel data regression was simulated using the R programming software.

METHOD

Data and Data Sources

This research is classified as quantitative research involving secondary data from Central Bureau of Statistics publications in 2023 entitled *Indikator Tujuan Pembangunan Berkelanjutan Indonesia* 2023. The data involved in this research are poverty data and its influencing factors, including access to clean water, gini ratio, the open unemployment rate, and the literacy rate from 2018 to 2023. The data involved as the research implementor is from 34 provinces in Indonesia. [3].

Variables

The variables used in this study are divided into independent and dependent variables, which are described in detail as follows:

Table 1. Variables

Dependent Variable	Y	Poverty Percentage (%)
Independent Variable	X_1	Clean Water Access (%)
	X_2	Gini Ratio (%)
	X_3	Open Unemployment Rate (%)
	X_4	Literacy Rate (%)

The dependent variable is based on the aim of this study, which focuses on the poverty rate. The determination of the independent variables is grounded in previous research, which assumed that they affect the poverty rate, particularly in education and economics aspects.

Descriptive Statistics

In this study, the implementation of descriptive statistics aims to describe the research variables, complemented by visualizations and analysis results, which are fundamental components of the research. The description of the research variables is intended to provide a data summary that helps facilitate a clearer understanding of the research data presentation [10].

Panel Data Regression

Panel data regression is an identification method implemented through cross-sectional and time-series data integration within a single equation [5]. In general, the panel data regression model can be expressed as follows [11]:

$$Y_{it} = \alpha_{it} + \boldsymbol{X}_{it}\boldsymbol{\beta} + u_{it};\ i = 1,2,\ldots,\mathrm{N};\ t = 1,2,\ldots,\mathrm{T}$$

Based on the equation above, the panel data regression model is classified into five types, which can be expressed as follows [12]:

(1) All slope coefficients and intercepts are constant across individuals and time

$$Y_{it} = \alpha^* + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

(2) The slope coefficients are constant, while the intercepts vary across individuals

$$Y_{it} = a_i^* + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

(3) The slope coefficients are constant, while the intercepts vary across individuals and time

$$Y_{it} = a_{it}^* + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

(4) All coefficients vary across individuals

$$Y_{it} = a_i^* + \sum_{k=1}^{K} \beta_{ki} X_{kit} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

(5) All coefficients vary across time and individuals

$$Y_{it} = a_{it}^* + \sum_{k=1}^{K} \beta_{kit} X_{kit} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

In its implementation, the panel data regression method generally uses three approaches, including [13]:

Common Effect Model (CEM)

The Common Effect Model assumes that the intercept values for each variable are the same, as are the slope coefficients for all cross-sectional units and time series, which can be expressed through the following equation [14]:

$$Y_{it} = \alpha + \boldsymbol{X}_{it}\boldsymbol{\beta} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

Fixed Effect Model (FEM)

The Fixed Effect Model assumes that the intercept values for each variable are different for each cross-sectional unit, and the slope is presumed to be constant, which can be expressed through the following equation [11]:

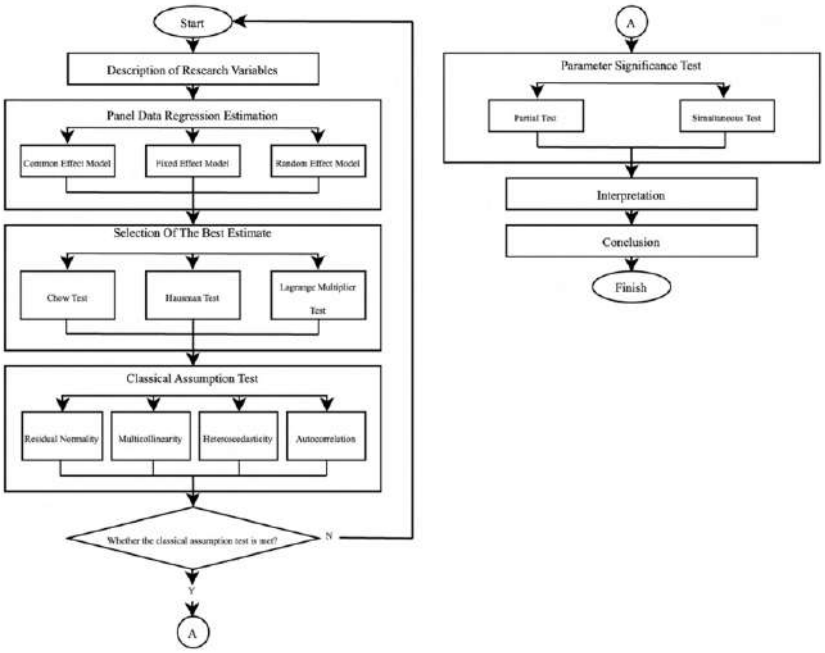
$$Y_{it} = \alpha_i + \boldsymbol{X}_{it}\boldsymbol{\beta} + u_{it};\ i = 1, 2, \ldots, \mathrm{N};\ t = 1, 2, \ldots, \mathrm{T}$$

Random Effect Model (REM)

The Random Effect Model approach assumes that the individual effects are random for all cross-sectional units, which can be expressed through the following equation [11]:

$$Y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kit} + u_{it} ; i = 1, 2, \dots, N ; t = 1, 2, \dots, T$$

Flowchart



Picture 1. Flowchart

RESULT AND DISCUSSION

Based on the research data used, further analysis is carried out with the following test results:

Description of Research Variables

The description of the research variables related to the poverty rate in Indonesia, along with the variables presumed to have an influence, can be identified using descriptive statistical methods, as shown in **Table 2.** to **Table 6.** below:

Table 2. Poverty Percentage

Year	Mean	Minimum	Province of Minimum Value	Maximum	Province of Maximum Value
2018	10.61%	3.55%	Jakarta	27.43%	Papua
2019	10.24%	3.42%	Jakarta	26.55%	Papua
2020	10.81%	4.45%	Bali	26.80%	Papua
2021	10.43%	4.56%	South Kalimantan	27.38%	Papua
2022	10.30%	4.53%	Bali	26.80%	Papua
2023	10.09%	4.25%	Bali	26.03%	Papua

Table 3. Clean Water Access

Year	Mean	Minimum	Province of Minimum Value	Maximum	Province of Maximum Value
2018	72.95%	49.37%	Bengkulu	90.90%	Bali
2019	84.23%	57.60%	Bengkulu	99.82%	Jakarta
2020	85.41%	64.47%	Bengkulu	99.84%	Jakarta
2021	86.68%	64.92%	Papua	99.86%	Jakarta
2022	87.64%	65.39%	Papua	98.42%	Bali
2023	88.19%	66.49%	Papua	99.42%	Jakarta

Table 4. Gini Ratio

Year	Mean	Minimum	Province of Minimum Value	Maximum	Province of Maximum Value
2018	0.35%	0.272%	Bangka Belitung Islands	0.422%	Yogyakarta
2019	0.35%	0.262%	Bangka Belitung Islands	0.428%	Yogyakarta
2020	0.38%	0.257%	Bangka Belitung Islands	0.437%	Yogyakarta
2021	0.35%	0.247%	Bangka Belitung Islands	0.436%	Yogyakarta
2022	0.34%	0.255%	Bangka Belitung Islands	0.459%	Yogyakarta
2023	0.34%	0.245%	Bangka Belitung Islands	0.245%	Yogyakarta

Table 5. Open Unemployment Rate

Year	Mean	Minimum	Province of Minimum Value	Maximum	Province of Maximum Value
2018	4.8%	1.40%	Bali	8.47%	Banten
2019	4.7%	1.57%	Bali	8.11%	Banten
2020	6.0%	1.32%	West Sulawesi	10.95%	Jakarta
2021	5.5%	3.01%	West Nusa Tenggara Gorontalo	9.91%	Riau Islands
2022	5.0%	2.34%	West Sulawesi	8.31%	West Java
2023	4.6%	2.27%	West Sulawesi	7.52%	Banten

Table 6. Literacy Rate

Year	Mean	Minimum	Province of Minimum Value	Maximum	Province of Maximum Value
2018	99.45%	88.44%	Papua	100%	Yogyakarta East Kalimantan
2019	99.56%	90.39%	Papua	100%	Yogyakarta East Kalimantan
2020	99.58%	90.78%	Papua	99.98%	Aceh
2021	99.59%	91.13%	Papua	99.99%	Aceh
2022	99.63%	92.04%	Papua	99.97%	West Java
2023	99.68%	93.57%	Papua	99.97%	Bangka Belitung Islands West Java

Based on the results of the identification of the description of the research variables, information was obtained that each independent and dependent variable from 2018 to 2023 experienced fluctuations identified through decreases and increases each year, thus creating dynamics for each research variable.

Panel Data Regression Estimation

This study implements the panel data regression method to estimate the effect of independent variables on the dependent variable using three approaches, namely CEM, FEM, and REM, as shown in **Table 7.** below:

Table 7. Panel Data Regression Estimation

	R-Squared	*p-value*
CEM	40.046%	< 0.05
FEM	33.031%	< 0.05
REM	28.568%	< 0.05

Selection of the Best Estimate

According to the panel data regression assessment, a suitable approach is needed to analyze the research data through the Chow, Hausman, and Lagrange Multiplier tests. Specifically, the Chow test determines the estimation between CEM and FEM, the Hausman test is aimed at choosing the estimation between FEM and REM, and the Lagrange Multiplier test determines the estimation between REM and CEM [12]. Meanwhile, the test results are informed through **Table 8**. as follows:

Table 8. Selection of the Best Estimate

	p-value	**Decision**	
Chow	< 0.05	Reject H_0	FEM
Hausman	> 0.05	Accept H_0	REM
Lagrange Multiplier	< 0.05	Reject H_0	REM

Based on the best estimation selection tests, the REM is the most appropriate approach to implement.

Classical Assumption Test

The classical assumption test is a requirement that must be fulfilled in conducting a test so that the estimation results are the Best Linear Unbiased Estimator (BLUE). The classical assumption tests include the residual normality, multicollinearity, heteroscedasticity, and autocorrelation test [12].

Residual Normality Test

The residual normality test can be implemented through the Kolmogorov-Smirnov test [12]. The test results state that the p-value is 0.3929 which underlies the decision to accept H_0. Therefore, it can be concluded that the test data residuals fulfill the assumption of residual normality.

Multicollinearity Test

The multicollinearity test for each independent variable can be determined using the VIF value [12]. The results of the VIF test are presented in **Table 9.** as follows:

Table 9. VIF

	VIF
Clean Water Access (X_1)	1.17
Gini Ratio (X_2)	1.14
Open Unemployment Rate (X_3)	1.03
Literacy Rate (X_4)	1.04

Heteroscedasticity Test

The heteroscedasticity test is used to identify differences in the variance of residuals [15]. The results showed a p-value < 0.05, indicating the presence of heteroscedasticity in the estimation process.

Autocorrelation Test

The autocorrelation test is intended to examine the correlation between the standard error at a specific lag time and its previous value. This test aims to avoid serial correlation issues, which can affect the efficiency of the test results [16]. The test results showed a p-value < 0.05, indicating that the estimation process faces autocorrelation issues.

Addressing Violations of Classical Assumptions

Based on the tests and analysis conducted, it was found that the estimation using the REM approach faced issues of heteroscedasticity and autocorrelation. Therefore, corrective measures were applied by using robust estimation methods. The results from the robust estimation are presented in **Table 10.** as follows:

Table 10. Robust Estimation

	p-value
Clean Water Access (X_1)	< 0.05
Gini Ratio (X_2)	< 0.05
Open Unemployment Rate (X_3)	< 0.05
Literacy Rate (X_4)	< 0.05

Significance Test

A significance test for parameters is necessary to analyze the effect of independent variables on the dependent variable simultaneously and partially. The significance test is divided into two types: simultaneous test and partial test [12].

Simultaneous Test

Based on the test using the REM estimation approach, a p-value of less than 0.05 was obtained, indicating that, simultaneously, access to clean water, gini ratio, open unemployment rate, and literacy rate have an impact on the poverty rate.

Partial Test

The partial test provides information on the partial effects on the poverty rate. The results of the partial test are presented in **Table 11.** as follows:

Table 11. Partial test

	p-value	Decision
$\hat{\beta}_1$	< 0.05	Reject H_0
$\hat{\beta}_2$	< 0.05	Reject H_0
$\hat{\beta}_3$	< 0.05	Reject H_0
$\hat{\beta}_4$	< 0.05	Reject H_0

Based on the partial test conducted, it can be concluded that, individually, access to clean water, gini ratio, open unemployment rate, and literacy rate influence the poverty rate.

Interpretation

Based on the tests conducted, the panel data regression equation obtained using the REM approach is as follows:

$$Y_{it} = 32.3479228 - 0.0174982X_{kit} + 9.2621172X_{kit} + 0.2296808X_{kit} - 0.2495794X_{kit}$$

The estimation of the test results can be interpreted as follows:

	Interpretation
$\hat{\beta}_1$	If the percentage of access to clean water increases by 1%, the poverty rate will decrease by 0.0174982%, assuming all other variables remain constant.
$\hat{\beta}_2$	If the percentage of the gini ratio increases by 1%, the poverty rate will increase by 9.2621172%, assuming all other variables remain constant.
$\hat{\beta}_3$	If open unemployment rate increases by 1%, the poverty rate will increase by 0.2296808%, assuming all other variables remain constant.
$\hat{\beta}_4$	If literacy rate increases by 1%, the poverty rate will decrease by 0.2495794%, assuming all other variables remain constant.

Coefficient of Determination

The conducted tests obtained an R-squared value of 28.568%. This R-squared value indicates the independent variables explain 28.568% of the variation in the poverty rate, while the explanation by other variables outside the scope of the study is 71.432%.

CONCLUSION

Based on the analysis results presented, the following conclusions as follows:

1. The poverty rate and its variables influencing it, such as access to clean water, the gini ratio, the unemployment rate, and the literacy rate, fluctuated across provinces in Indonesia from 2018 to 2023.
2. The regression equation estimation using the Random Effect Model approach can be stated as follows:
$$Y_{it} = 32.3479228 - 0.0174982X_{kit} + 9.2621172X_{kit} + 0.2296808X_{kit} - 0.2495794X_{kit}$$
3. Based on the estimation results, all independent variables influence the poverty rate, both simultaneously and partially.
4. The R-squared value of 28.568% indicates that the ability of the independent variables to explain the percentage of poverty is 28.568%. Indirectly, it can be assumed that government programs in the economy and education have contributed to poverty alleviation by optimizing access to clean water, the gini ratio, the open unemployment rate, and the literacy rate.

REFERENCE

[1] Wardhana, A. and Kharisma, B. 2019. Peran Pengeluaran Sektor Pendidikan dan Sektor Kesehatan Terhadap Kemiskinan di Indonesia. *Jurnal Ekonomi dan Bisnis Universitas Udayana*, **8**(12), 1343-1366.

[2] Pramesty, M. H. E., Ibrahim, C. A., Rahmawati, N., Amin, M. I. N. A, Hapsery, A., and Meimela, A., 2023, Pemodelan Kemiskinan di Indonesia dengan Metode Structural Equation Modelling-Partial Least Square (SEM-PLS). *Prosiding Seminar Nasional Hasil Riset dan Pengabdian*, **5**, 746-753.

[3] Badan Pusat Statistik. 2023. *Indikator Tujuan Pembangunan Berkelanjutan Indonesia 2023*. Indonesia: Badan Pusat Statistik.

[4] Murdiyana and Mulyana. 2017. Analisis Kebijakan Pengentasan Kemiskinan di Indonesia. *Jurnal Politik Pemerintahan*, **10**(1), 73–96.

[5] Ghozi, S. and Hermansyah, H. 2018. Analisis Regresi Data Panel Profitabilitas Bank Pembangunan Daerah (BPD) di Indonesia. *Jurnal Matematika*, **8**(1), 1–12.

[6] Yesi, D. and Juairiyah, O. 2021. Sebaran Tingkat Kemiskinan dan Tingkat Akses Air Bersih di Sumatera Selatan. *Dinamika Lingkungan Indonesia*, **8**(1), 11–16.

[7] Budiono, S. and Purba, J. T. 2022. Reducing Poverty Strategy Through Educational Participation, Clean, Water, and Sanitation in Indonesia. *Jurnal Ekonomi dan Bisnis*, **25**(1), 177–198.

[8] Endrawati, D., Nujum, S., and Selong, A. 2023. Pengaruh Pertumbuhan Ekonomi, Rasio Gini, dan Indeks Pembangunan Manusia Terhadap Tingkat Kemiskinan Indonesia 2017-2022. *Jurnal Pendidikan Tambusai*, **7**(3), 20144–20151.

[9] Sagala, R., Harlen, Utami, B. C. 2023. Pengaruh Pertumbuhan Ekonomi dan Tingkat Pengangguran Terbuka (TPT) Terhadap Kemiskinan di Kabupaten Pelalawan. *Jurnal Niara*, **15**(3), 514–524.

[10] Tarigan, M. and Silaban, D. F. 2024. Statistika Deskriptif. *Jintan: Jurnal Ilmu Keperawatan*, **4**(2), 187-195.

[11] Nandita, D. A., Alamsyah, L. B., Jati, E. P., and Widodo, E. 2019. Regresi Data Panel untuk Mengetahui Faktor-faktor yang Mempengaruhi PDRB di Provinsi DIY Tahun 2011-2015. *Indonesian Journal of Applied Statistics*, **2**(1), 42-52.

[12] Venosia, D., Suliyanto., Sediono., and Chamidah, N. 2022. Pemodelan Persentase Kepesertaan BPJS Non Penerima Bantuan Iuran dengan Pendekatan Regresi Data Panel. *Jurnal Ilmiah Teori dan Aplikasi Statistika*, **15**(1), 116–126.

[13] Firdaus, Y. A., Ngatini., and Wijaya, S. U. 2023. Pemodelan Regresi Data Panel Harga Beras di Wilayah Indonesia Bagian Barat. *Jurnal Ilmiah Teori dan Aplikasi Statistika*, **16**(2), 486–498.

[14] Rahmadeni, R. and Wulandari, N. 2017. Analisis Faktor-faktor yang Mempengaruhi Inflasi pada Kota Metropolitan di Indonesia dengan Menggunakan Analisis Regresi Data Panel. *Jurnal Sains Matematika dan Statistika*, **3**(2), 34–42.

[15] Indri, F. Z. and Putra, G. H. 2022. Pengaruh Ukuran Perusahaan dan Konsentrasi Pasar Terhadap Kualitas Laporan Keuangan pada Perusahaan Sektor Industri Barang Konsumsi yang Terdaftar di Bursa Efek Indonesia pada Tahun 2016-2020. *Jurnal Ilmu Manajemen Ekonomi dan Kewirausahaan*, **2**(2), 1-17.

[16] Khosropour, A. 2017. A Panel Data Analysis of The Relationship Between Corporate Social Responsibility and Earnings Management: Evidence from Iran. *Revista Quid*, 2423–2431.

Forecasting Stock Prices Using a Nonlinear Approach with the Exponential Smooth Transition Autoregressive (ESTAR) Model

Allan Ruhui Fatmah Sari [(1)], Dwi Arman Prasetya [(2),] Trimono[(3)]

[1,2,3]Universitas Pembangunan Nasional Veteran Jawa Timur

Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur

e-mail: 210831010007@student.upnjatim.ac.id[(1)], arman.prasetya.sada@upnjatim.ac.id[(2)], trimono.stat@upnjatim.ac.id[(3)]

ABSTRAK

Peningkatan minat masyarakat Indonesia dalam berinvestasi pada aset keuangan tunggal telah mendorong pertumbuhan pasar modal secara signifikan. Namun, tingginya volatilitas pasar membawa risiko kerugian yang perlu diantisipasi. Salah satu model yang relevan dalam menangani permasalahan tersebut yaitu menggunakan model nonlinier *Exponential Smooth Transition Autoregressive* (ESTAR). ESTAR merupakan perluasan dari model *Autoregressive* (AR) yang menggunakan transisi lebih halus untuk menangani data deret waktu yang tidak linear. Penelitian ini bertujuan untuk memprediksi harga saham menggunakan model nonlinier ESTAR guna membantu investor menghadapi ketidakpastian pasar dan mengelola risiko jangka pendek. Data yang digunakan berupa harga penutupan harian saham PT Bank Central Asia Tbk, periode Januari 2022 hingga Desember 2024. Metodologi penelitian mencakup uji stasioneritas, estimasi parameter AR(*p*), estimasi parameter model ESTAR(*p,d*), hingga evaluasi akurasi prediksi menggunakan *Mean Absolute Percentage Error* (MAPE). Hasil penelitian menunjukkan bahwa model AR(1) merupakan model dengan orde terbaik dan model ESTAR(1,1) sebagai model akhir yang optimal. Evaluasi hasil prediksi untuk periode satu bulan ke depan, menunjukkan bahwa nilai MAPE sebesar 2,79% yang mengindikasikan performa model dalam memprediksi harga saham sangat baik.
Kata kunci: Deret waktu; ESTAR; Prediksi Saham

ABSTRACT

Increased interest among Indonesians in investing in a single financial asset has driven significant growth in the capital market. However, high market volatility brings the risk of loss that needs to be anticipated. One of the relevant models in dealing with these problems is using the nonlinear Exponential Smooth Transition Autoregressive (ESTAR) model. ESTAR is an extension of the Autoregressive (AR) model that uses smoother transitions to handle nonlinear time series data. This study aims to predict stock prices using the ESTAR nonlinear model to help investors deal with market uncertainty and manage short-term risk. The data used is the daily closing price of PT Bank Central Asia Tbk shares, for the period January 2022 to December 2024. The research methodology includes stationarity test, AR(p) parameter estimation, ESTAR(p,d) model parameter estimation, and prediction accuracy evaluation using Mean Absolute Percentage Error (MAPE). The results show that the AR(1) model is the best order model and the ESTAR(1,1) model is the final optimal model. Evaluation of the prediction results for the next one month period, shows that the MAPE value is 2.79% which indicates the model's performance in predicting stock prices is very good.
***Keywords**: Time Series; ESTAR; Stock Prediction*

INTRODUCTION

In recent years, the Indonesian stock market has often experienced significant fluctuations influenced by various economic, social and political sectors [1]. These fluctuations cause stock prices to become uncontrollable, resulting in increased stock market volatility, thus triggering an increased risk of loss to investors. Stock volatility is caused by overall uncertainty, such as the COVID-19 pandemic, unfavorable world trade activities, and changes in monetary policy [2]. Then, the phenomenon of declining purchasing power of the upper middle class is one of the reasons why stock prices will experience up and down movements every day [3]. This happens because instead of buying, people now have a tendency to save and the value of inflation is high [4].

This complex situation occurs in one of the company's shares in Indonesia, such as PT Bank Central Asia Tbk. The company is engaged in the financial services industry sector that focuses on commercial and consumer banking. BCA has a large influence in the Indonesian capital market which reflects investor confidence in national economic stability with the stock code BBCA.JK. The share price chart of PT Bank Central Asia Tbk over the past year shows a decline in April which does indicate a change in purchasing power. This volatile movement explains market conditions and can be one of the indicators that investors should pay attention to. Therefore, proper risk management is required by reducing losses and the uncertainty of stock price movements. The most important factor in making a risk management policy is to consider the stock price factor and the predicted risk of loss.

Under circumstances of stock market uncertainty, the linear approach is often unable to capture nonlinear changes. Therefore, a nonlinear approach that can capture better patterns and provide accurate predictions is needed as a solution to the problem. The data used is historical stock price data which is an important and difficult form of time series data in data science and machine learning because it requires accurate and reliable modeling for data that is very complex and changes over time [5], [6]. One of the time series models to predict stock prices is using the Threshold Autoregressive (TAR) model. Then the TAR model is generalized to get a smoother transition than Autoregressive (AR) or Smooth Transition Autoregressive (STAR). There are two types of STAR, the first is Exponential Smooth Transition Autoregressive (ESTAR) and the second is Logistic Smooth Transition Autoregressive (LSTAR) [7], [8]. Stock price data that has an up and down movement by showing a smooth non-linear pattern of change and experiencing symmetrical changes is more suitable for the ESTAR model. Meanwhile, LSTAR is more suitable for data that experience asymmetric changes.

Although many studies on stock price forecasting have been conducted, most of the existing studies use linear models, such as ARIMA, LSTM or non-linear approaches in general without utilizing the ESTAR model. Based on these problems, this study aims to forecast the stock price of PT Bank Central Asia Tbk (BBCA.JK) and fill the gap in the literature in applying nonlinear models, especially the ESTAR model. This approach is expected to make a significant contribution in helping investors face challenges in decision-making amid high market volatility.

METHOD

This study uses historical stock data on the company PT Bank Central Asia Tbk with the company code BBCA.JK. Data taken from January 2022 - December 2024 from the Investing website (www.investing.com), where the data for January 2022 - October 2024 is insample and the

data for November - December 2024 is outsample. The historical stock price dataset used in this study is the closing price variable (close). In this research, Python is used to implement stock prediction using the ESTAR model. The following is a flowchart of the research to be carried out in Figure 1.

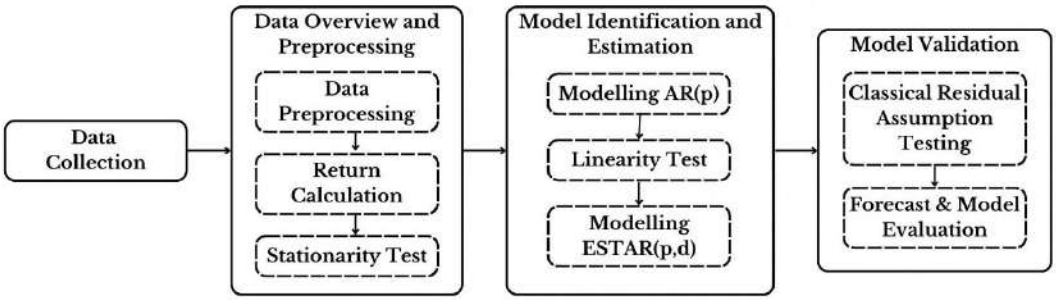


Figure 1. Flowchart of ESTAR modeling

1. Data Overview and Preprocessing

a. Data Preprocessing

Data preprocessing is an important process to improve data quality and data reliability [9]. This step has three processes, namely performing time stamp conversion, checking data type, checking missing values, checking duplicate data, and separating data as insample and outsample. After this process, it will be checked whether the time series data has stationary properties or not.

b. Return Calculation

Stock returns are defined as the returns that investors get from previous capital investments. Return can be done by calculating the natural logarithm, which is multiplied by the ratio between the price at time t and the price of the previous time period.

c. Stationarity Test

The characteristics of stationary data on variance have a lower limit and an upper limit that shows a number or $\lambda = 1$. Meanwhile, stationary data on the mean will fluctuate around the mean line which is close to 0. The purpose of the stationarity test is to fulfill the prerequisites of a time series model that has stationary properties, namely the average and variance do not change over time [10]. Augmented Dickey-Fuller (ADF) test is employed as a statistical method to formally test for the presence of unit roots [11].

2. Model Identification and Estimation

a. Model Selection and Parameter Estimation of AR(p)

Autoregressive model is a time series data regression modeling that relates the actual observation value to the previous observation value [12]. AR parameter estimates are obtained using the Least Square Method, which reduces the sum of the following residual squares:

$$\sum_{t=2}^{T} a_t^2 = SSE = \sum_{t=2}^{T} (X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p})^2 \quad (1)$$

Furthermore, to select the best model of the AR(p) order, we can use the Akaike Info Criterion (AIC) value [13]. According to Van Dijk (2000) [14], the AIC formula is as follows.

$$AIC = T \ln \hat{\sigma}_a^2 + 2k, \text{ where } \hat{\sigma}_a^2 = SSE = \sum_{t=1}^{T} \hat{a}_t^2 \quad (2)$$

b. Linearity Test

The purpose of the LM_3 test is to detect whether or not the model is nonlinear in time series data. The linearity test can be conducted using the Lagrange Multiplier (LM) test method, specifically LM_3 [15], derived from auxiliary regression. The estimation of the auxiliary regression model is performed using Ordinary Least Squares (OLS). Testing using LM_3 is carried out as follows.

H_0: $\phi_{1,i} = \phi_{2,i}$; the model is linear.

H_1: $\phi_{1,i} \neq \phi_{2,i}$; with at least one i $\in$ {1,2,...,p} nonlinear model.

Test statistic:

$$LM_3 = T \frac{(SSE_0 - SSE_1)}{SSE_0} \quad (3)$$

Description:

SSE_0 : Sum of squares of AR(p) model residuals

SSE_1 : Sum of squares of model residuals from auxiliary regression

The test rejection criterion is if H_0 is rejected with a statistical value of $LM_3 > \chi^2_{3(p+1)}$

c. ESTAR Estimation and Modeling

The ESTAR model can be estimated using the Nonlinear Least Square (NLS) method which is suitable for the nonlinear structure of the ESTAR model because it can estimate the parameters γ (smoothness) and c (threshold). This method is carried out for the process of finding parameter values using numerical methods, namely the Gauss-Newton method to iterate on the estimation. According to Terasvirta (1994) [16], writing the function on the exponential transition as follows [17].

$$G(s_t; \gamma; c) = 1 - \exp(-\gamma(s_t - c)^2), \gamma > 0 \quad (4)$$

So the following ESTAR model is obtained.

$$X_t = \boldsymbol{\phi_1' Z_t}\left(1 - (1 - \exp(-\gamma(s_t - c)^2))\right) + \boldsymbol{\phi_2' Z_t}(1 - \exp(-\gamma(s_t - c)^2)) + \varepsilon_t \quad (5)$$

3. Model Validation

a. Classical Residual Assumption Testing

In the classical assumption testing of residuals, there are three types of tests that need to be conducted to ensure the validity of the model. First, the autocorrelation assumption test is carried out to determine whether the model exhibits a correlation between a period t and the previous period (t-1) [18], [19]. Second, the heteroscedasticity assumption test is performed to assess whether the regression model between observations shows differences in variances and residuals [20]. Third,

the normality assumption test is conducted to evaluate whether the residual values in the model are normally distributed [21], [22].

b. Forecast and Model Evaluation

Stock price prediction involves forecasting future stock values using historical data and appropriate modeling techniques such as time series analysis or machine learning algorithms. To evaluate the performance of the prediction models, the Mean Absolute Percentage Error (MAPE) is utilized, providing a measure of the average deviation between the predicted and actual stock prices in percentage terms [23], [24], [25], [26].

RESULT AND DISCUSSION

Data Preprocessing

The following figure presents a graphical representation of the price trajectory of BBCA.JK shares.

Figure 2. Close price plot line BBCA.JK

Figure 2 shows a plot line visualization of BBCA.JK stock closing price data. The figure shows the movement of the closing share price for the period January 2022 - December 2024. The smallest share price is 7000, the highest share price is 10950 and the average share price is 8954.29. In this process, check the date data type to datetime64[ns]. Then, after checking the data, the results do not occur missing values and duplicate data. Then, separating data in the January 2022 - October 2024 range as insample data and data in the November - December 2024 range as outsample data.

Return

In the picture below is the movement of BBCA.JK stock price returns.

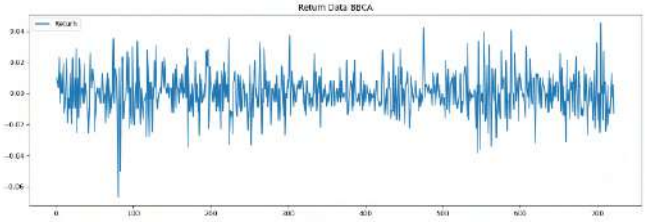


Figure 3. BBCA.JK stock return chart

Figure 3 shows that the return value of BBCA.JK shares has a stationary condition which can then be tested using the ADF test.

Stationarity Test

The results of stationarity testing using the ADF test at the 5% significance level can be seen in the following table.

Table 1. Stasionerity Test with ADF

ADF-statistic	p-value	Critical value
-21.7301	0.00	-2.865568

Based on the table, it can be concluded that the ADF statistic value is smaller than the critical values and the ADF test p-value is smaller than $\alpha = 0.05$, which means that it succeeds in rejecting H_0, namely the data does not have unit roots or stationary data.

Estimating and Testing AR (p) Model Parameters

The process of building an ARIMA model is done from the previous return data. Then the visualization of ACF and PACF is done to determine the order (lag) of the optimal AR model by eliminating the influence of the lag between previous values, as follows.

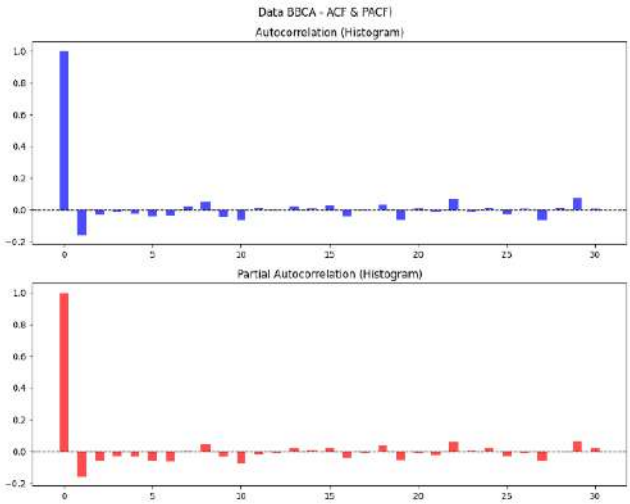


Figure 4. ACF and PACF plots

Then to build an AR model using the library from statsmodels.tsa.arima.model import ARIMA and identify the best AR(p) model with the AIC method. The result is that the AR(1) model is the best model. Furthermore, estimation of the model is carried out using the OLS method.

Table 2. AR(1) Model parameter identification and estimation

AR Model	AIC	Significant Test
(1,0)	**-3995.362**	Yes
(2,0)	-3994.998	No
(3,0)	-3993.991	No
⋮	⋮	⋮

Based on this table, the transition variable used in the STAR model is obtained from the AR(1) model with an AIC value of -3995.362 and a p-value that is less than α = 0.05.

Linearity Test

Furthermore, linearity testing is carried out to determine whether the model requires a linear or nonlinear model using the LM_3 test with auxiliary regression. The estimation of the auxiliary regression model is done using OLS. The following are the estimation results of the auxiliary regression model with the transition variable $s_t = X_{t-1}$.

Table 3. Auxiliary regression model estimation with transition variable

Parameter	Coefficient	t-value	Probability
Intercept	0.0001	10.482	0.000
$\boldsymbol{\beta_{1,1}}$	-0.0022	-1.730	0.084
$\boldsymbol{\beta_{2,1}}$	0.1233	2.964	0.003
$\boldsymbol{\beta_{3,1}}$	2.2336	1.965	0.050

The following is the LM_3 test table with a significance level of 5%

Table 4. LM_3 test result

LM-statistic	p-value	Critical value (Chi-Square)
680.596	0.0000	12.591

Based on these results, it can be obtained that the LM-statistic value is greater than the critical value and which means that it succeeds in rejecting H_0 so that the model is nonlinear.

ESTAR Estimation and Modeling

The best model estimation result is AR(1) so that the following form is obtained.

$$Y_t = -0.1582Y_{t-1} + e_t$$

After knowing the order used is 1, to determine the delay value for the ESTAR model using AIC and obtained the best delay is d = 1. ESTAR (1,1) using the Nonlinear Least Square method is as follows.

Table 5. Parameter estimation of ESTAR(1,1) model

Parameter	Coefficient	t-value	Probability
$\boldsymbol{\varphi_{1,0}}$	-0.000930	-0.013671	0.989097
$\boldsymbol{\varphi_{1,1}}$	-0.114007	-0.028132	0.977565
$\boldsymbol{\varphi_{2,0}}$	0.915124	0.083459	0.933511
$\boldsymbol{\varphi_{2,1}}$	17.929593	0.085525	0.931870
$\boldsymbol{\gamma}$	6.525581	0.091928	0.926783
$\boldsymbol{c}$	0.013083	0.036304	0.971051

So that the ESTAR(1,1) model obtained is as follows:

$$X_t = (-0.000930 - 0.114007X_{t-1})\left(1 - (1 - \exp(6.525581(X_{t-1} - (0.013083))^2))\right) + (0.915124 + 3.064531X_{t-1})(1 - \exp(6.525581(X_{t-1} - (0.013083))^2)) + \varepsilon_t$$

Classical Residual Assumption Test

1. Autocorrelation Assumption Test

Testing autocorrelation in this study uses the Ljung Box-Pierce test. The following are the test results:

Table 6. Residual autocorrelation test results

Lag	Q-Stats	p-value
1	0.018032	0.893179
2	3.240676	0.197832
⋮	⋮	⋮
19	21.561696	0.306610
20	21.566417	0.364494

The test uses a significance level of 5% and shows that at the 1st to 20th lag there is no significant autocorrelation in the model.

2. Heteroscedastisity Assumption Test

In testing heteroscedasticity using the Lagrange Multiplier test where the results show that using a significance level of 5%, the LM value = 2.7457 and p-value = 1.0000 which is more than the α = 0.05 value so that the residual data model is homogeneous or there is no heteroscedasticity and there is no ARCH effect in the residuals.

3. Normality Assumption Test

Next, conduct a normality test using the Kolmogorov-Smirnov test. Based on this test using a 5% significance level, the results obtained are the statistical value of *D* = 0.0431 and the p-value = 0.1549 where the p-value is more than α = 0.05. This means that the residual data of the ESTAR(1,1) model is proven to be normally distributed.

Stock Price Prediction and MAPE Calculation

The ESTAR (1,1) model used is as follows:

$$X_t = (-0.000930 - 0.114007X_{t-1})\left(1 - (1 - \exp(6.525581(X_{t-1} - (0.013083))^2))\right) + (0.915124 + 3.064531X_{t-1})(1 - \exp(6.525581(X_{t-1} - (0.013083))^2)) + \varepsilon_t$$

The prediction results of the daily stock price of PT Bank Central Asia Tbk for the next four months are as follows.

Table 7. Comparison of stock price prediction results with original data

Date	Data Original	Data Prediction
01/11/2024	10425	10290.33
04/11/2024	10375	10281.69
⋮	⋮	⋮
30/12/2024	9800	10311.11

31/12/2024	9675	10311.88

In the prediction results of the model, further prediction evaluation is carried out using Mean Absolute Percentage Error (MAPE). The MAPE value shows a fairly small result of 2.79% in the model so that the model can be very good at predicting. Based on the prediction results, the ESTAR(1,1) model produces a relatively constant and small BBCA stock return prediction. This shows that the model is able to capture the gradual nonlinear transition characteristics in accordance with the basic characteristics of the model that accommodates nonlinear transitions smoothly between regimes. So that the model provides an indication that in the final period of 2024, the BBCA.JK stock market tends to experience volatility that is not extreme and is in a relatively stable adjustment phase. This situation reflects the attitude of investors who are still cautious in making decisions because the economic conditions after the pandemic have not fully stabilized.

CONCLUSION

This research successfully developed a stock price prediction model using the ESTAR approach to overcome the complexity of financial time series data. By applying the ESTAR model, this study achieved a MAPE of 2.79%, which indicates a high level of accuracy in forecasting stock price movements. This finding emphasizes the effectiveness of incorporating nonlinear dynamics in financial forecasting, which allows for better alignment between forecasted prices and actual market behaviour. However, the model has an important limitation in that it does not account for external economic shocks or the influence of global financial indicators, both of which can have a significant impact on stock prices. As a result, its predictive performance is limited under highly dynamic market conditions. To improve its robustness, future research could integrate the ESTAR model with hybrid approaches, such as artificial neural networks (ANN), to improve its adaptability to complex and nonlinear patterns. In addition, exploring multi-asset forecasting could expand the applicability of this model in the broader context of portfolio analysis.

REFERENCE

[1] M. Qibtiyah and P. Widodo, "Analisis Fluktuasi Ekonomi Di Indonesia Terhadap Output Gap Potensial Tahun 1993 - 2022," *J. Dev. Econ. Digit.*, vol. 3, no. 1, pp. 31–45, 2024.

[2] R. Rosihan, I. Nurhayati, and R. S. Aminda, "Analisis Volatilitas Harga Saham Terhadap Indeks Harga Saham Gabungan (Ihsg) Periode Maret 2019 – Februari 2021," *Bus. Manag. Anal. J. BMAJ*, vol. 5, no. 2, pp. 175–188, Oct. 2022, doi: 10.24176/bmaj.v5i2.7667.

[3] J. Yutanesy and R. Suhendah, "Perubahan Harga, Volume Saham, dan Kapitalisasi Pasar Selama COVID-19 pada Sektor Keuangan," *J. Ekon.*, vol. 27, no. 03, pp. 159–181, Mar. 2022, doi: 10.24912/je.v27i03.871.

[4] E. Nurkhanifah and S. Arifin, "Analisis Dampak Menurunnya Daya Beli Di Lingkungan Masyarakat Indonesia Akibat Inflasi," *J. Sahmiyya*, vol. 2, no. 1, pp. 240–248, 2023.

[5] Trimono Trimono, Aviolla Terza Damaliana, and Irma Amanda Putri, "Modelling of Return of S&P 500 Using the Non Linear Generalized Autoregressive Conditional Heteroscedasticity (NGARCH) Model," in *Nusantara Science and Technology Proceedings*, Galaxy Science, May 2024. doi: 10.11594/nstp.2024.4110.

[6] M. M. Al Haromainy, D. A. Prasetya, and A. P. Sari, "Improving Performance of RNN-Based Models With Genetic Algorithm Optimization For Time Series Data," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 16–24, Jun. 2023, doi: 10.38043/tiers.v4i1.4326.

[7] E. N. Aidoo, R. T. Ampofo, G. E. Awashie, S. K. Appiah, and A. O. Adebanji, "Modelling Covid-19 Incidence in the African Sub-Region Using Smooth Transition Autoregressive Model," *Model. Earth Syst. Environ.*, vol. 8, no. 1, pp. 961–966, Mar. 2022, doi: 10.1007/s40808-021-01136-1.

[8] C. T. Krisanti, N. Herawati, A. Sutrisno, and N. Nusyirwan, "Logistic Smooth Transition Autoregressive (LSTAR) and Exponential Smooth Transition Autoregressive (ESTAR) Methods in Predicting the Exchange Rate of Farmers in Lampung Province, Indonesia," *Int. J. Multidiscip. Res. Anal.*, vol. 7, no. 07, pp. 3209–3215, Jul. 2024, doi: 10.47191/ijmra/v7-i07-18.

[9] M. Sui, C. Zhang, L. Zhou, S. Liao, and C. Wei, "An Ensemble Approach to Stock Price Prediction Using Deep Learning and Time Series Models," Sep. 26, 2024, *Business, Economics and Management*. doi: 10.20944/preprints202409.2077.v1.

[10] D. Sulistiowati, M. S. Syahrul, M. S. Syahrul, and I. Rina, "Forecasting PT Triputra Agro Persada Tbk (TAPG) Share Prices Using Multivariate Time Series Analysis," *J Stat. J. Ilm. Teori Dan Apl. Stat.*, vol. 16, no. 2, pp. 594–605, Dec. 2023, doi: 10.36456/jstat.vol16.no2.a8344.

[11] M. I. Rizki and T. A. Taqiyyuddin, "Penerapan Model SARIMA untuk Memprediksi Tingkat Inflasi di Indonesia," *J. Sains Mat. Dan Stat.*, vol. 7, no. 2, pp. 62–72, Aug. 2021, doi: 10.24014/jsms.v7i2.13168.

[12] A. Khoerunnisa, I. M. Nur, and P. R. Arum, "Metode Markov Switching Autoregressive (MSAR) untuk Peramalan Indeks Saham Syariah Indonesia (ISSI)," *Pros. Semin. UNIMUS*, vol. 5, pp. 608–623, 2022.

[13] N. Salsabila and A. Oktaviarina, "Peramalan PDRB Di Jawa Timur Menggunakan Model Arimax Dengan Variabel Eksogen Ekspor-Impor," *Math Unesa J. Ilm. Mat.*, vol. 12, no. 1, pp. 208–218, 2024.

[14] D. Van Dijk, T. Terasvirta, and P. H. Franses, "Smooth Transition Autoregressive Models - A Survey of Recent Developments," *Econom. Inst. Res. Rep.*, vol. EI2000-23/A, 2000.

[15] M. Odelia, D. A. I. Maruddani, and H. Yasin, "Peramalan Harga Saham Dengan Metode Logistic Smooth Transition Autoregressive (LSTAR) (Studi Kasus Pada Harga Saham Mingguan Pt. Bank Mandiri Tbk Periode 03 Januari 2011 Sampai 24 Desember 2018)," *J. Gaussian*, vol. 9, no. 4, pp. 391–401, Dec. 2020, doi: 10.14710/j.gauss.v9i4.29403.

[16] T. Terasvirta, "Spesification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *J. Am. Stat. Assoc.*, vol. 89, no. 425, 1994.

[17] S. Lestari, D. Kusnandar, and S. Aprizkiyandari, "Pemodelan Data Return Saham Menggunakan Metode Smooth Transition Autoregressive," *Bul. Ilm. Math Stat Dan Ter. Bimaster*, vol. 10, no. 4, pp. 455–462, 2021.

[18] G.- Mardiatmoko, "Pentingnya Uji Asumsi Klasik Pada Analisis Regresi Linier Berganda," *BAREKENG J. Ilmu Mat. Dan Terap.*, vol. 14, no. 3, pp. 333–342, Oct. 2020, doi: 10.30598/barekengvol14iss3pp333-342.

[19] M. I. Rizki, T. Ammar, F. Fitriyani, and S. Fasya, "Peramalan Indeks Harga Saham PT Verena Multi Finance Tbk Dengan Metode Pemodelan ARIMA Dan ARCH-GARCH," *J Stat. J. Ilm. Teori Dan Apl. Stat.*, vol. 14, no. 1, pp. 11–23, Jul. 2021, doi: 10.36456/jstat.vol14.no1.a3774.

[20] F. Z. Indri and G. H. Putra, "Pengaruh Ukuran Perusahaan dan Konsentrasi Pasar Terhadap Kualitas Laporan Keuangan pada Perusahaan Sektor Industri Barang Konsumsi yang Terdaftar di Bursa Efek Indonesia pada Tahun 2016-2020," *J. Ilmu Manaj. Ekon. Dan Kewirausahaan*, vol. 2, no. 2, pp. 236–252, Jun. 2022, doi: 10.55606/jimek.v2i2.242.

[21] A. Q. Sari, Y. L. Sukestiyarno, and A. Agoestanto, "Batasan Prasyarat Uji Normalitas Dan Uji Homogenitas Pada Model Regresi Linear," *Unnes J. Math.*, vol. 6, no. 2, pp. 168–177, 2017.
[22] F. N. Hayati, D. Nurlaily, and E. Pusporani, "Peramalan Data Ekspor Non Migas Provinsi Kalimantan Timur Menggunakan Univariate Time Series," *J Stat.*, vol. 14, no. 2, pp. 59–66, Jan. 2022, doi: 10.36456/jstat.vol14.no2.a3858.
[23] E. Purnaningrum and M. Athoillah, "SVM Approach for Forecasting International Tourism Arrival in East Java," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1863/1/012060.
[24] F. Roshafara, "Forecasting Average Rice Prices at Milling Level According to Quality Using Support Vector Regression," *J Stat. J. Ilm. Teori Dan Apl. Stat.*, vol. 17, no. 1, pp. 664–671, Jul. 2024, doi: 10.36456/jstat.vol17.no1.a9245.
[25] T. Trimono, A. Sonhaji, and U. Mukhaiyar, "Forecasting Farmer Exchange Rate in Central Java Province Using Vector Integrated Moving Average," *Media Stat.*, vol. 13, no. 2, pp. 182–193, Dec. 2020, doi: 10.14710/medstat.13.2.182-193.
[26] A. T. Damaliana, A. Muhaimin, and D. A. Prasetya, "Forecasting the Occupancy Rate of Star Hotels in Bali Using the XGBoost and SVR Methods," *J. Stat. Univ. Muhammadiyah Semarang*, vol. 12, no. 1, pp. 24–33, 2024.

Geographically Weighted Negative Binomial Regression (GWNBR) Modeling In Infant Mortality Rate Cases In South Sulawesi

Nurul Fajril Ilmah(1), Bobby Poerwanto(2), Hardianti Hafid(3)

[1,2,3]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,

Universitas Negeri Makassar

Kampus UNM Parangtambung, Jalan Daeng Tata Makassar

e-mail: nurulilmahfajril@gmail.com(1), bobby_poerwanto@unm.ac.id(2) ,
hardiantihf@unm.ac.id(3)

ABSTRAK

Geographically Weighted Negative Binomial Regression (GWNBR) adalah salah satu metode untuk memodelkan data cacah yang mengalami overdispersi dan mempunyai heterogenitas spasial. Sulawesi Selatan merupakan salah satu provinsi yang mengalami peningkatan pada kasus angka kematian bayi. Oleh karena itu, penelitian ini bertujuan memperoleh model yang lebih baik dalam memetakan faktor-faktor yang mempengaruhi kasus angka kematian bayi di Provinsi Sulawesi Selatan. Metode yang digunakan pada penelitian ini adalah GWNBR dengan menggunakan *Adaptive Tricube Kernel* sebagai pembobot. Hasil penelitian menunjukkan bahwa model GWNBR dengan pembobot *Adaptive Tricube Kernel* menghasilkan nilai AIC terkecil yaitu sebesar 223,4447 sehingga lebih efektif digunakan dalam memodelkan kasus angka kematian bayi di Provinsi Sulawesi Selatan. Variabel-variabel yang berpengaruh signifikan terhadap kasus angka kematian bayi adalah X_1 (Persentase ASI Eksklusif), X_2 (Persentase Inisiasi Menyusu Dini), X_3 (Cakupan Kunjungan Bayi Lengkap), X_4 (Persentase Pemberian Vitamin A), X_5 (Jumlah Puskesmas), X_6 (Persentase Bayi Berat Lahir Rendah), X_7 (Cakupan Persalinan di Fasilitasi Pelayanan Kesehatan), dan X_8 (Pemberian Tablet Tambah Darah pada Ibu Hamil).

Kata kunci: GWNBR, Adaptive Tricube Kernel, Kematian Bayi

ABSTRACT

Geographically Weighted Negative Binomial Regression (GWNBR) is a method used to model count data that exhibit overdispersion and spatial heterogeneity. South Sulawesi is one of the provinces experiencing an increase in infant mortality cases. Therefore, this study aims to obtain a better model for mapping the factors that influence infant mortality cases in South Sulawesi Province. The method used in this study is GWNBR with an Adaptive Tricube Kernel as the weighting function. The results show that the GWNBR model with Adaptive Tricube Kernel weighting produces the smallest AIC value, which is 223.4447, making it more effective for modeling infant mortality cases in South Sulawesi Province. The variables significantly affecting infant mortality cases include X_1 (Percentage of Exclusive Breastfeeding), X_2 (Percentage of Early Initiation of Breastfeeding), X_3 (Complete Baby Visit Coverage), X_4 (Percentage of Vitamin A Supplementation), X_5 (Number of Community Health Centers), X_6 (Percentage of Low Birth Weight Babies), X_7 (Delivery Coverage in Health Service Facilities), and X_8 (Iron Tablet Supplementation to Pregnant Women).

***Keywords**: GWNBR, Adaptive Tricube Kernel, Infant Mortality*

INTRODUCTION

Poisson regression is used to analyze the reliationship between predictor variables and data that follow a Poisson distribution, assuming equidispersion, where the mean is equal to the variance. If the variance exceeds the mean (overdispersion), the Poisson model is no longer appropriate [1], [2]. One solution is to use negative binomial regression, which introduces a dispersion parameter to handle this condition[3]. Additionally, if spatial effects are present in the data, a spatial effect test is required. Spatial effects occur when observations are dependent on neighboring areas. They are divided into spatial dependence, which is handled with an area-based approach, and spatial heterogeneity, which uses a point-based approach. These effects are represented by location coordinates or weighting [4].

One of the analytical tools to handle data cases by considering spatial effects is the Geographically Weighted Regression (GWR) model [5]. The Geographically Weighted Negative Binomial Regression (GWNBR) model is applied when modeling discrete data that exhibit overdispersion and spatial heterogeneity [6].

The infant mortality rate (IMR) is influenced by geographical, socio-cultural, and economic factors that vary by region, leading to spatial heterogeneity. According to Tobler's First Law of Geography, as explained by [7], everything is related, but nearby things are more strongly related than distant ones, which is evident in the clustering of infant mortality in certain areas.

Globally, the IMR has been declining, with a rate of approximately 27 per 1,000 live births in 2020 [8]. In Indonesia, it decreased from 17.6 per 1,000 live births in 2020 to 16.9 in 2022. However, it still does not meet the target standards for mortality reduction[9]. In South Sulawesi, the IMR increased from 5 per 1,000 live births in 2020 to 8 per 1,000 live births in 2022, signaling a need for more effective efforts to reduce infant mortality[10].

Previous studies on the GWNBR method have been conducted by several researchers. Pratiwi, Pramoedyo, Astutik, Astuti and Fauwziyah [11] examined the factors influencing stunting in Malang Regency using Geographically Weighted Negative Binomial Regression (GWNBR). The results showed that five variables significantly affected stunting: Access to durable sanitary latrines, availability of integrated health posts, exclusive breastfeeding, population density, and community empowerment. Based on the smallest AIC criterion, GWNBR was found to be the most suitable method for modeling stunting cases. Another study by Delvia, Mustafid and Yasin [12] addressed overdispersion in poverty rates using GWNBR. The modeling results indicated that the most influential factors were the unemployment rate, economic growth, and the percentage of households occupying non-owned housing. Additionally, Rais and Haris [13] investigated the factors influencing pneumonia in children under five in Central Sulawesi Province. The modeling results showed that the significant factors included exclusive breastfeeding, the proportion of children receiving complete basic immunization, and the percentage of children receiving Vitamin A. and the coverage of early childhood health services. Although the GWNBR method has been used in previous studies, its application in analyzing the factors influencing infant mortality in South Sulawesi has not been conducted before. Therefore, the researcher will use the GWNBR approach to identify the factors affecting infant mortality in South Sulawesi Province.

METHOD

Data Source

The data used in this study is secondary data, specifically infant mortality rate (IMR) records in South Sulawesi, obtained from the South Sulawesi Provincial Health Office in 2023.

Research Variables

This study involves two types of variables: response variables (Y) and predictor variables (X). The response variable used is the infant mortality rate (Y), while the predictor variables include factors that are believed to influence the infant mortality rate in 24 districts/cities in South Sulawesi Province. Below is a further explanation of the predictor variables used in this study.

X_1 = Percentage of Babies Given Exclusive Breastfeeding
X_2 = Percentage Early Initiation of Breastfeeding
X_3 = Coverage of Complete Baby Visits
X_4 = Percentage of Vitamin A Supplementation
X_5 = Number of Community Health Centers
X_6 = Percentage of Low Birth Weight
X_7 = Coverage of Deliveries in Health Care Facilities
X_8 = Percentage of Pregnant Women Consuming Iron Supplement Tablets

Research Produce

The procedures to be followed in this study are as follows:

1. Collecting various data sources and information that will be used in the research.
2. Recapping the infant mortality data in South Sulawesi Province. At this stage, the data will be sourced from the South Sulawesi Provincial Health Office.
3. Processing the data using R-Studio software with the Geographically Weighted Negative Binomial Regression (GWNBR) method.
4. Drawing conclusions.
5. Compiling the research report

Data Analysis Technique

The data analysis techniques commonly used in this study are as follows:

1. Describing the Data
   a. Characteristics of the infant mortality cases in South Sulawesi in 2023 and the factors that are suspected to influence them.
   b. Multicollinearity testing by examining the VIF values[14].

$$VIF = \frac{1}{1 - R_j^2}$$

Where:
R_j^2 : The Coefficient of determination between variable X_j and other predictor variables.

2. Poisson Regression Modeling
   a. Parameter estimation of the Poisson regression model.[15].

$$L(\beta) = \frac{\exp\left(-\sum_{i=1}^{n} \exp\left(x_i^T \beta\right)\right)\left(\exp \sum_{i=1}^{n} y_i x_i^T \beta\right))}{\prod_{i=1}^{n} y_i!}$$

b. Testing the Poisson regression model parameters for significance test on the model [16]
   - Simultaneous test

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

   Where:
   $\ln(\hat{\omega})$: The likelihood function for the model excluding predictor variables.
   $\ln(\hat{\Omega})$: Likelihood function for the model without excluding predictor variables.
   - partial tests.

$$Z_{hit} = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$$

   Where:
   $\hat{\beta}_k$: Coefficient of the k-th predictor variable model
   $se(\hat{\beta}_k)$: Standard error of the maximum likelihood estimate

3. Overdispersion Testing [16]
4. Negative Binomial Regression Modeling
   a. Parameter estimation of the Negative Binomial regression model.[17].

$$L(\beta, \theta) = \prod_{i=1}^{n}\left[\frac{\Gamma(\mathrm{y_i} + \theta^{-1})}{\Gamma(\theta^{-1})\Gamma(y_i + 1)}\left(\frac{1}{1 + \theta\mu_i}\right)^{\frac{1}{\theta_i}}\left(\frac{\theta\mu_i}{1 + \theta\mu_i}\right)^{y_i}\right]$$

   b. Conducting significance testing of the negative binomial regression model parameters both simultaneously and partially [15]
5. Testing spatial effects using spatial heterogeneity with the Breusch-Pagan test [18].

$$BP = \left(\frac{1}{2}\right) f^T Z(Z^T Z)^{-1} Z^T f \sim \chi_p^2$$

   where, Vector elements f is $f_i = \frac{\varepsilon_i^2}{\sigma^2} - 1$
6. GWNBR Modeling
   a. Calculating The Euclidean distance between observation points based on their geographical coordinates. [19].

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

   b. Obtaining the optimal bandwidth for each observation location using Cross Validation (CV)[20].

$$CV = \sum_{i=1}^{n} (y_i - y_{\neq 1}(h))^2$$

   c. Calculating the weight matrix using the Adaptive Tricube Kernel[21].

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h}\right)^3\right]^3 & , for\ d_{ij} \le h \\ 0 & , others \end{cases}$$

Where:

d_{ij} : Euclidean distance between locations(u_i, v_j) to the location (u_i, v_j)

h : Bandwidth

d. Estimating the parameters [15] and structuring them into the model

$$L(\beta_{u_i v_i}, \theta_i | y_i, x_i) = \prod_{i=1}^{n} \left(\prod_{r=0}^{y_i - 1} \left(r + \frac{1}{\theta_i}\right) \right) \frac{1}{(y_i!)} \left(\frac{1}{1 + \theta_i \mu_i}\right)^{\frac{1}{\theta}} \left(\frac{\theta_i \mu_i}{1 + \theta_i \mu_i}\right)^{y_i}$$

e. Testing the GWNBR model parameters for significance test on the model.

$$Z_{hit} = \frac{\hat{\beta}_j(u_i v_i)}{se\left(\hat{\beta}_j(u_i, v_i)\right)}$$

$\hat{\beta}_j$: Coefficient of the j-th predictor variable model

$se(\hat{\beta}_j)$: Standard error of the maximum likelihood estimate

7. Model Interpretation

RESULT AND DISCUSSION

Descriptive Analysis

Descriptive statistics illustrate the distribution of infant mortality cases (IMR) in South Sulawesi in 2023. The total number of IMR cases was 1,431, with an average of 59.62 cases. The minimum number of cases was 15, the maximum was 220, and the standard deviation was 41.68. The descriptive statistics of the predictor variables, including the mean, maximum value, minimum value, and standard deviation, are presented in Table 1.

Table 1. Descriptive Statistics of Predictor Variables

Variable	Mean	Min	Maks	Stdev
X_1	69,55	37,75	85,76	10,77
X_2	86,35	66,63	97,85	9,53
X_3	108,15	68,99	181,64	24,77
X_4	92,64	82,48	98,84	4,17
X_5	19,67	8,00	47,00	8,58
X_6	6,65	3,370	10,150	1,75
X_7	91,80	65,63	121,97	15,09
X_8	56,74	28,09	98,81	22,02

Multicollinearity Test

The VIF values for each predictor variable are presented below.

Table 2. Multicollinearity

Variable	VIF
Percentage of Babies Given Exclusive Breastfeeding (X_1)	1,733
Early Initiation of Breastfeeding (X_2)	1,845
Coverage of Complete Baby Visits (X_3)	1,900
Percentage of Vitamin A Supplementation (X_4)	1,293
Number of Community Health Centers (X_5)	1,458
Percentage of Low Birth Weight (X_6)	1,882
Coverage of Deliveries in Health Care Facilities (X_7)	1,677
Percentage of Pregnant Women Consuming Iron Supplement Tablets (X_8)	1,319

Poisson Regression Modeling

The results obtained can be observed in Table 3.

Table 3. Estimation and Testing of Poisson Regression Model Parameters

	$Estimation$	$Std.Error$	Z_{Score}	$P-Value$	Significant
(Intercept)	4,678	0,788	5,937	$2{,}90 \times 10^{-9}$	Significant
X_1	0,023	0,004	6,425	$1{,}32 \times 10^{-10}$	Significant
X_2	$-$ 0,015	0,004	$-4{,}322$	$1{,}55 \times 10^{-5}$	Significant
X_3	0,002	0,002	1,539	0,124	Not Significant
X_4	$-0{,}027$	0,008	$-3{,}573$	0,0003	Significant
X_5	0,035	0,003	10,966	$< 2 \times 10^{-16}$	Significant
X_6	$-0{,}042$	0,021	$-2{,}000$	0,045	Significant
X_7	0,008	0,002	3,652	0,0002	Significant
X_8	0,001	0,001	0,534	0,594	Not Significant
Devians:140,60				Df :15	
AIC: 296,769					

Simultaneous testing of Poisson regression parameters with the following hypotheses:

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

$H_1 =$ At least one $\beta_k \neq 0; k = 1,2, \ldots, p$

Based on the test statistic values in Table 4, the decision is to reject H_0 at a significance level of 0,05. The deviance value $D(\hat{B}) = 140{,}60 > \boldsymbol{x^2_{(0{,}05,9)}} = 16{,}919$ This indicates that at least one predictor variable significantly influences the response variable. The next step is to conduct individual parameter testing. The hypotheses used in this test are as follows:

$H_0: \beta_k = 0$

$H_1: \beta_k \neq 0$

Based on the results of the partial testing at a 5% significance level, it was found that all variables, except X_3 and X_8 have $P - Value < \alpha = 5\%$. This means that the variables X_1, X_2, X_4, X_5, X_6 dan X_7 significantly influence the model. Therefore, the Poisson regression model formed is as follows:

$$\hat{\mu}_i = \textbf{exp}\ (4{,}678 + 0{,}023\mathrm{X}_1 - \ 0{,}015\mathrm{X}_2 - \ 0{,}027\mathrm{X}_4 + 0{,}035\mathrm{X}_5 - 0{,}042\mathrm{X}_6 + 0{,}008X_7)$$

Overdispersion

The results of the overdispersion test are presented in the Table 4 below.

Table 4. Dispersion Parameter of Poisson Regression

$\boldsymbol{Devians}$	$\boldsymbol{Df}$	$\boldsymbol{\theta}$
140,60	15	9,373

Based on the overdispersion test, it was found that the deviance divided by its degrees of freedom is greater than 1. Therefore, it can be concluded that the data Fails to satisfy the equidispersion assumption and experiences overdispersion.

Negative Binomial Regression Modeling

The results obtained can be seen in Table 5.

Table 5. Estimasi dan Testing of Negative Binomial Regression

	$\boldsymbol{Estimation}$	$\boldsymbol{Std.Error}$	$\boldsymbol{Z_{Score}}$	$\boldsymbol{P-Value}$	**Significant**
(Intercept)	4,784	1,933	2,475	0,013	Significant
X_1	0,022	0,009	2,527	0,011	Significant
X_2	−0,015	0,009	−1,505	0,132	Not Significant
X_3	0,001	0,004	0,223	0,823	Not Significant
X_4	−0,028	0,019	−1,475	0,140	Not Significant
X_5	0,031	0,009	3,176	0,001	Significant
X_6	−0,024	0,055	−0,435	0,663	Not Significant
X_7	0,011	0,006	1,840	0,066	Not Significant
X_8	0,0002	0,004	0,076	0,939	Not Significant
Devians:24,695				Df:15	
AIC: 225,701					

Simultaneous testing of Negative Binomial regression parameters with the following hypotheses:

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

$H_1 =$ At least one $\beta_k \neq 0; k = 1,2, \ldots, p$

Based on Table 5, H₀ is rejected at $\alpha = 5\%$ since $D(\hat{B})$ = 24,695 > 16,919, indicating that at least one predictor variable significantly influences the response variable, thus partial testing is conducted. The next step is to conduct individual parameter testing. The hypotheses used in this test are as follows:

$H_0: \beta_k = 0$

$H_1: \beta_k \neq 0$

Based on the results of the partial test at a 5% significance level, it was found that only two variables, X_1 dan X_5, have $P - Value < \alpha = 5\%$, meaning that these variables significantly influence the response variable. Hence, the resulting negative binomial regression model is as follows:

$$\hat{\mu}_i = \mathbf{exp}\ (4{,}784 + 0{,}022\mathrm{X}_1 - 0{,}031\mathrm{X}_5)$$

Testing Spatial Heterogeneity

The test statistics for heterogeneity using the Breusch-Pagan test, along with the results, are shown in Table 6.

Table 6. Breusch Pagan Test

Breusch Pagan Test	$x^2_{(0,05,8)}$	*P-Value*	Description
17,77	15,507	0,023	Reject H_0

The hypotheses for this test are as follows:

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

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

Based on Table 7, Ho is rejected at $\alpha = 5\%$ since $BP = 17{,}77 > 15{,}507$ and $P - Value = 0{,}023 < 0{,}05$, indicating spatial heterogeneity across observation points.

GWNBR Modeling

The GWNBR modeling is analyzed, but first, the Euclidean distance between the observed regions is calculated. The obtained Euclidean distances are used to find the bandwidth values and weights using the Adaptive Tricube Kernel. Subsequently, GWNBR modeling is conducted, which includes testing the model parameters for GWNBR.

a. Simultaneous Test

The hypotheses for the simultaneous test are as follows:

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

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

Table 7. Simultaneous Significance Test of GWNBR Model Parameters

$D(\widehat{B})$	$x^2_{(0,05,9)}$	Description
24,695	16,919	Reject H_0

Based on the simultaneous testing results at a 5% significance level, the value of nilai $D(\hat{B}) = 294770{,}7 > x^2_{(0,05,9)} = 16{,}919$, leading to the rejection of H_0. This indicates that at least one predictor variable significantly influences the model

b. Partial Test

The hypotheses for the partial test are as follows:

$H_0: \beta_j(u_i, v_i) = 0$

$H_1: \beta_j(u_i, v_i) \neq 0$

Partial test results show that at a 5% significance level, a parameter is significant if $Z_{\left(\frac{0.05}{2}\right)} > 1{,}96$.

Significant variables vary across regions.

Table 8. Grouping of Parameter Significance Results

Number	District/City	Significant Variable
1	Barru, Bone, Bulukumba, Enrekang, Gowa, Jeneponto, Makassar, Maros, Pangkep, Pare-pare,	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8$

	Pinrang, Selayar, Sidrap, Sinjai, Soppeng, Takalar, dan Bantaeng	
2	East Luwu , North Luwu, Luwu, Palopo, North Toraja, dan Toraja	$X_1, X_2, X_3, X_4, X_5, X_7, X_8$
3	Wajo	$X_1, X_3, X_4, X_5, X_6, X_7, X_8$

Table 8 presents the classification of districts/cities based on the significance of predictor variables influencing infant mortality, showing three variations of influencing factor combinations across regions. As seen in Figure 1, the distribution map of the formed clusters exhibits a closely related pattern.

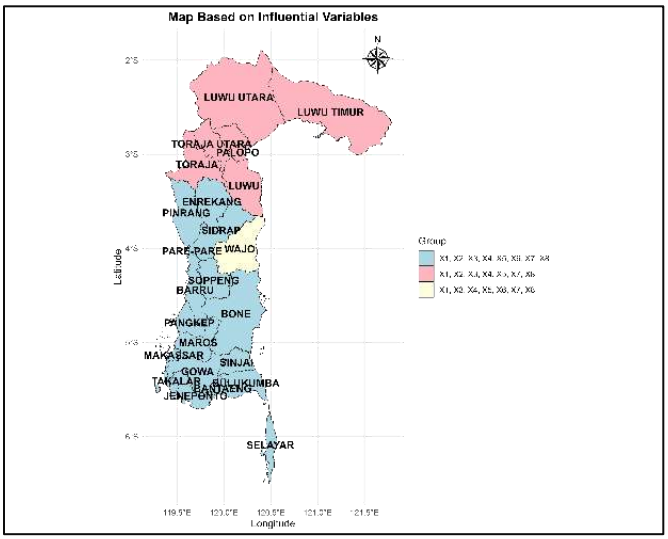


Figure 1. Map of District/City Grouping Distribution in South Sulawesi Province Based on Influential Variables

Model Interpretation

According to Table 8, one of the GWNBR models with Adaptive Tricube Kernel weighting for districts/cities in South Sulawesi Province is as follows :

$$\hat{\mu}_{Makassar} = \exp(4{,}350 - 0{,}273X_1 - 0{,}420X_2 + 0{,}013X_3 - 1{,}050X_4 - 0{,}212X_5 + 1{,}238X_6 + 0{,}152X_7 - 0{,}160X_8)$$

Based on the GWNBR model in Makassar City, it can be inferred that for each one-unit increase in the percentage of Babies Given Exclusive Breastfeeding (X_1) reduces infant mortality by exp(-0,273) = 0,761 times, this is consistent with research in Susanto and Mustikawati [6]. Similarly, early initiation of breastfeeding (X_2) decreases infant mortality by exp (-0,420) = 0,657 times, aligning with the fingding in Husnah, Sakdiah and Andayani [22]. Conversely, an increase in complete infant visit coverage (X_3) actually increases infant mortality by exp (0,013) = 1,013 times. This result is less realistic, as an increase in complete infant visit coverage should reduce leads to a rise in infant mortality. This occurs due to poor quality of services during visits, incomplete examinations, or inadequate follow-up. Furthermore, each increase in vitamin A supplementation (X_4) reduces leads to a rise in infant mortality by exp(-1,050) = 0,350 times, while the percentage of low birth weight infants (X_6) leads to an increase in infant mortality by

exp(1,238) = 3,449 times this finding is consistent with the study in Ismail, Utami, and Haris[23]. The addition of community health centers (X_5) reduces the number of infant deaths by exp(0,212) = 1,236 times this finding aligns with the research in Cabral, Udus, Jamlean, Pramesti and Anuraga [24]. Facility-based delivery coverage (X_7) leads to a rise in infant mortality by exp(0,152) = 1,164 times. This is because not all healthcare facilities have sufficient personnel or equipment, resulting in suboptimal delivery care services. Percentage of Pregnant Women Consuming Iron Supplement Tablets (X_8) reduces infant mortality by exp(-0,160) = 0.852 times by lowering the risk of maternal anemia. This finding is in accordance with study in Maryam and Muslimah [25].

CONCLUSION

The GWNBR model with the Adaptive Tricube Kernel weighting function developed to analyze factors affecting infant mortality rates in South Sulawesi is as follows:

$$\hat{\mu}_{Makassar} = \exp(4{,}350 - 0{,}273X_1 - 0{,}420X_2 + 0{,}013X_3 - 1{,}050X_4 - 0{,}212X_5 + 1{,}238X_6 + 0{,}152X_7 - 0{,}160X_8)$$

The factors influencing infant mortality include exclusive breastfeeding, early initiation of breastfeeding, complete baby visit coverage, vitamin A supplementation, number of health centers, low birth weight, facility-based deliveries, and iron tablet consumption by pregnant women. This model can be used for health program planning and it is recommended to compare various weighting functions to obtain the best model.

REFERENCES

[1] A. C. Cameron and P. K. Trivedi, *Regression Analysis Of Count Data*, First. Cambridge: Cambridge University Press, 1998.

[2] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan pengujian hipotesis statistika dasar dengan software r," *BUDIMAS: Jurnal Pengabdian Masyarakat*, vol. 3, no. 2, pp. 327–334, 2021.

[3] R. F. Ramadhan and R. Kurniawan, "Pemodelan Data Kematian Bayi Dengan Geographically Weighted Negative Binomial Regression," *Media Statistika*, vol. 9, no. 2, p. 95, 2017, doi: 10.14710/medstat.9.2.95-106.

[4] M. F. Itsnaini, Sugiman, and Sunarmi, "Estimasi Parameter Model Regresi Spasial Dengan Metode Geographically Weighted Poisson Regression," *UNNES Journal of Mathematics*, vol. 8, no. 2, pp. 21–31, 2019.

[5] N. Lutfiani and S. Mariani, "Pemodelan Geographically Weighted Regression (GWR) dengan Fungsi Pembobot Kernel Gaussian dan Bi-square," *UNNES Journal of Mathematics*, vol. 5, no. 1, pp. 82–91, 2019.

[6] T. Susanto and E. Mustikawati P. H, "Pemodelan Geographically Weighted Negative Binomial Regression (GWNBR) untuk Kasus Kematian Bayi Di Provinsi Jawa Tengah," 2020.

[7] O. Schanberger and C. A. Gotway, "Handbook of Applied Economic Statistics," in *Chapman & Hall/CRC*, B. P. Carlin, C. Chatfield, M. Tenner, and J. Zidek, Eds., New York: Chapman & Hall/CRC, 2005, ch. Statistic, p. 26.

[8] The World Bank, "Mortality rate, infant (per 1,000 live births)." Accessed: Mar. 31, 2024. [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.IMRT.IN

[9] I. Aliska, A. S. E. Putri, and M. Ramadani, "Determinan Kematian Bayi Ditinjau dari Perilaku Kesehatan Ibu : Tinjauan Literatur," *Jurnal Epidemiologi Kesehatan Indonesia*, vol. 7, no. 1, p. 25, 2023, doi: 10.7454/epidkes.v7i1.6689.

[10] P. S. S. Dinas Kesehatan, *Laporan Kinerja Tahun 2022*. Sulawesi Selatan, 2022.
[11] E. Pratiwi, H. Pramoedyo, S. Astutik, and F. Fauwziyah, "Modeling Geographically Weighted Negative Binomial Regression (GWNBR) on Stunting Incidence in Malang Regency," *Jurnal Matematika, Statistika dan Komputasi*, vol. 19, no. 1, pp. 163–171, 2022, doi: 10.20956/j.v19i1.21757.
[12] N. Delvia, M. Mustafid, and H. Yasin, "Geographically Weighted Negative Binomial Regression Untuk Menangani Overdispersi Pada Jumlah Penduduk Miskin," *Jurnal Gaussian*, vol. 10, no. 4, pp. 532–543, 2021, doi: 10.14710/j.gauss.v10i4.33106.
[13] Z. Rais and A. S. Haris, "GEOGRAPHICALLY WEIGHTED NEGATIVE BINOMIAL REGRESSION (GWNBR) IN MODELING THE RISK FACTORS OF PNEUMONIA DISEASE AMONG TODDLERS IN THE CENTRAL," vol. 5, no. 3, pp. 118–131, 2023, doi: 10.35580/variansiunm151.
[14] R. R. Hocking, *Methods and Application of Linear Models*. New York: John Wiley & Sons, Ltd, 1996.
[15] J. L. Fleiss, B. Levin, and C. P. Myunghee, *Statistical Methods for Rates and Proportions*, Third. New York: Wiley, 2003.
[16] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman & Hall, 1989. doi: 10.1007/978-1-4899-3242-6.
[17] A. Ricardo and T. Carvalho, *Geographically Weighted Negative Binomial Regression – Incorporating Overdispersion*. New York: Springer Science, 2013.
[18] L. Anselin and A. K. Bera, "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics," in *Handbook of Applied Economic Statistics*, A. Ullah, Ed., CRC Press, 1998, pp. 237–290.
[19] C. Chasco, I. Garcia, and J. Vicens, "Modeling spatial variations in household disposable income with Geographically Weighted Regression," Munich Personal RePEc Archive, 1682, 2007.
[20] S. Kartika and G. Kholijah, "Penggunaan Metode Geograhically Weighted Regression (GWR) Untuk Mengestimasi Faktor Dominan yang Mempengaruhi Penduduk Miskin di Provinsi Jambi," *Journal of Matematics: Theory and Applications*, vol. 2, no. 2, pp. 37–45, 2020.
[21] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression*. England: John Wiley & Sons Ltd., 2002.
[22] Husnah, Sakdiah, and H. Andayani, "Dampak Inisiasi Menyusui Dini Terhadap Penurunan Angka Kematian Bayi," *Jurnal Kedokteran Nanggroe Medika*, vol. 1, no. 938, pp. 6–37, 2024.
[23] I. M. Ismail, T. W. Utami, and M. Al Haris, "Pemodelan Jumlah Kematian Bayi di Provinsi Jawa Barat dengan Pendekatan Geographically Weighted Negative Binomial Regression (GWNBR)," *Program Studi Statistik, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Muhammadiyah Semarang*, 2019.
[24] A. H. Cabral, M. Y. Udus, S. F. Jamlean, W. Pramesti, and G. Anuraga, "Pemodelan Faktor yang Mempengaruhi Angka Kematian Bayi di Jawa Timur dengan Menggunakan Geographically Weighted Regression," *Snhrp: Seminar Nasional Hasil Riset dan Pengabdian*, no. 2019: Seminar Nasional Hasil Riset dan Pengabdian Ke-II (SNHRP-II), pp. 37–49, 2019.
[25] S. Maryam and E. A. Muslimah, "Analisis Riwayat Tablet Tambah Darah pada Ibu Hamil dengan Anemia di Indonesia (Data RISKESDAS 2018)," *Jurnal Ilmiah Ilmu Kebidanan*, vol. Vol. 10, pp. 1–8, 2018.

Cluster Analysis Using the Ward Algorithm for Grouping Regency / City in Central Java Province Based on Poverty Indicators 2023

Agung Supriyono[(1)], Atika Nurani Ambarwati [(2)]

Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang.

Jl. Prof. Dr. Hamka Km 01 No. 17 Tambakaji Ngaliyan

e-mail: agungsupriyono2004@gmail.com[(1)], atika.nurani@gmail.com[(2)]

ABSTRAK

Kemiskinan masih menjadi isu utama di Provinsi Jawa Tengah, dengan tingkat kemiskinan yang tercatat sebesar 10,23% pada tahun 2023. Penanggulangan kemiskinan yang belum merata di berbagai wilayah turut menambah kompleksitas permasalahan ini. Penelitian ini bertujuan untuk mengklasifikasikan kabupaten/kota di Jawa Tengah berdasarkan indikator-indikator kemiskinan, dengan menggunakan pendekatan analisis klaster hierarki melalui algoritma Ward. Data yang digunakan merupakan data sekunder dari Badan Pusat Statistik (BPS), yang mencakup enam variabel yang memengaruhi kemiskinan. Salah satunya adalah persentase penduduk miskin (X1), yang merepresentasikan proporsi penduduk di bawah garis kemiskinan dan menjadi indikator penting dalam menilai kesejahteraan suatu daerah. Hasil penelitian menunjukkan bahwa penerapan analisis klaster hierarki melalui algoritma Ward menghasilkan empat klaster kabupaten/kota berdasarkan indikator kemiskinan tahun 2023, masing-masing dengan karakteristik yang berbeda. Klaster pertama terdiri dari daerah-daerah dengan tingkat kemiskinan tinggi serta keterbatasan dalam aspek pendidikan, kesehatan, dan infrastruktur ekonomi. Wilayah dalam klaster ini memerlukan perhatian khusus dalam penyusunan kebijakan. Berdasarkan karakteristik yang diperoleh dari masing-masing klaster, temuan ini dapat dimanfaatkan untuk merancang kebijakan pembangunan yang lebih terarah dan berbasis data, seperti pengalokasian anggaran, program penanggulangan kemiskinan, serta peningkatan akses terhadap layanan dasar sesuai dengan kondisi tiap klaster.

Kata kunci : Kemiskinan, Analisis Hierarki, Pendekatan Ward, Jawa Tengah

ABSTRACT

Poverty is still a major issue in Central Java Province, with the poverty rate recorded at 10.23% in 2023. Uneven poverty reduction in various regions adds to the complexity of this problem. This study aims to classify districts/cities in Central Java based on poverty indicators, using a hierarchical cluster analysis approach through Ward's algorithm. The data used is secondary data from the Central Bureau of Statistics (BPS), which includes six variables that affect poverty. One them is the percentage poor people (X1), which represents the proportion of the population below the poverty line and is an important indicator in assessing the welfare of a region. The results show that the application of hierarchical cluster analysis through Ward's algorithm produces four clusters of districts/cities based on poverty indicators in 2023, each with different characteristics. The first cluster consists regions with high poverty rates and limitations in education, health, and economic infrastructure. Regions in this cluster require special attention in policy formulation. Based on the characteristics obtained from each cluster, these findings can be used to design more targeted and data-based development policies, such as budget allocation, poverty reduction programs, and improving access to basic services according to the conditions each cluster.

Keywords: *Poverty, hierarchical analysis, Ward method, Central Java*

INTRODUCTION

Poverty is one of the challenges faced by various countries, especially in developing countries, namely Indonesia [1]. The problem of poverty has become a priority that must be addressed immediately and needs more attention. Because it jeopardizes the economic development and social stability of a country [2]. Poverty is a major challenge faced by developing countries such as Indonesia because it impacts economic development and social stability. According to BPS data in 2023, 9.36% or around 25.9 million Indonesians still live in poverty, struggling to access basic needs such as food, clean water, sanitation, housing, education, and health services. Java, as the most populous region, accounts for the largest number of poor people, at 13.94 million or more than half of the national total. Among the provinces in Java, Central Java recorded the highest poverty percentage at 10.23%, exceeding the national average. Several regions in the province, such as Kebumen, Brebes, and Wonosobo, are still classified as areas of extreme poverty, reflecting the welfare gap between urban and rural areas and the unevenness of poverty reduction efforts.

Some previous studies show that the percentage of poor people, population, open unemployment rate, expected years of schooling, per capita expenditure, and life expectancy have a significant effect on poverty [3],[4]. Therefore, this study uses these variables as variables to group districts/cities based on the factors that influence poverty. This grouping aims to make it easier for the government to design a more targeted poverty alleviation strategy. To facilitate decision making for the Central Java Provincial government in overcoming this poverty problem, of course, the right approach is needed to map the regions according to their characteristics.

Ward's approach was chosen in this study because it can reduce the number of error squares in the clustering process, thus creating clusters with high internal homogeneity [5]. Research conducted by [6] with the title "Analisis Cluster dengan Average Linkage Method dan Ward's Method pada Pengelompokan Kabupaten/Kota Di Provinsi Sumatera Utara Berdasarkan Indikator Indeks pembangunan Manusia Tahun 2022" stated that the ward's method algorithm proved superior because it has a lower Sum of Squared Errors (SSE) value, indicating that this method is able to form denser and more homogeneous clusters. In addition, evaluation of the dendrogram shows that the ward's method produces more uniform and consistent clusters. This approach aims to minimize the variance within clusters, Ward's method produces more homogeneous and compact groups, allowing the identification of areas with high poverty rates that also have limitations in education, health, and infrastructure. The results of this clustering are expected to provide a more structured picture of the condition of the region so that it can be used as a basis for formulating more targeted development policies [7].

ALGORITHMS

This research uses a quantitative Algorithm with a hierarchical cluster analysis Algorithm using the Ward Algorithm. The steps taken in this research include:

Data Source

The type of data used in this study are secondary data, which comes from the Central Java Provincial Statistics Agency Publication in 2023.

Research Variables

The variables used in this study only use predictor variables (X). Predictive variables are factors that have the potential to have an impact on Poverty in Central Java Province. The following are the variables:

Table 1. Risearch Variables

Variables	Describtions	Scale
X1	Percentage of Poor Population	Ratio
X2	Total Population	Ratio
X3	Open Unemployment Rate	Ratio
X4	Expected Years of Schooling	Ratio
X5	Per capita expenditure	Ratio
X6	Life Expectancy Rate	Ratio

Data Processing

The Algorithm used in this research uses the literature Algorithm or literature study with the following steps [8]:

1. Conduct a Descriptive Analysis of the Poverty Condition in Central Java Province in 2023.
2. Test assumptions in Cluster, In clustering, there are two assumptions that need to be met, namely:
   a. The sample represents the population, the sample used in cluster analysis must be able to represent the population to be explained, because this analysis is said to be good if the sample is representative. To find out that the sample can represent the population can be seen from the Kaiser Mayer Olkin (KMO) value, if the KMO value obtained is> 0.5 then the sample is sufficient to be analyzed [9].
   b. Multicollinearity, used to determine the existence of a linear relationship between independent variables. The VIF (Variance Inflation Factor) value can be used to determine the presence or absence of multicollinearity. If the VIF value < 10.00, then there are no symptoms of multicollinearity and vice versa [10], [11].

   $$VIF_i = \frac{1}{1-R_i^2} \quad (1)$$

   Description:
   R^2 = Koefisien determinan
   i = The i-th independent variable
3. Measuring the similarity between objects with Euclidean Distance, Euclid distance is used to measure the distance from the data object to the cluster center [12].

   $$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \quad (2)$$

   Description:
   d_{ij} = Euclid distance of i-th data object and j-th data object
   p = Number of parameters used
   x_{ik} = The i-th data object on the k-th variable
   x_{jk} = j-th data object on the k-th variable

4. The clustering process in the ward method uses Sum Square Error (SSE) to measure the homogeneity between two objects based on the least amount of squared error and also to measure the quality of the cluster [13].

   $SSE = \sum_{j=1}^{p}(\sum_{i=1}^{n} x_{ij}^2 - \frac{1}{n}(\sum_{i=1}^{n} x_{ij})^2)$ (3)

   Description:

   x_{ij} = The value of the j-th variable for the i-th object

   p = Number of variables measured

   n = Number of objects in the formed cluster
5. Determining the number of clusters, aims to group data into internally homogeneous clusters by minimizing variation or squared error within clusters [9].
6. Interpretation of cluster results, at the interpretation stage the average (centroid) is used. Through this centroid value, to explain the purpose or characteristics of the cluster.
7. Conclusion

Figure 1. Flow chart

RESULTS AND DISCUSSION

Descriptive Analysis

Descriptive analysis was conducted to determine the general description of Poverty conditions in Central Java Province and the characteristics of each indicator or variable used.

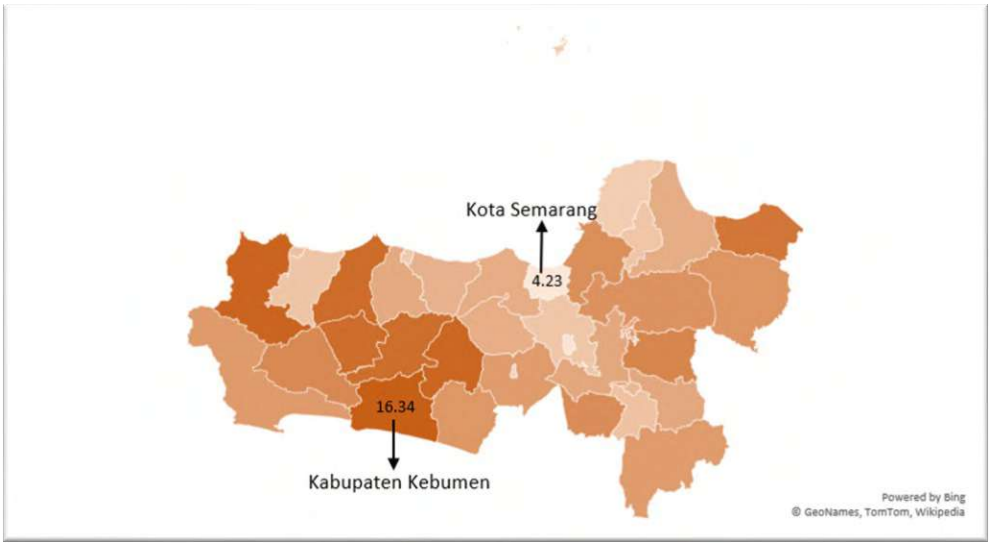


Figure 2. District/City Poverty Rates in Central Java 2023

It is also known that the lowest poverty condition is in Semarang City with a value of 4.23%, while the highest poverty condition is in Kebumen district with a value of 16.34%, which shows that the poverty rate in Central Java experiences significant inequality. This shows that poverty reduction has not been well-targeted in areas with high poverty rates.

Assumptions in Cluster Analysis

1. The sample is representative of the population
   To find out that the sample can represent the population can be seen from the Kaiser Mayer Olkin (KMO) value, if the KMO value obtained is > 0.5 then the assumption is fulfilled.

Table 2. Kaiser Mayer Olkin value

KMO and Bartlett's Test	
Kaiser-Meyer-Olkin	,698

   Based on table 2, the KMO value is 0.69 where the KMO value of 0.69 > 0.5, it can be concluded that the sample can represent the population and can be used for further analysis.
2. Multicollinearity
   The coefficient of determination of each variable is obtained by making the variable that you want to know the coefficient of determination as the dependent variable and the remaining variables as independent variables. If the Variance Inflation Factor (VIF) value is < 10.00, then the assumption is fulfilled.

Table 3. Varians Inflation Factor value

Variable	VIF value
Percentage of Poor Population	1,000
Total Population	1,608
Open Unemployment Rate	3,713
Expected Years of Schooling	2,424
Per capita expenditure	2,868
Life Expectancy Rate	3,678

Since the VIF value of each variable is less than 10.00 in the previous table, it can be said that there are no variables that show signs of multicollinearity.

Measuring similarity between objects with Euclidean Distance

In calculating the similarity of objects (districts), the Algorithm used is the Euclidean distance, there are 35 districts that will be calculated for similarity. Calculation of the Euclidean distance matrix using Equation (2). An example of calculating the distance between X1 and X2 is as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{35} (x_{ik} - x_{2k})^2} = 3{,}112$$

Using the same calculation, the distance between City 1 and City 3 and so on is also obtained. The smaller the distance value between two observations, the more similar the two observations are.

Ward Algorithm Cluster Process

The clustering process in this study groups regencies/cities in Central Java Province based on their characteristics using the ward Algorithm. The results of clustering using the ward Algorithm can be seen with a dendrogram diagram, as follows:

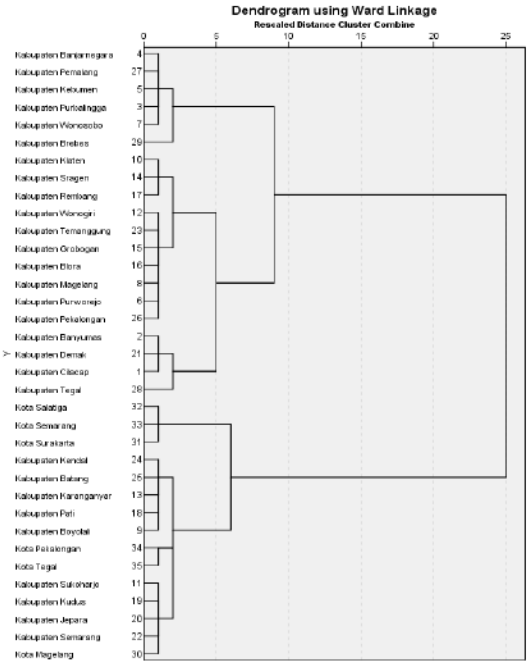


Figure 3. Dendrogram Diagram Results

Based on the dendrogram analysis, there are three possible numbers of clusters that can be formed, namely four, three, and two clusters. In the formation of four clusters, we get a first cluster consisting of 6 districts, a second cluster with 14 districts, a third cluster that includes 3 districts, and a fourth cluster consisting of 12 districts. If we group into three clusters, the first cluster will still consist of 6 districts, the second cluster of 14 districts, while the third cluster will include 15 districts. Finally, if only two clusters were formed, the first cluster would include 20 districts and the second cluster would consist of 15 districts.

Determining the Number of Clusters

In this study, the number of clusters is determined based on the Sum of Squared Errors (SSE) value. The SSE value shows how far each data in the group is from the center of the group. The smaller the SSE value in a cluster, the more homogeneous the objects in the cluster are, so the quality of the clustering is getting better.

Table 4. Sum Squared Error value

Cluster	SSE Value
1	3112610.36
2	13170017.30
3	321280.27
4	13936095.40
1	3112610.36
2	13170017.30
3	48419657.81
1	18952231.33
2	48419657.81

From table 4, it can be seen that the lowest Sum squared error cluster value of each cluster in the number of clusters formed is in the grouping with 4 clusters. So it can be concluded that in this study the object is classified with 4 groups. The members of each cluster formed are as follows:

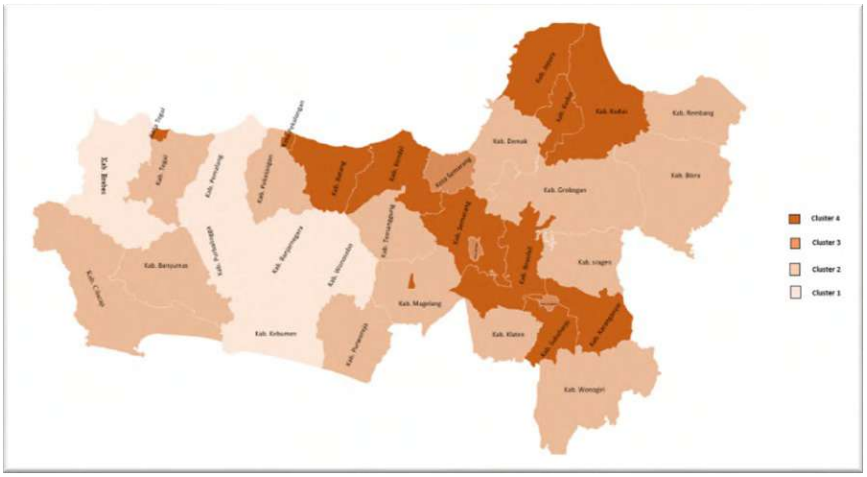


Figure 4. Cluster Interpretation Results

Interpretation of Cluster Visualization Results

Interpreting cluster results involves the centroid value, which is the average value of each variable on the objects in the cluster.

Table 5. Cluster Interpretation Result

Variable	Cluster 1	Cluster 2	Cluster 3	Clusster 4
X1	15,43	11,25	4,77	8,03
X2	1,31	1,21	0,8	0,83
X3	12,24	12,81	15,29	13,16
X4	6,24	4,48	5,04	4,57
X5	10,51	11,31	16,31	12,54
X6	72,94	74,92	77,82	76,08

Based on table 5, the characteristics (average) of each group formed on the Poverty indicator in Central Java in 2023 are obtained.

- Cluster 1 consists of areas with high poverty rates, high open unemployment, and relatively low expected years of schooling. These areas are generally rural areas with limited access to basic services and formal employment. The Central Java BPS report (2023) confirms that the development imbalance between coastal and mountainous areas is still the main cause of the high poverty rate in this region.
- Cluster 2 includes areas with moderate poverty, large populations, but lower unemployment than Cluster 1. This may indicate that informal economic activity is quite developed, but has not been able to lift per capita expenditure significantly. The study [14] shows that limitations in the quality of education and access to MSME capital are still obstacles to poverty alleviation in these areas.
- Cluster 3, which consists only of large cities, shows the lowest poverty indicators and the highest life expectancy, reflecting better access to health services, education, and formal sector employment opportunities. However, the high unemployment in this cluster suggests that rapid urbanization is not always followed by adequate employment. This is consistent with the World Bank's (2022) finding that Indonesia's major cities face the challenge of "jobless growth".
- Cluster 4 consists of areas with low poverty rates and high expected years of schooling, but significant unemployment. This may indicate a mismatch between education and local labor market needs. According to the Ministry of National Development Planning/Bappenas (2021), vocational education and job training reforms are urgently needed in these areas to improve labor productivity.

CONCLUSIONS

Based on the analysis and discussion presented in the previous chapter, the following conclusions can be drawn:

1. Based on the results of the hierarchical cluster analysis research that applies Ward's Algorithm, the grouping of districts/cities in Central Java Province based on poverty indicators in 2023 resulted in 4 (four) clusters formed according to their respective characteristics.
2. Based on the dendrogram analysis, four distinct clusters were identified. The first cluster includes six districts/cities, while the second cluster consists of fourteen districts/cities. The

third cluster consists of three kabupaten/kota, and the fourth cluster consists of twelve kabupaten/kota. The following are the details of each of these clusters.

a. Cluster 1 includes Banjarnegara Regency, Pemalang Regency, Kenumen Regency, Purbalingga Regency, Wonosobo Regency, Brebes Regency. Cluster 1 requires special attention from the government in the form of direct interventions, such as increasing access to education and job creation. Poverty alleviation programs based on economic empowerment can be implemented in this region.
b. Cluster 2 includes Klaten Regency, Sragen Regency, Rembang Regency, Wonogiri Regency, Temanggung Regency, Grobogan Regency, Blora Regency, Magelang Regency, Purworejo, Pekalongan Regency, Banyumas Regency, Demak Regency, Cilacap Regency, Tegal Regency. Cluster 2 shows that poverty in some regions is still high despite lower unemployment rates. This indicates the need to improve access to business capital and job skills training.
c. Cluster 3 includes Surakarta City, Salatiga City, Semarang City. Cluster 3 reflects better socioeconomic conditions than the other clusters, but still faces challenges in unemployment. Therefore, strengthening the industrial sector and investing in productive labor can be the main solutions.
d. Cluster 4 includes Kendal Regency, Batang Regency, Karanganyar Regency, Pati Regency, Boyolali Regency, Pekalongan City, Tegal City, Sukoharjo Regency, Kudus Regency, Jepara Regency, Semarang Regency, Tegal City. Cluster 4 has a low poverty rate, but still faces challenges in terms of education. Therefore, policies that focus more on improving the quality of education and access to scholarships are needed.

Based on the characteristics obtained from each cluster, these findings can be used to design more targeted and data-driven development policies, such as budget allocation, poverty reduction programs, and improving access to basic services in accordance with the conditions of each cluster.

SUGGEST

Suggestions for future research suggest that clustering be done by considering the spatial aspect or location of the region. For example, geographically close regions may have similar poverty conditions. This can be analyzed using the spatial cluster method, so that it can be used to see whether poverty is evenly distributed or concentrated in certain areas. By adding a spatial approach, the research results are expected to be more accurate and can help the government in making more targeted policies, especially in areas that are close to each other and have similar poverty problems.

REFERENCES

[1] A. Salsabila, A. Fitrianto, and M. A. Aliu, "Association of Poverty Categories , Educational Characteristics , and Area of Residence in Indonesia Using a Three-Way Log-Linear Model," vol. 17, no. 1, pp. 624–634, 2024.
[2] M. Irfan, A. Samsir, M. Jamli, M. Syafri, and S. Astuty, "Faktor-Faktor Yang Mempengaruhi Tingkat Kemiskinan di Provinsi Sulawesi Selatan," *Ekonodinamika Jurnal Ekonomi Dinamis*, vol. 6, no. 2, pp. 182–197, 2024.
[3] A. Valiant Kevin, A. Bhinadi, and A. Syari'udin, "Pengaruh Pdrb, Angka Harapan Hidup, Dan Rata Rata Lama Sekolah Terhadap Kemiskinan Di Kabupaten/Kota Provinsi Jawa

Tengah Tahun 2013-2021," *SIBATIK JOURNAL: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan*, vol. 1, no. 12, pp. 2959–2968, 2022, doi: 10.54443/sibatik.v1i12.482.

[4] S. Yulianto, A. Z. Utami, and A. N. Ambarwati, "Perbandingan Model Spasial dalam Permasalahan Kemiskinan di Provinsi Jawa Timur," vol. 24, no. 2, pp. 143–150, 2024.

[5] I. Insiyah, M. Khasanah, and T. P. Hendarsyah, "Penerapan Metode Ward Clustering Untuk Pengelompokkan Daerah Rawan Kriminalitas Di Jawa Timur Tahun 2021," *Jurnal Statistika dan Komputasi*, vol. 2, no. 1, pp. 44–54, 2023, doi: 10.32665/statkom.v2i1.1664.

[6] putri nazwa Maharani, "Jurnal Pendidikan Inklusif Analisis Cluster Dengan Average Linkage Method," vol. 8, no. 12, pp. 48–67, 2024.

[7] M. J. Budiman and Fanny Jouke Doringin, "Jurnal Ilmu Komputer," *Biomaterials*, vol. 07, no. 12, pp. 85–90, 2023.

[8] S. Bao, Arman, La Gubu, W. Somayasa, Bahriddin, and Agusrawati, "Analisis Cluster Terhadap Tingkat Pencemaran Udara Pada Sektor Industri Di Sulawesi Tenggara," *Jurnal Matematika Komputasi dan Statistika*, vol. 4, no. 1, pp. 547–557, 2024, doi: 10.33772/jmks.v4i1.79.

[9] Y. I. Harnanto, A. Rusgiyono, and T. Wuryandari, "Penerapan Analisis Klaster Metode Ward Terhadap Kabupaten/Kota Di Jawa Tengah Berdasarkan Pengguna Alat Kontrasepsi," *Jurnal Gaussian*, vol. 6, no. 4, pp. 528–537, 2017.

[10] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "Metode K-Medoids Clustering Dengan Validasi Silhouette Index Dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018)," *Jurnal Gaussian*, vol. 8, no. 2, pp. 161–170, 2019, doi: 10.14710/j.gauss.v8i2.26640.

[11] R. K. Putri, M. Athoillah, and A. Haqiqiyah, "Analisis Faktor Yang Mempengaruhi Ketepatan Kelulusan Mahasiswa Dengan Algoritma Regresi Linear," *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 671–680, 2024.

[12] I. Imasdiani, I. Purnamasari, and F. D. T. Amijaya, "Perbandingan Hasil Analisis Cluster Dengan Menggunakan Metode Average Linkage Dan Metode Ward," *Eksponensial*, vol. 13, no. 1, p. 9, 2022, doi: 10.30872/eksponensial.v13i1.875.

[13] E. Saraswi, H. Perdana, and A. Fakhrunnisa, "Distribusi Tenaga Kesehatan di Kalimantan Barat Menggunakan Metode Ward," *Jurnal Forum Analisis Statistik (FORMASI)*, vol. 4, no. 1, pp. 39–48, 2024, doi: 10.57059/formasi.v4i1.73.

[14] T. Angriani *et al.*, "Peran UMKM Dalam Menanggulangi Kemiskinan Di Provinsi Kalimantan Tengah," vol. 2, no. 1, pp. 12–21, 2024.

Implementation of Geographically and Temporally Weighted Regression with Cross Validation and Generalized Cross Validation Methods for Deforestation Modeling in Kalimantan

Gema Khusnul Ma'rifah(1), Mohammad Idhom (2), Trimono (3)

[1,2,3]Program Studi Sains Data, Fakultas Ilmu Komputer,

Universitas Pembangunan Nasional Veteran Jawa Timur

Jl. Rungkut Madya, Surabaya, Jawa Timur

e-mail: 21083010034@student.upnjatim.ac.id (1), idhom@upnjatim.ac.id(2),
trimono.stat@upnjatim.ac.id 3)

ABSTRAK

Deforestasi di Indonesia mendapatkan sorotan di tingkat nasional hingga internasional yang menimbulkan dampak ekologi, ekonomi, dan sosial. Kalimantan termasuk wilayah dengan tingkat deforestasi yang tinggi dipicu oleh berbagai faktor yang bervariasi antar lokasi serta waktu. Penelitian ini bertujuan untuk memodelkan laju deforestasi di Kalimantan selama periode 2014 hingga 2022 menggunakan metode Geographically and Temporally Weighted Regression (GTWR). Model diuji dengan kombinasi fungsi pembobot Fixed dan Adaptive Gaussian Kernel serta metode penentuan bandwidth Cross Validation (CV) dan Generalized Cross Validation (GCV). Hasil menunjukkan bahwa model terbaik adalah GTWR dengan Fixed Gaussian Kernel dan GCV yang didasarkan pada nilai R^2 dan AIC. Analisis spasial-temporal menunjukkan bahwa tidak ada variabel signifikan pada 2014, kebakaran hutan signifikan pada 2019, dan pada tahun-tahun lainnya kedua variabel tersebut signifikan secara luas. Temuan ini memberikan wawasan terhadap dinamika spasial dan temporal faktor-faktor deforestasi untuk mendukung kebijakan berbasis wilayah.

Kata kunci: GTWR; Cross Validation; Generalized Cross Validation

ABSTRACT

Deforestation in Indonesia has received national and international attention for its ecological, economic and social impacts. Kalimantan is an area with high deforestation rates triggered by various factors that vary between locations and time. This study aims to model the deforestation rate in Kalimantan during the period 2014 to 2022 using the Geographically and Temporally Weighted Regression (GTWR) method. The model was tested with a combination of Fixed and Adaptive Gaussian Kernel weighting functions and Cross Validation (CV) and Generalized Cross Validation (GCV) bandwidth determination methods. The results show that the best model is GTWR with Fixed Gaussian Kernel and GCV based on R^2 and AIC values. Spatial-temporal analysis shows that neither variable is significant in 2014, forest fires are significant in 2019, and in other years both variables are broadly significant. The findings provide insights into the spatial and temporal dynamics of deforestation factors to support area-based policies.

Keywords: *GTWR; Cross Validation; Generalized Cross Validation*

INTRODUCTION

The rate of deforestation in Indonesia is one of the polemics that has received attention at the national and international levels. Indonesia ranks fifth in global tree cover loss from 2001 to 2023 and second in humid tropical primary forest loss from 2002 to 2023 [1]. Indonesia experienced the highest deforestation rate globally between 1996 and 2000 at 3.5 million hectares and decreased from 2002 to 2014 by 0.75 million hectares per year with the lowest point in 2022 at 104,000 hectares [2]. This decrease in deforestation rate was successfully suppressed by the implementation of the New Forest Moratorium and Reforestation Program. However, this decline has not succeeded in overcoming the various impacts caused both in terms of ecology, economy, and social such as changes in air temperature, loss of biodiversity, increased flooding, soil erosion, and increased emissions resulting in global warming.

Kalimantan recorded a deforestation rate of 1.12% annually from 1990 to 2014 [3]. The five provinces in Kalimantan are included in the 10 provinces with the worst deforestation in 2023 because almost half of the total deforestation in Indonesia occurred in Kalimantan with an area of 124,611 hectares out of 230,760 hectares [4]. Meanwhile, East Kalimantan is predicted to experience deforestation of up to 2,749.16 hectares in the next 10 years as a result of the move of Indonesia's new capital city [5]. This states that Kalimantan is one of the regions in Indonesia that continues to experience deforestation with the possibility of increasing deforestation rates caused by the relocation of the national capital.

Mining and logging activities as one of the significant economic activities have a major influence on ecological damage and high deforestation in Kalimantan along with the influence of population density [6]. The relocation of the national capital also triggers the potential for deforestation as a result of migration by changing the function of forest land into residential, industrial and office land [7]. The conversion of primary forests into oil palm plantations to fulfill market needs is also one of the causes of deforestation in Kalimantan [8]. Another cause of deforestation in Kalimantan that has a wide impact is forest and land fires [9].

Causal factors of deforestation exhibit both spatial and temporal variations, requiring in-depth analysis that considers both. Geographically Weighted Regression (GWR) and Multiscale Geographically Weighted Regression (MGWR) methods are able to capture spatial variations, but do not account for temporal changes. To overcome this limitation, the Geographically and Temporally Weighted Regression (GTWR) method is used because it is able to capture both spatial and temporal variations. Some previous studies have shown that GTWR is superior to global regression and GWR such as in modeling protected forest deforestation in Indonesia and in the analysis of the construction sector in Java Island using bisquare weight functions [10] [11]. Another study used GTWR Robust to overcome the violation of the normality assumption in the analysis of deforestation in Sumatra [12]. Research also showed that using Gaussian and Bisquare kernels gave almost similar results, while the Great Circle Distance approach with an exponential kernel still produced the best model in the analysis of the Human Development Index in West Java [13] [14].

Based on this, research on the GTWR method has shown development in various fields by discussing different aspects, especially related to the use of weight functions. However, there is no research that explicitly discusses the selection of methods to determine the optimum bandwidth of the weight function. The choice of method between Cross Validation (CV) and Generalized Cross Validation (GCV) in determining the optimum bandwidth affects the resulting estimate. Therefore,

this study uses Geographically and Temporally Weighted Regression with CV and GCV against two different weighting functions namely Fixed Gaussian Kernel and Adaptive Gaussian Kernel in the context of deforestation in Kalimantan. This research is expected to provide a deeper understanding of the causal factors and patterns of deforestation that varies between regions in Kalimantan each year.

METHOD

Data Collection

The research used several variables including deforestation factors from 2014 to 2022. All data used in this study is sourced from the Badan Pusat Statistika (BPS). The variables used in this study consist of Deforestation (Y), Population Density (X_1), Area of Forest and Land Fires (X_2), Production of Oil Palm Plantation Crops (X_3), Area of Oil Palm Plantation Crops (X_4), GDP Mining and Quarrying (X_5), and GDP Subcategory of Forestry and Logging (X_6). The total amount of data used in this study is 45 rows of data consisting of 5 provinces in Kalimantan, namely West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, and North Kalimantan.

Preprocessing Data

Data preprocessing is a series of stages carried out in preparing the quality of data before further analysis is carried out. The preprocessing carried out consists of Cleaning data; Data Integration; Handling Missing Values is done using the K-Nearest Neighbor Imputation (KNNI) [15] karena KNNI mampu mempertahankan struktur lokal dan pola kompleks dalam data spasial-temporal. KNNI dipilih daripada imputasi rata-rata yang cenderung terlalu menyederhanakan atau metode Multiple Imputation by Chained Equations (MICE) yang membutuhkan asumsi distribusi dan iterasi kompleks yang tidak diperlukan dalam konteks ini; and Handling Reforestation by changing the negative deforestation value, which means that reforestation is greater than deforestation, so to avoid misunderstanding the analysis, the value is changed to 0.

Linear Regression

A multiple linear regression model is used to explain the relationship between the dependent variable and several independent variables [16]. Simultaneous tests are conducted with the F-test [17], while partial tests use the t-test [18]. Classical assumption testing is conducted to ensure the validity of the model, including normality test using Kolmogorov-Smirnov [19], multicollinearity with VIF < 10, heteroscedasticity using Glejser test, and autocorrelation using Durbin-Watson test [20].

Spatial and Temporal Heterogeneity Testing

The Breusch-Pagan Test can be used in testing spatial heterogeneity [21]. The temporal heterogeneity test can be tested using a boxplot [12].

Geographically and Temporally Weighted Regression (GTWR)

The GTWR model is an extension of the GWR model. GTWR integrates spatial and temporal data in the weight matrix. The GTWR model is formulated in equation (1) [22].

$$Y_i = \beta_0(u_i, v_i, t_i) + \sum_{k=1}^{p} \beta_k (u_i, v_i, t_i) x_{ik} + \varepsilon_i \quad (1)$$

Y_i is the dependent variable with k as independent variables at location (u_i, v_i) with time t_i. Parameter estimation $\hat{\beta}(u_i, v_i, t_i)$ with the i -th point in GTWR is shown in equation (2) [22].

$$\hat{\beta}(u_i, v_i, t_i) = \left(X^T W(u_i, v_i, t_i)\right)^{-1} X^T W(u_i, v_i, t_i) Y \quad (2)$$

Let n be the number of observations and $W(u_i, v_i, t_i) = diag(\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in})$. The element $\alpha_{i1}(1 \leq j \leq n)$ represents the spatial-temporal distance function at (u_i, v_i, t_i). The following equation defines the spatial-temporal distance function that combines the spatial and temporal distance characteristics in equation (3) [22]:

$$(d^{ST})^2 = \lambda (d^S)^2 + \mu (d^T)^2 \quad (3)$$

The spatial and temporal distance functions are denoted by d^S and d^T with scale factors λ and μ that balance the different impacts of spatial and temporal distance measurements. Hence the Euclidean distance in equation (4) [22]:

$$\left(d_{ij}^{ST}\right)^2 = \lambda \left\{ \left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2 \right\} + \mu \left(t_i - t_j\right)^2 \quad (4)$$

With $\lambda \neq 0$ let τ be the ratio parameter of $\frac{\mu}{\lambda}$, then the equation becomes equation (5) [22]:

$$\frac{\left(d_{ij}^{ST}\right)^2}{\lambda} = \left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2 + \tau \left(t_i - t_j\right)^2 \quad (5)$$

With the effect of temporal distance, the spatial distance can be enlarged or minimized by the influence of the parameter τ. The model used for spatial weighting factors in GTWR uses the Fixed Gaussian Kernel (6) and Adaptive Gaussian Kernel (7) kernel function as follows [23]:

$$w_i(u_i, v_i) = exp\left[-\frac{1}{2} \left({d_{ij}}/{h} \right)^2 \right] \quad (6)$$

$$w_i(u_i, v_i) = exp\left[-\frac{1}{2} \left({d_{ij}}/{h_{i(q)}} \right)^2 \right] \quad (7)$$

The optimal bandwidth selection uses two methods namely Cross Validation (CV) and Generalized Cross Validation formulated in (8)(9) [24].

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2 \quad (8)$$

$$GCV = \sum_{i=1}^{n} [y_i - \hat{y}_{=i}(b)]^2 / (n - v_1)^2 \quad (9)$$

Model Evaluation

Model evaluation is performed using R^2 and Akaike Information Criterion (AIC) evaluation metrics is defined in equation (10)(11) [25].

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \quad (10)$$

$$AIC = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n + tr(G) \quad (11)$$

RESULT AND DISCUSSION

Descriptive Analysis

Table 1 contains an overview value of each variable in this research.

Table 1. Descriptive Analysis of Research Variable

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Min	0	8.00	82.2	167.7	50.3	6334	706

1st Qu	11564	17.00	1743.8	1343.0	479.3	14715	1425
Median	27240	28.00	7646.0	2192.6	1078.8	25572	2336
Mean	45206	39.07	47662.4	2950.2	1014.3	77116	3090
3rdQu	69585	37.00	47432.6	4100.9	1434.5	42459	4236
Std.dev	42986.9	35.2	102940	2386.6	668.4	115150	2242.7
max	186642	113.00	583833.4	8600.9	2205.8	490046	7488

Table 1 shows a very scattered distribution of the dependent variable (Y) with possible outliers and some areas with very high values. The independent variable (X_1) has a fairly wide spread indicating a positively skewed distribution. Variable (X_2) indicates extreme outliers and a highly abnormal distribution. The variable (X_3) indicates a very scattered distribution with possible outliers and shows a distribution that tends to be right (positively skewed). Variable (X_4) indicates a slightly negative skewed distribution, although not generally and is generally widely distributed. Variable (X_5) has a very wide and extreme data distribution indicating extreme dispersion and strong outlier potential. The variable (X_6) has a more moderate distribution than X_2 and X_5, but has a wide spread and is slightly positively skewed.

Linear Regression Modeling

The first step is to estimate the parameters using Ordinary Least Square (OLS) to explain the relationship between the independent variables and the dependent variable. The results of multiple linear regression model parameter estimation can be seen in Table 2.

Table 2. Parameter Estimation Value

Variable	Coefficient	Standard Error	t-Value	p-Value
Intercept	106200	32440	3.274	0.00226
X_1	-849.40	289.90	-2.930	0.00570
X_2	0.164	0.0529	3.100	0.00364
X_3	-9.434	6.107	-1.545	0.13069
X_4	6.743	19.63	0.344	0.73310
X_5	0.1553	0.1385	1.122	0.26906
X_6	-8.649	8.502	-1.017	0.31546

Furthermore, the simultaneous test shows that the F test statistical value is 4.8 and the p-value is 0.000985. The decision to reject H_0 means that simultaneously the independent variable has a significant influence on the dependent variable. In the partial test, the results show that variables X_1, and X_2 have a significant influence on the dependent variable with a p-value < 0.05. The variables X_3, X_4, X_5, X_6 are not significant because the p-value > 0.05 so they do not make a significant contribution to the dependent variable.

The first linear regression residual assumption test is the normality test with the test results showing a D value of 0.14315 with a p-value of 0.2869 which has a value greater than the 0.05 significance level. So it can be concluded that the data is normally distributed. The multicollinearity test results show that the variable that has a VIF value above 10 is variable X_6. In the heteroscedasticity test, the Glejser test shows a p-value < 0.05 on variable X_1 with a value of

0.02150 while X_2, X_3, X_4, X_5, and X_6 do not show significant heteroscedasticity in the model. The autocorrelation test results show a DW value of 1.8286 with a p-value of 0.1162. Because the p-value > 0.05, it is concluded that there is significant autocorrelation in the residuals of the regression model.

Based on the previous parameter estimation test results, variables X_3, X_4, X_5, and X_6 were excluded from the model due to their low statistical significance (p-value > 0.05), indicating a minimal contribution to the model. Additionally, variable X_6 showed signs of multicollinearity, which could affect the stability of coefficient estimates. Removing non-significant variables helps simplify the model without significantly reducing its predictive power. Since several assumption tests were not met, re-modeling was necessary, and new parameter estimates are presented in Table 3.

Table 3. Parameter Estimation Value

Variable	Coefficient	Standard Error	t-Value	p-Value
Intercept	57960	8589	6.749	3.33e-08
X_1	-513.50	156.70	-3.278	0.00211
X_2	0.1515	0.05346	2.835	0.00703

In the simultaneous test, it was found that the F test statistical value was 9.361 and the p-value was 0.0004345. The decision rejects H_0, which means that simultaneously the independent variable has a significant effect on the dependent variable. In the partial test, the results show that the variables X_1 and X_2 have a significant effect on the dependent variable with a p-value of 0.00211 and 0.00703. The normality test shows a D value of 0.10947 with a p-value of 0.6143 which has a value greater than the 0.05 significance level. The multicollinearity test results show that the variables have a VIF value below 10. The Glejser test shows a p-value < 0.05 on variable X_1 with a value of 0.0000551, while X_2 does not show significant heteroscedasticity in the model. The autocorrelation test results show a DW value of 0.98756 with a p-value of 0.0001141. Because the p-value < 0.05, it is concluded that there is significant autocorrelation in the residuals of the regression model.

Spatial and Temporal Heterogeneity Testing

Testing spatial heterogeneity using Breusch-Pagan results in a BP value of 8.6302 with a p-value of 0.01337. Because the p-value < 0.05, there is spatial heterogeneity. The temporal heterogeneity test is shown in Figure 1.

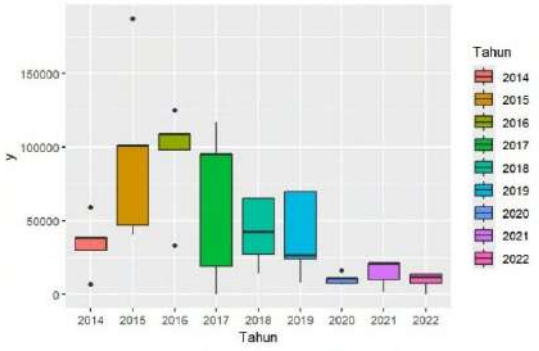


Figure 1. Boxplot of Deforestation in Kalimantan 2014-2022

Figure 1 shows the diversity of the dependent variable (Y) from year to year which illustrates significant fluctuations. Significant differences in the data distribution of some years indicate inhomogeneity in the Deforestation variable which needs to be further analyzed especially the factors that may affect deforestation during those years.

Geographically and Temporally Weighted Regression (GTWR)

Table 4 shows the parameter estimation summary of the Geographically Temporally Weighted Regression model with four kernel configurations: Gaussian Fixed, Gaussian Fixed GCV, Gaussian Adaptive, and Gaussian Adaptive GCV. Coefficient estimates for each variable are presented based on the minimum, maximum, mean, and standard deviation values.

Table 4. Summary of Coefficient Estimates

Models	Variable	Min	Max	Mean	Std. Deviation
Gaussian Fixed	Intercept	52889.34	91798.33	72528.19	13352.6
	X_1	-768.9184	-460.2715	-617.0524	96.53012
	X_2	-0.041245	0.2173	0.1303034	0.069369
Gaussian Fixed GCV	Intercept	55884.45	93006.57	73014.79	13376.2
	X_1	-780.0198	-499.2047	-623.4417	92.64175
	X_2	-0.12756	0.218129	0.1203035	0.094433
Gaussian Adaptive	Intercept	54247.05	65553.6	61295.97	3570.865
	X_1	-574.2845	-478.0567	-539.3515	28.28268
	X_2	0.13508	0.1748805	0.1536025	0.01171776
Gaussian Adaptive GCV	Intercept	57248.89	59723.09	58788.04	823.7398
	X_1	-527.8837	-507.1715	-520.1057	6.571252
	X_2	0.1474668	0.1576054	0.1521811	0.003044776

The GTWR model was compared with the linear regression model using the F test to see whether GTWR provides a significant improvement in explaining the data. Based on the tests conducted, the GTWR model with fixed bandwidth both CV and GCV are superior to the linear regression model because it provides a significant F value (p-value <0.05). In contrast, the GTWR model with adaptive bandwidth did not show a significant advantage over the linear regression model due to the high p-value. Furthermore, to see the spatial and temporal variations produced by the model, a partial test was conducted, the results of which are shown in Figure 2.

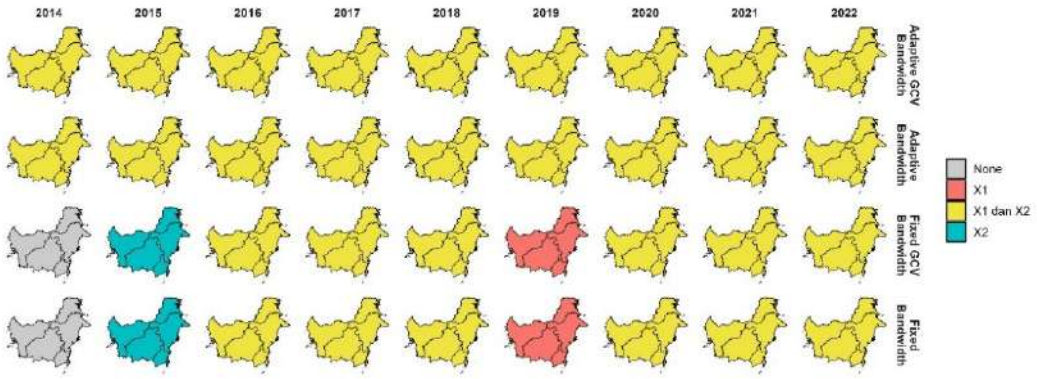


Figure 2. Distribution map based on significant variables

The GTWR model shows that the effect of population density and forest and land fires on deforestation is spatially and temporally variable depending on the year and location. In 2014, the Gaussian Fixed and Gaussian Fixed GCV models did not identify statistically significant predictors. The variable X2 (forest and land fires) started to show a significant effect in 2015. In 2016, both X1 (population density) and X2 became significant in more locations. From 2017-2022, all four models consistently showed that both X1 and X2 variables were statistically significant in most provinces and years, indicating that population density and fires were the main drivers of deforestation. In 2019, the Gaussian Fixed and Gaussian Fixed GCV models only show significance in X1, while the Adaptive model shows that both predictors are significant in all provinces in Kalimantan.

Population density tends to be the dominant predictor in years or areas with development activity, settlement expansion, or urbanization that could potentially increase pressure on forest areas. Forest and land fires are more significant in years with extreme weather conditions such as El Nino or when there is an increase in hotspots. This also reflects uneven fire control policies across provinces. The increased influence of X1 after 2019 can also be attributed to the planned relocation of the capital city to East Kalimantan which triggers population growth and new land activities in the vicinity.

Model Evaluation

Based on the evaluation model based on the R^2 and AIC values of each model, it can be concluded that the Gaussian Fixed GCV model is the model that performs the best modeling of the data. The value of the evaluation model can be seen in Table 5.

Table 5. Model Evaluation

Models	R^2	AIC
Gaussian Fixed	0.4503620	1067.479
Gaussian Fixed GCV	0.4599574	1067.243
Gaussian Adaptive	0.3564924	1070.633
Gaussian Adaptive GCV	0.3203041	1072.631

CONCLUSION

Geographically and Temporally Weighted Regression (GTWR) modeling in the case of deforestation in Kalimantan during the period 2014-2022 was carried out with two types of spatial weighting functions namely Gaussian Fixed and Gaussian Adaptive, each with two bandwidth selection methods Cross Validation (CV) and Generalized Cross Validation (GCV). The evaluation results show that the Gaussian Fixed Kernel model with GCV bandwidth provides the best performance with an R^2 value of 0.4599574 and AIC of 1067.243. This study is limited to the use of Gaussian Fixed and Adaptive kernels as well as CV and GCV methods in bandwidth selection. Future research can develop models with other kernel functions such as Bisquare and consider adding relevant variables to improve model accuracy.

REFERENCE

[1] M. Weisse and E. Goldman, "Top 10 Lists," World Resources Institute. Accessed: Aug. 20, 2024. [Online]. Available: https://research.wri.org/id/node/77

[2] KLHK, "Pengendalian Deforestasi dan Karhutla di Indonesia," PPID Kementerian Lingkungan Hidup dan Kehutanan Pejabat Pengelola Informasi dan Dokumentasi. Accessed: Aug. 20, 2024. [Online]. Available: https://ppid.menlhk.go.id/berita/siaran-pers/7594/pengendalian-deforestasi-dan-karhutla-di-indonesia

[3] Yusuf Aguswan, "POLA DEGRADASI DAN DEFORESTASI DI KESATUAN HIDROLOGIS GAMBUT (KHG) PROVINSI KALIMANTAN TENGAH TAHUN 2016 -2017," *Jurnal Hutan Tropika (Tropical Forest Journal)*, vol. XIV, no. 2, pp. 89–98, Dec. 2019, doi: 10.36873/jht.v14i2.1151.

[4] D. L. Pristiandaru, "10 Provinsi dengan Deforestasi Terparah 2023, Mayoritas di Kalimantan," Kompas.com. Accessed: Aug. 20, 2024. [Online]. Available: https://lestari.kompas.com/read/2024/03/28/090000586/10-provinsi-dengan-deforestasi-terparah-2023-mayoritas-di-kalimantan

[5] A. H. Nawawi and Evangs Mailoa, "Prediksi Lahan Deforestasi Dan Reforestasi Hutan Kalimantan Timur Dengan Metode Rantai Markov," *Decode: Jurnal Pendidikan Teknologi Informasi*, vol. 4, no. 1, pp. 251–259, Feb. 2024, doi: 10.51454/decode.v4i1.268.

[6] D. N. Isnaini *et al.*, "Determinants of Deforestation in Kalimantan," in *Seminar Nasional Official Statistics*, 2021, pp. 978–988. doi: 10.34123/semnasoffstat.v2020i1.570.

[7] T. Berlianty and T. Meiliana, "Potensi Deforestasi di Pulau Kalimantan: Pro dan Kontra Migrasi," *International Journal of Demos (IJD)*, vol. 5, no. 2, pp. 279–290, Jun. 2023, doi: 10.37950/ijd.v5i2.426.

[8] N. I. Fawzi and M. Y. Iswari, "Analisis Heat Island pada Perkebunan Kelapa Sawit: Studi Kasus di Kabupaten Kayong Utara, Kalimantan Barat," *Jurnal Wilayah dan Lingkungan*, vol. 8, no. 2, pp. 106–115, Aug. 2020, doi: 10.14710/jwl.8.2.106-115.

[9] F. Ulandari and R. Kurniawan, "PERBANDINGAN ALGORITMA LSDBC DAN DBSCAN PADA PEMETAAN DAERAH RAWAN KEBAKARAN HUTAN: Studi Kasus di Pulau Sumatera, Kalimantan, Sulawesi, dan Papua," *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 12, no. 2, pp. 25–30, Dec. 2020, doi: 10.34123/jurnalasks.v12i2.281.

[10] T. Z. Adiningrum, A. Prahutama, R. Santoso, D. Statistika, F. Sains, and D. Matematika, "PEMODELAN DEFORESTASI HUTAN LINDUNG DI INDONESIA MENGGUNAKAN MODEL GEOGRAPHICALLY AND TEMPORALLY WEIGHTED REGRESSION (GTWR)," *Jurnal Gaussian*, vol. 7, no. 3, pp. 314–325, Aug. 2018, doi: 10.14710/j.gauss.7.3.314-325.

[11] S. Haryanto, M. N. Aidi, and A. Djuraidah, "Analysis of the Geographically and Temporally Weighted Regression (GTWR) of the GRDP the Construction Sector in Java Island," *Forum Geografi*, vol. 22, no. 1, pp. 130–139, 2019, doi: 10.23917/forgeo.v33i1.7332.

[12] A. D. Putra and S. I. Oktora, "Spatial-Temporal Analysis of Deforestation in Sumatera Island 2011-2019," in *Proceedings of The International Conference on Data Science and Official Statistics*, 2022, pp. 590–609. doi: 10.34123/icdsos.v2021i1.202.

[13] Harianto, W. H. Nugroho, and E. Sumarminingsih, "Geographically and Temporally Weighted Regression Model with Gaussian Kernel Weighted Function and Bisquare Kernel Weighted Function," *IOP Conf Ser Mater Sci Eng*, vol. 1115, no. 1, p. 1, Mar. 2021, doi: 10.1088/1757-899x/1115/1/012063.

[14] A. D. Rahmawati, A. H. Wigena, and M. N. Aidi, "Influencing factors for the human development index in West Java using geographically and temporally weighted regression kernel functions," *Jurnal Pendidikan Geografi: Kajian, Teori, dan Praktek dalam Bidang Pendidikan dan Ilmu Geografi*, vol. 28, no. 2, p. 228, Jun. 2023, doi: 10.17977/um017v28i22023p228-241.

[15] Miraati Laia, "Analisis Kinerja Algoritma K-Nearest Neighbor Imputation (KNNI) Untuk Missing Value Pada Klasifikasi Data Mining," *Journal of Informatics, Electrical and Electronics Engineering*, vol. 2, no. 3, pp. 92–98, Mar. 2023, doi: 10.47065/jieee.v2i3.891.

[16] T. Maulana Fahrudin, P. Aji Riyantoko, K. Maulida Hindrayani, and I. Gede Susrama Mas Diyasa, "Daily Forecasting for Antam's Certified Gold Bullion Prices in 2018-2020 using Polynomial Regression and Double Exponential Smoothing," *Journal of International Conference Proceedings*, vol. 3, no. 4, pp. 45–53, 2020, doi: 10.32535/jicp.v3i4.1009.

[17] T. Arifianto, Y. A. Pangestu, D. S. Oktaria, L. S. Moonlight, and D. I. Pratiwi, "Prediksi Daya Pada Panel Surya Menggunakan Metode Time Series dan Analisis Regresi," *Jurnal Ilmiah Intech : Information Technology Journal of UMUS*, vol. 4, no. 01, pp. 52–63, May 2022, doi: 10.46772/intech.v4i01.674.

[18] Y. Widyaningsih, G. P. Arum, and K. Prawira, "APLIKASI K-FOLD CROSS VALIDATION DALAM PENENTUAN MODEL REGRESI BINOMIAL NEGATIF TERBAIK," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 15, no. 2, pp. 315–322, Jun. 2021, doi: 10.30598/barekengvol15iss2pp315-322.

[19] T. Trimono, D. I. Asih Maruddani, and D. Ispriyanti, "PEMODELAN HARGA SAHAM DENGAN GEOMETRIC BROWNIAN MOTION DAN VALUE AT RISK PT CIPUTRA DEVELOPMENT Tbk," *JURNAL GAUSSIAN*, vol. 6, no. 2, pp. 261–270, 2017, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[20] G. Mardiatmoko, "PENTINGNYA UJI ASUMSI KLASIK PADA ANALISIS REGRESI LINIER BERGANDA (STUDI KASUS PENYUSUNAN PERSAMAAN ALLOMETRIK KENARI MUDA [CANARIUM INDICUM L.])," *BAREKENG: Jurnal Ilmu Matematika*

dan Terapan, vol. 14, no. 3, pp. 333–342, Oct. 2020, doi: 10.30598/barekengvol14iss3pp333-342.

[21] Deviani, B. Y. Wulandari, Askariyah, I. J. Fitri, and S. Hariati, "ANALISIS SPASIAL BERBASIS PEMETAAN MENGGUNAKAN GEOGRAPHICALLY WEIGHTED REGRESSION (GWR) PADA SEBARAN JUMLAH WISMA DI KABUPATEN LOMBOK BARAT," *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 660–670, Aug. 2024, doi: 10.46306/lb.v5i2.545.

[22] B. Huang, B. Wu, and M. Barry, "Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices," *International Journal of Geographical Information Science*, vol. 24, no. 3, pp. 383–401, Mar. 2010, doi: 10.1080/13658810802672469.

[23] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression the analysis of spatially varying relationships*. England: John Wiley & Sons Ltd, 2002.

[24] A. T. Damaliana, I. Nyoman Budiantara, and V. Ratnasari, "Comparing between mgcv and agcv methods to choose the optimal knot points in semiparametric regression with spline truncated using longitudinal data," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/3/032003.

[25] A. Muhaimin, H. Prabowo, and Suhartono, "Model Selection for Forecasting Rainfall Dataset," *IJDASEA (International Journal of Data Science, Engineering, and Analytics)*, vol. 1, pp. 1–10, 2021, doi: 10.3390/xxxxx.

Comparison of Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) Methods in Determining Factors Affecting Tuberculosis Cases in Indonesia

Awaliyatul Uswah [(1)], Jose Rizal [(2)], Yulian Fauzi [(3)]

[1,2,3] Master of Statistics Study Program, University of Bengkulu, Bengkulu

WR. Supratman Street, Kandang Limun Village, Muara Bangkahulu District, Bengkulu City

e-mail: awaliawell359@gmail.com[(1)], jrizal@unib.ac.id[(2)], yulianfauzi@unib.ac.id[(3)]

ABSTRAK

Model Geographically Weighted Generalized Poisson Regression (GWGPR) dan Geographically Weighted Negative Binomial Regression (GWNBR) pada temuan penelitian ini menunjukkan hasil efektif dalam memodelkan data insiden tuberkulosis (TB) yang dicirikan oleh overdispersi dan heterogenitas spasial. Meskipun kedua model menghasilkan statistik kecocokan yang sebanding, seperti yang ditunjukkan oleh nilai Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) yang hampir identik, GWGPR menunjukkan sensitivitas yang lebih tinggi terhadap variabilitas regional, sebagaimana dibuktikan oleh pembentukan empat klaster provinsi yang berbeda berdasarkan variabel prediktor yang signifikan, dibandingkan dengan hanya dua klaster yang diidentifikasi oleh model GWNBR. Hal ini menunjukkan bahwa GWGPR dapat menawarkan pemahaman yang lebih berwarna tentang efek spasial dalam data epidemiologi. Pada beberapa variabel yaitu prevalensi merokok, kelembaban rata-rata tahunan, jumlah hari hujan, persentase penduduk yang melaporkan keluhan kesehatan, dan persentase penemuan dan pengobatan TB yang terbukti signifikan secara konsisten di semua provinsi dalam kedua pendekatan pemodelan. Hasil ini mendukung pentingnya teknik pemodelan yang terbobot secara geografis tidak hanya untuk meningkatkan akurasi prediktif tetapi juga untuk menginformasikan pengaruh kesehatan masyarakat berdasarkan wilayah. Dengan demikian, penggunaan model yang adaptif secara spasial seperti GWGPR dapat mendukung strategi pengendalian penyakit yang lebih terarah dan efektif dengan menyelaraskan respons kebijakan kesehatan dengan kebijakan penanggulangan TB di wilayah setempat.

Kata kunci: *AIC, BIC,* GWGPR, GWNBR, heterogenitas spasial, overdispersi, tuberkulosis

ABSTRACT

The findings of this study demonstrate that both the Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR) models are effective in modeling tuberculosis (TB) incidence data characterized by overdispersion and spatial heterogeneity. Although both models yield comparable fit statistics—as indicated by nearly identical Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values—GWGPR exhibits a higher sensitivity to regional variability, as evidenced by the formation of four distinct provincial clusters based on significant predictor variables, compared to only two clusters identified by the GWNBR model. This suggests that GWGPR may offer a more nuanced understanding of spatial effects in epidemiological data. Furthermore, several covariates; namely smoking prevalence, average annual humidity, number of rainy days, reported health complaints, and TB case detection and treatment coverage, emerged as consistently significant across all provinces in both modeling approaches. The recurrence of these variables across spatially disaggregated models highlights their fundamental role in influencing TB transmission dynamics at a national scale. Accordingly, the use of spatially adaptive models such as GWGPR can support more

targeted and effective disease control strategies by aligning health policy responses with the localized determinants of TB burden.
Keywords: *AIC, BIC, GWGPR, GWNBR, overdispersion, spatial heterogeneity, tuberculosis*

INTRODUCTION

Modeling count data is a common approach in statistical analysis when the dependent variable represents discrete, such as the number of disease cases. However, a frequent challenge in modeling count data is the presence of overdispersion, a condition in which the variance exceeds the mean [1]. This violates the key assumption of the Poisson regression model, which assumes equal mean and variance, leading to biased parameter estimates and inefficient inferences. To address this, extended count models such as Negative Binomial Regression (NBR) and Generalized Poisson Regression (GPR) have been developed [2]. These models introduce additional parameters to account for overdispersion and have been widely used in various applied contexts [1], [3].

Beyond overdispersion, spatial structure in the data presents another modelling challenge. In many real-world scenarios, observations are not independent but exhibit spatial correlation—values in one region may influence or resemble those in nearby regions. This is particularly relevant in public health, where disease spread and risk factors are often geographically patterned. To account for such spatial heterogeneity, extensions of GPR and NBR have been developed: Geographically Weighted Generalized Poisson Regression (GWGPR) and Geographically Weighted Negative Binomial Regression (GWNBR)**.** These geographically weighted models incorporate spatial coordinates (e.g., latitude and longitude) and allow model parameters to vary across locations, enabling more localized and context-sensitive analysis.

A relevant and pressing application of these spatial count models is in analysing tuberculosis (TB), a chronic infectious disease primarily affecting the lungs and caused by *Mycobacterium tuberculosis*. Transmission occurs through inhalation of airborne droplets, making the spread of TB highly sensitive to both environmental conditions and population density. Given its mode of transmission and the complex interaction between biological and ecological factors, TB incidence exhibits significant spatial variation. Previous studies have shown that environmental, climatic, and socioeconomic conditions vary markedly across Indonesian provinces, contributing to spatial heterogeneity in epidemic cases[1].

In light of these challenges, this study proposes a comparative spatial regression analysis using the GWGPR and GWNBR models to explore factors associated with TB incidence across Indonesia in 2021. Both models will incorporate adaptive bi-square kernel weighting to reflect spatial proximity. Model performance will be evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to determine the most suitable approach for addressing both modelling. This analysis aims to provide more accurate insights into TB determinants while supporting the development of geographically targeted public health interventions. the study intends to identify which approach better reflects the spatial dynamics of TB and to uncover key environmental and health-related predictors influencing its spread across regions.

METHOD

Before applying the GWGPR and GWNBR models, a multicollinearity diagnostic was conducted to ensure the independence of predictor variables. Multicollinearity occurs when two or

more independent variables are highly correlated, leading to unstable coefficient estimates and inflated standard errors. Identifying multicollinearity in independent variables based on the Variance Inflation Factor (VIF) value with the following equation [4], [5]:

$$VIF = \frac{1}{1 - R_k^2}$$

The estimation of Poisson regression model parameters using the Maximum Likelihood Estimation (MLE) method by maximizing the likelihood function, which measures how likely it is to observe the given sample data under different parameter values. [6].

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

The assumption in poisson regression is that the mean and variance of the response variable are equal. Overdispersion occurs when the variance significantly exceeds the mean, which can invalidate the model's standard errors and lead to misleading inferences. Checking for overdispersion using the formula for the Pearson chi-square statistical test [7]:

$$\theta = \frac{\chi^2}{n-p}, \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mu_i}$$

The parameter estimation of generalized Poisson regression using the MLE method with the likelihood function [8]:

$$L(\theta, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \theta exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)} \right)^{y_i} \frac{(1 + \theta y_i)^{y_i - 1}}{y_i!} exp\left[\frac{-exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)(1 + \theta y_i)}{1 + \theta exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)} \right]$$

The parameter estimation of the Negative Binomial Regression (NBR) model can be done using the MLE method. The likelihood function of NBR [9]:

$$L(\beta, \theta) = \prod_{i=1}^{n} \left(\frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1}) + \Gamma(y_i + 1)} \left(\frac{1}{1 + \theta\mu_i} \right)^{\theta^{-1}} \left(\frac{\theta\mu_i}{1 + \theta\mu_i} \right)^{y_i} \right)$$

A statistically significant and positive Moran's I value would suggest that provinces with high (or low) TB counts tend to be near others with similar values, supporting the use of spatial modeling approaches. Moran's index test in testing spatial dependency of data, namely whether or not there is a spatial influence on the data [10].

$$Z(I) = \frac{I - E\,(I)}{\sqrt{Var(I)}}$$

To identify spatial heterogeneity, we assess whether the relationship between variables varies across different spatial locations. Spatial heterogeneity implies that the data-generating process differs across geographic regions, violating the assumption of global homogeneity in standard regression models.Breusch Pagan test in identifying spatial heterogeneity in data [10].

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

Calculating the distance with the Haversine equation between observation locations based on geographical position. The Haversine equation can be formulated as follows [11] & [12]:

$$a = sin^2\left(\frac{\Delta lat}{2}\right) + \cos(lat_1) \cdot cos \cdot sin^2\left(\frac{\Delta long}{2}\right)$$

$$d_{ij} = 2r \cdot arc\sin\left(\sqrt{a}\right)$$

The optimal bandwidth for each observation in the geographically weighted models was determined using the cross-validation (CV) method [13]. This process identified the bandwidth that minimized the sum of squared prediction errors, thereby ensuring the best balance between local sensitivity and model generalizability. Optimum bandwidth for each observation location using Cross Validation (CV) can be written as follows [14]:

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

The selected bandwidth was then applied using an adaptive bisquare kernel, allowing each location to use a neighborhood size suited to its spatial context. Determining weights with the adaptive bisquare kernel function [15].

$$W_{ij}(u_i, v_i) = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h_i}\right)^2\right)^2, & d_{ij} \leq h_i \\ 0, & d_{ij} > h_i \end{cases}$$

With d_{ij} calculated by two Haversine distance equations between location (u_i, v_i) to location (u_j, v_j) and h is the bandwidth .

Geographically Weighted Generalized Poisson Regression (GWGPR) parameter estimation is performed using the MLE method as follows:

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

Conducting Geographically Weighted Negative Binomial Regression (GWNBR) modeling. The GWNBR method can be formulated with [16]:

$$y_i \sim NB\left[t_i\left(\sum_k \beta_k(u_i, v_i)x_{ik}\right), \theta(u_i, v_i)\right]$$

The estimation of the GWNBR coefficient parameters was performed using the MLE method. The weighted natural logarithm function for the GWNBR model is:

$$ln\, L(\boldsymbol{\beta}(u_i, v_i), \theta_i | y_i x_i)$$
$$= \sum_{n=1}^{n} w_{ij}\,(u_i, v_i)\left\{ln\left[\frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1}) + \Gamma(y_i + 1)}\right] - (y_i + \theta^{-1})ln(1+\theta\mu_i) + y_i ln(\theta\mu_i)\right\}$$

Simultaneous significance test using Maximum Likelihood Ratio Test. Hypothesis for the Poisson regression method, GPR, and NBR are as follows:

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

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

Meanwhile, for GWGPR and GWNBR there are spatial elements (u_i, v_i), so the hypothesis is:

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

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

Statistical test (Likelihood Ratio Test):

$$D(\hat{\beta}) = -2ln\left|\frac{L(\widehat{\omega})}{L(\widehat{\Omega})}\right| = 2\ ln\left(L(\widehat{\Omega}) - L(\widehat{\omega})\right)$$

The function $L(\widehat{\omega})$ the maximum likelihood value for a simple model without involving predictor variables and $L(\widehat{\Omega})$ for a complete model. Reject H_0 if the value $D(\hat{\beta}) > \chi^2_{(\alpha,k)}$ which means that there is at least one parameter in the model that has a significant effect on the model. The value of $D(\hat{\beta})$ is the deviation, the smaller the value, the smaller the error produced by the model, so that the model becomes more precise.

Partial testing is carried out to determine which parameters have a significant effect on the model. In the Poisson, GPR, and NBR regression methods with the following test hypotheses:

$H_0 : \beta_j = 0$

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

The test statistics used follow the z distribution, namely:

$$Z_{hit} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

Meanwhile, for GWGPR and GWNBR there are spatial elements (u_i, v_i), so the hypothesis is:

$H_0 : \beta_j(u_i, v_i) = 0$

$H_1 : \beta_j(u_i, v_i) \neq 0, j = 1,2, \dots, k$

Test statistics:

$$Z_{hit} = \frac{\hat{\beta}_j(u_i, v_i)}{se\left(\hat{\beta}_j(u_i, v_i)\right)}$$

The rejection criterion for the partial test hypothesis, namely reject H_0 if the value of $|Z| > Z_{\alpha/2}$ which means that the parameter has a significant effect on the model.

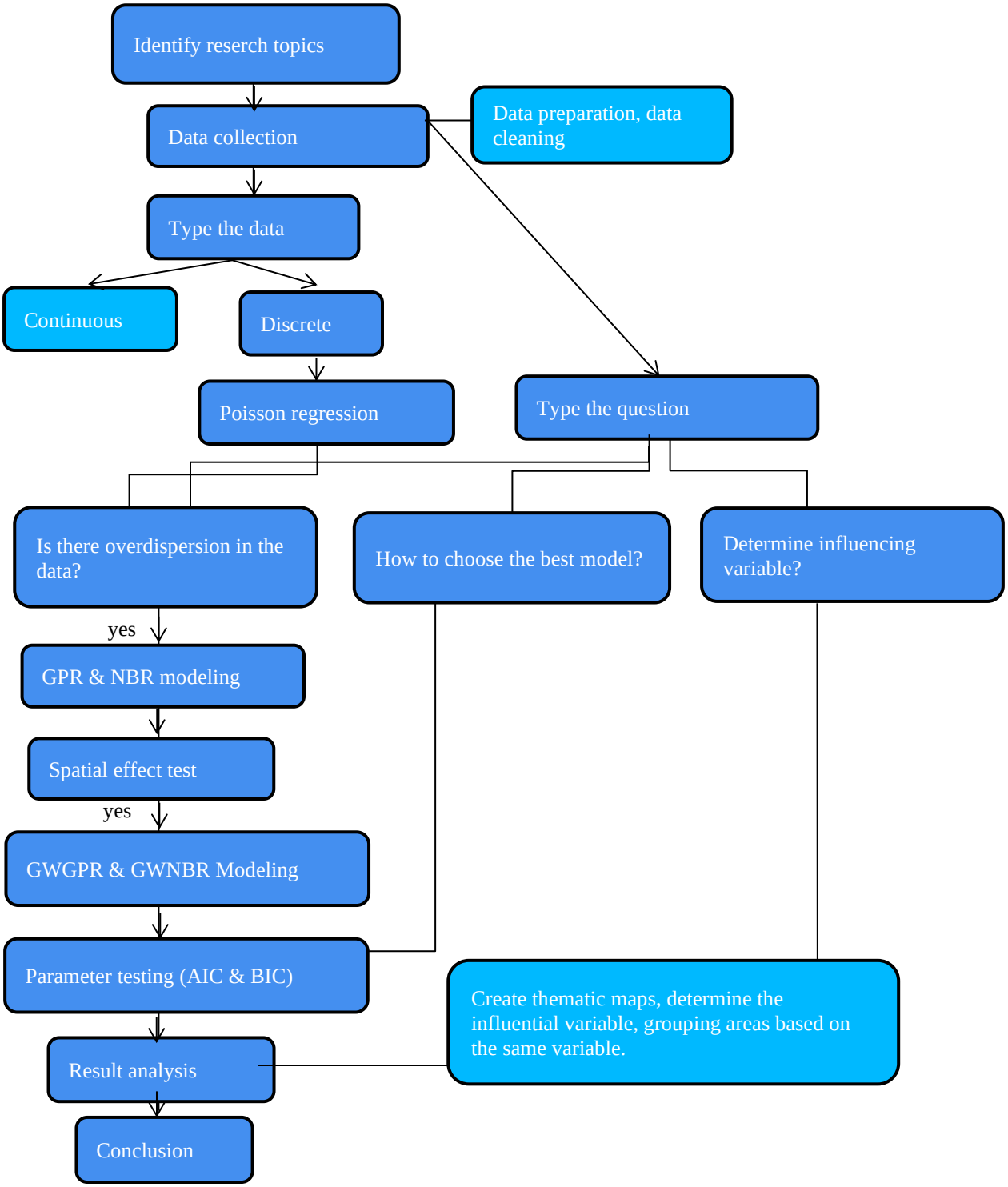
Comparing the goodness of fit values of the GWGPR and GWNBR models with two distances on the adaptive bi-square weighting based on the AIC and BIC criteria. According to [17] the equation for AIC is as follows:

$$AIC = -2logL(\hat{\beta}) + 2k$$

Meanwhile, the BIC formula is as follows [18]:

$$BIC = -2logL + log(n)k$$

RESEARCH FLOWCHART



Gambar 1. Flowchart of Research

RESEARCH VARIABLE

The variables used in this study are presented below.

Y: Number of positive tuberculosis (TB) cases
X_1: Percentage of smokers
X_2: Average annual temperature (°C)
X_3: Average annual humidity percentage
X_4: Average wind speed
X_5: Number of rainy days in a year
X_6: Average sunlight exposure
X_7: Population density
X_8: Percentage of population reporting health complaints
X_9: Percentage of TB detection and treatment
X_{10}: Percentage of poor population
X_{11}: Number of medical personnel
X_{12}: Number of healthcare workers
X_{13}: Number of community health centers (Puskesmas)

RESULT AND DISCUSSION

The number of TB cases in Indonesia in 2021 ranked second highest in the world after India. The trend of the number of TB cases has continued to increase since 2018-2021 [19]. The average number of TB cases in Indonesia in 2021 was 13,036 cases, with the highest cases in West Java Province, namely 103,253 cases, followed by Central Java Province with 47,430 cases and East Java with 47,398 cases, while the lowest cases were in North Sulawesi Province with 1,050 cases. The following is a graph of the number of TB cases in Indonesia:

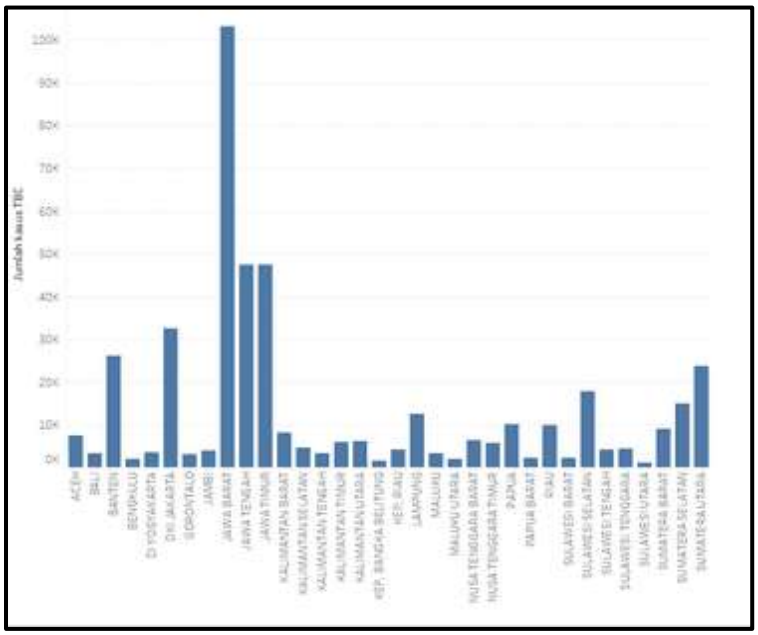


Figure 2. Graph of the number of TB cases per province in Indonesia in 2021

Multicollinearity testing is used to determine whether the independent variables meet the assumption that they are not correlated with each other so the model formed is a regression model in which there is no perfect relationship between the independent variables.

Table 1. VIF Values of Independent Variables

Variable	VIF value	Variable	VIF value
X_1	1.723	X_8	2.631
X_2	3.461	X_9	2.042
X_3	3.657	X_{10}	2.426
X_4	2.565	X_{11}	58.791
X_5	2.396	X_{12}	59.311
X_6	2.904	X_{13}	50.975
X_7	14.856		

Based on the VIF Value in Table 1, it is known that of the 13 variables tested, 9 of them have a VIF value < 10. Four variables produce VIF values > 10, namely X_7, X_{11}, X_{12}, X_{13}. While the variables that have VIF values < 10 namely X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9 and X_{10}.

The result of the Pearson chi-square overdispersion test is 9536.552 which is greater than 1. So it is concluded that in the Poisson regression for TB case data in Indonesia in 2021, there was overdispersion. The occurrence of overdispersion in Poisson regression, the steps to handle it are to form a Generalized Poisson Regression and Negative Binomial Regression model.

Table 2. Estimated values of GPR model parameters

Parameter	Estimate	Std. Error	Z Value
$\hat{\beta}_0$	-4838	0,125	-38,804
$\hat{\beta}_1$	-0,014	0,038	-0,370
$\hat{\beta}_2$	-0,142	0,236	-0,600
$\hat{\beta}_3$	-0,083	0,056	-1,469
$\hat{\beta}_4$	-0,271	0,178	-1,518
$\hat{\beta}_5$	-0,007	0,005	-1,317
$\hat{\beta}_6$	-0,006	0,019	-0,309
$\hat{\beta}_8$	0,011	0,031	0,346
$\hat{\beta}_9$	0,057	0,014	4,116
$\hat{\beta}_{10}$	0,000	0,022	0,000
Devians	95.685		
AIC	691.59		

Table 3. Estimated values of the NBR model parameters

Parameter	Estimate	Std. Error	Z Value
$\hat{\beta}_0$	21,638558	7.1212245	3,039
$\hat{\beta}_1$	0,021245	0,044055	0,482
$\hat{\beta}_2$	-0,228496	0,162272	-1,408
$\hat{\beta}_3$	-0,082547	0,047744	-1,729
$\hat{\beta}_4$	-0,178384	0,182653	-0,977
$\hat{\beta}_5$	-0,007680	0,005102	-1,505
$\hat{\beta}_6$	0,001284	0,017675	-0,073
$\hat{\beta}_8$	0,014469	0,028613	0,506
$\hat{\beta}_9$	0,053197	0,010223	5,203
$\hat{\beta}_{10}$	-0,025263	0,025393	-0,995
Devians	95.685		
AIC	697.52		

The spatial effects tested as a condition for spatial regression are the spatial dependency test and spatial heterogeneity test. The results of spatial dependency test with Moran Index obtained $p\text{-}value = 0{,}000073$. With a significance level of $\alpha = 5\%$ the conclusion is rejected H_0 meaning that there is spatial dependency between locations, between observations of one location and other adjacent locations that influence each other.

Spatial heterogeneity test with Breusch Pagan test obtained $BP = 18{,}227$ and $p\text{-}value = 0{,}03263 < \alpha = 0{,}05$. The results of this value indicate that there is spatial heterogeneity or diversity between regions so that the parameters produced in each region can vary. Spatial heterogeneity conditions with overdispersion can be modeled using the GWGPR and GWNBR methods.

The estimation of GWGPR model parameters is obtained by including spatial weighting in its calculations using the Newton-Raphson iteration method. The estimation of GWGPR model parameters with adaptive Bisquare Kernel spatial weighting in each province obtained the value of Parameter estimation at each research location (u_i, v_i), where $i = 1, 2, 3, .., 4$.

Simultaneous parameter testing based on the deviation value is 11158, and a level greater than the value of $\chi^2_{(0.05,9)} = 16{,}919$. So reject H_0 hich means that there is at least one independent variable that has a significant effect on the dependent variable in each GWGPR model.

The partial test results produce different parameters in several provinces. Based on this difference, groups can be made where the number of TB cases that occur is influenced by the same variables.

Table 4. Grouping of provinces that have the same significant independent variables from the GWGPR model

No	Province	Significant Variables
1	Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, South Kalimantan, East Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi, Maluku, West Papua, Papua	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}$
2	Central Kalimantan	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9$
3	North Kalimantan	X_1, X_3, X_5, X_8, X_9
4	North Sulawesi, Gorontalo, North Maluku	$X_1, X_3, X_4, X_5, X_6, X_8, X_9, X_{10}$

The following is a visualization of the variable grouping in the GWGPR model:

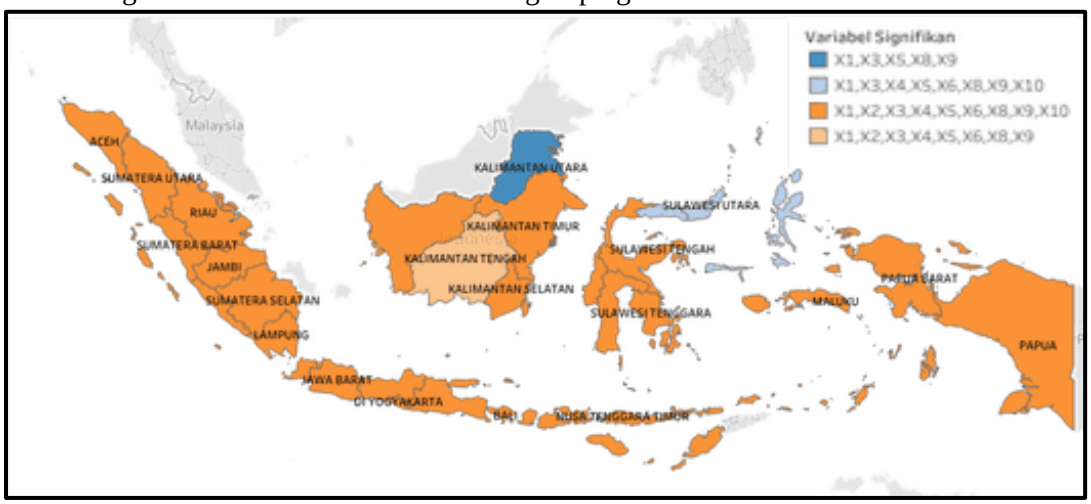


Figure 3. Grouping of provinces based on the significance of the same variables from the GWGPR model.

GWNBR parameter estimation uses the MLE method with the Newton-Raphson numerical iteration method, as well as geographic coordinate weighting of the region. The weighting components used are geographic location based on the spatial weighting matrix of the bandwidth and distance values of each location so that they have different parameters and describe the special or local properties of the model.

Simultaneous parameter testing is on the results of the deviation value $= 17549{,}5$ which is greater than the value of $\chi^2_{(0.05,9)} = 16{,}91898$. So reject H_0 which means that there is at least one independent variable that has a significant effect on the dependent variable in the GWNBR model. Grouping of provinces based on the similarity of independent variables that have a significant effect based on the partial test of the GWNBR model in the case of TBC.

Table 5. Grouping of provinces that have the same significant independent variables from the GWNBR model

No	Province	Significant Variables
1	Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, South Kalimantan, East Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi, Maluku, West Papua, Papua, Central Kalimantan, North Sulawesi, Gorontalo, North Maluku	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9$
2	North Kalimantan	$X_1, X_2, X_3, X_4, X_5, X_6, X_8, X_9$

The following is a map visualization of the variable grouping in the GWNBR model:

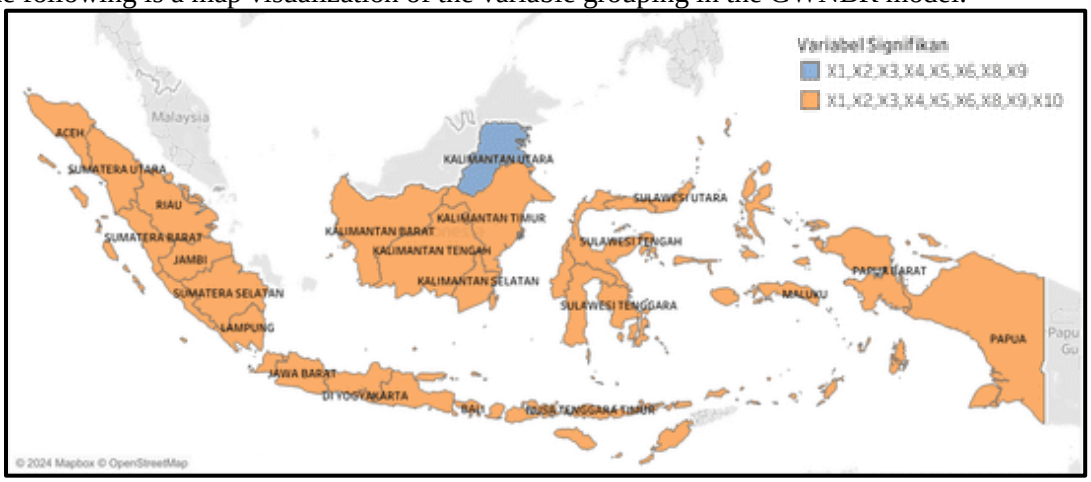


Figure 4. Grouping of provinces based on the same significant variables from the GWNBR model

The following is a comparison table of the model goodness of fit test results.

Table 6. Comparison AIC dan BIC model

No	Model	AIC	BIC
1	Poisson Regression	216152	216167.3
2	GPR	691.590	708.380
3	NBR	697.522	714.312
5	GWGPR	706.570	723.360
7	GWNBR	706.566	723.356

Based on Table 6, the smallest AIC and BIC values are the GPR model and the largest AIC and BIC values are the Poisson Regression. However, this model is less appropriate for use in this study because of the overdispersion conditions in the regression and there are spatial effects on the data. Therefore, the more appropriate models to use are the GWGPR and GWNBR models. Based on Table 8, the AIC and BIC values of the GWGPR and GWNBR models are not significantly different. Although both models yield comparable fit statistics, as indicated by nearly identical AIC and BIC values. GWGPR exhibits a higher sensitivity to regional variability, as evidenced by the formation of four distinct provincial clusters based on significant predictor variables, compared to only two clusters identified by the GWNBR model. This suggests that GWGPR may offer a more nuanced understanding of spatial effects in epidemiological data.

Variables X_2, X_4, X_6, X_{10} in several areas have significant differences in the partial test of the GWGPR and GWNBR models. Among them are variable X_2 (average annual temperature) in the provinces of North Kalimantan, North Sulawesi, Gorontalo, and North Maluku; variable X_4 (average wind speed) in North Kalimantan Province; variable X_6 (average sunshine) in North Kalimantan Province; variable X_{10} (percentage of poor population) in North Kalimantan and Central Kalimantan Provinces.

In several regions of Kalimantan Island, certain variables were found to be statistically insignificant, including the percentage of the population living in poverty (x_{10}). This can be attributed to the relatively low poverty rates in Kalimantan compared to other regions in Indonesia. In 2021, the national average poverty rate in Indonesia was 10.42%, while Central Kalimantan and North Kalimantan recorded significantly lower rates at 5.16% and 6.83%, respectively. Additionally, the number of tuberculosis (TB) cases across all provinces in Kalimantan was relatively low compared to the national average. The national average TB case count was 13.036, whereas the highest number of TB cases in Kalimantan was recorded in West Kalimantan (8,067 cases), and the lowest in Central Kalimantan (3,193 cases). Furthermore, other environmental variables such as average annual temperature (x_2), average humidity (x_3), average wind speed (x_4), and average solar radiation (x_6) were also found to be insignificant in several regions. These variables were generally associated with provinces that exhibited TB case counts below the national average, as reflected in the regional groupings based on the estimated coefficient values. This indicates that in areas with relatively low TB incidence, such variables may not play a dominant role in influencing TB prevalence.

This is an interesting finding from this final assignment, from several provinces producing the same variables, but in several other provinces, there are different significant variables. This is interesting to study further, where both models are theoretically quite significant in producing almost the same model goodness of fit test. However the empirical data found by the author is not enough evidence to support this theory.

The interpretation of the GWGPR and GWNBR models is based on spatially distributed estimates, visualized through thematic maps to highlight provincial-level patterns and characteristics. These spatial visualizations serve as references for policy-making, particularly in identifying variables with a higher estimated risk of influencing tuberculosis (TB) cases.

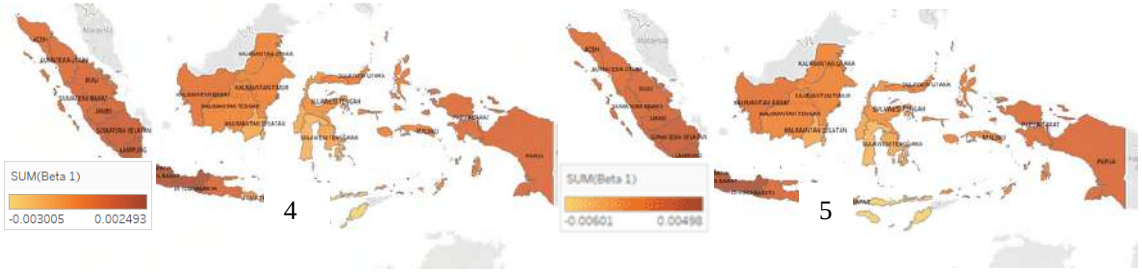


Figure 5. coefficient of the percentage variable of smokers in the GWGPR model
Figure 6. coefficient of the percentage variable of smokers in the GWNBR model

As shown in figure 5 & 6, darker brown shades on the maps represent provinces with higher estimated parameter values, while lighter shades indicate lower estimates. This color gradient underscores the spatial variability of regression coefficients across geographic regions. The variable representing the percentage of smokers shows that the darkest brown areas are primarily located in Sumatra and parts of Java, with the highest estimate found in the Special Region of Yogyakarta (DIY). The estimated coefficients for the smoking percentage in DIY are 0.002493 using GWGPR and 0.00498 using GWNBR. These estimates imply that a 1% increase in smoking prevalence is associated with an increase in TB cases by a factor of $\exp(0.002493) \approx 1.0025$ under GWGPR and

exp(0.00498) ≈ 1.005 under GWNBR, assuming other covariates remain constant. This result highlights the practical importance of tobacco control in mitigating TB incidence in high-risk provinces.

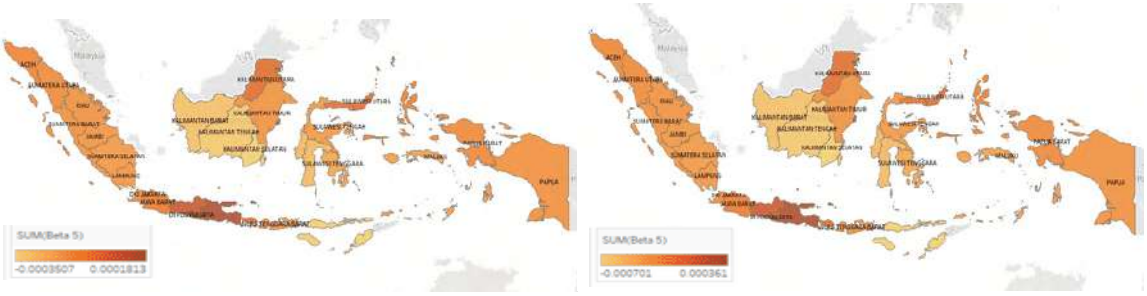


Figure 7. coefficient of the number of rainy days in the GWGPR model
Figure 8. coefficient of the number of rainy days in the GWNBR model

In the contrary, the variable for the number of rainy days (Figures 7 & 8) is predominantly light brown in areas such as Nusa Tenggara and parts of Kalimantan, indicating lower estimated values. The lowest estimates are found in East Nusa Tenggara, with coefficient values of -0.0003507 (GWGPR) and -0.000701 (GWNBR). These results suggest that an increase of one day in annual rainfall correlates with a decrease in TB cases by a factor of exp(-0.0003507) ≈ 0.9996 (GWGPR) and exp(-0.000701) ≈ 0.9993 (GWNBR), respectively—approximately equivalent to no meaningful change. These small effect sizes indicate that although the relationship is statistically detectable, the practical impact is minimal under these models.

Overall, the spatial heterogeneity captured by GWGPR and GWNBR models reveals that certain covariates exhibit both statistically and practically significant impacts on TB incidence, and that these effects vary geographically. These findings underscore the value of geographically weighted models in informing targeted, region-specific public health interventions.

This study is limited to data from the year 2021, and as such, the findings may not be generalizable to other time periods due to potential temporal variations in the influencing factors and disease patterns. Future research could explore the use of geographically weighted regression models that incorporate a broader range of social determinants, such as education level, housing conditions, and access to healthcare services, to enhance understanding of the spatial dynamics influencing TB incidence.

CONCLUSION

Based on the analysis conducted, several important conclusions can be drawn. The GWGPR (Geographically Weighted Generalized Poisson Regression) and GWNBR (Geographically Weighted Negative Binomial Regression) models are both applicable to modelling TB case data in Indonesia. The comparison of model fit statistics, particularly the AIC and BIC values, indicates no substantial difference between the two methods, suggesting that either model may be appropriately used to account for spatial variation in TB cases. However, GWGPR exhibits a higher sensitivity to regional variability, this suggests that GWGPR may offer a more nuanced understanding of spatial effects in epidemiological data

The analysis also revealed that the factors influencing TB incidence vary across provinces due to spatial heterogeneity and spatial dependency. Nevertheless, certain variables consistently showed a significant effect across all provinces in both the GWGPR and GWNBR models. These include the percentage of smokers, average annual humidity, number of rainy days per year, the percentage of the population reporting health complaints, and the percentage of TB detection and treatment coverage.

While the findings provide valuable insights into the spatial dynamics of TB in Indonesia, the study is limited by the use of cross-sectional data from 2021, which may not capture temporal trends or be generalizable across other years. Future research should consider longitudinal data and incorporate a broader set of social and environmental determinants. Furthermore, external validation with data from other time periods or regions would enhance the robustness and generalizability of the model outcomes.

REFERENCE

[1] N. Delvia, M. Mustafid, and H. Yasin, "Geographically Weighted Negative Binomial Regression Untuk Menangani Overdispersi Pada Jumlah Penduduk Miskin," *Jurnal Gaussian*, vol. 10, no. 4, pp. 532–543, 2021.

[2] M. Ririanti and R. D. Guntur, "Generalized Poisson Regression Modeling on the Number of Infant Deaths in East Nusa Tenggara Province in 2022," vol. 17, no. 2, pp. 779–788, 2024.

[3] F. Fitriani and M. Athoillah, "Penulisan Karya Tulis Ilmiah Bidang Sains Data," 2024.

[4] R. R. Hocking, *Methods and Applications of Linear Models*, Second Edi., vol. 39, no. 3. New York: John Wiley & Sons, Inc., 1997.

[5] R. K. Putri, M. Athoillah, and A. Haqiqiyah, "Analisis Faktor Yang Mempengaruhi Ketepatan Kelulusan Mahasiswa Dengan Algoritma Regresi Linear," *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 5, no. 2, pp. 671–680, 2024.

[6] R. E. Caraka and H. Yasin, *Geographically Weighted Regression Analysis*. Yogyakarta: Mobius, 2017.

[7] P. Mccullagh and J. A. Nelder, *Generalized Linear Models*, Second Edi., vol. 28, no. 1. London: Chapman and Hall, 1989.

[8] F. Famoye, "Restricted generalized poisson regression model," *Communications in Statistics - Theory and Methods*, vol. 22, no. 5, pp. 1335–1354, 2014.

[9] W. Greene, "Functional forms for the negative binomial model for count data," *Economics Letters*, vol. 99, no. 3, pp. 585–590, 2008.

[10] Breusch and Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979.

[11] B. D. Kifana and M. Abdurohman, "Great Circle Distance Methode for Improving Operational Control System Based on GPS Tracking System," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 4, no. 04, pp. 647–662, 2012.

[12] Ö. G. Esenbuğa, A. Akoğuz, E. Çolak, B. Varol, and B. Erol, "Comparison of Principal Geodetic Distance Calculation Methods for Automated Province Assignment in Turkey,"

16th International Multidisciplinary Scientific GeoConference SGEM2016, Informatics, Geoinformatics and Remote Sensing, vol. 2, no. June, 2016.

[13] S. Alfiani and P. R. Arum, "Pemodelan Pertumbuhan Ekonomi di Jawa Barat Menggunakan Metode Geographically Weighted Panel Regression," *J statistika*, vol. 15, no. 2, pp. 219–227, 2022.

[14] A. Stewart Fotheringham, Chris Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. UK: John Wiley & Sons Ltd, 2015.

[15] A. Djuraidah, H. Djalihu, and A. M. Soleh, "Mixed Geographically and Temporally Weighted Autoregressive to Modeling the Levels of Poverty Population in Java in 2012-2018," *Journal of Physics: Conference Series*, vol. 1863, no. 1, 2021.

[16] M. Y. Darsyah, "Pemodelan Geographically Weighted Negative Binomial Regression (GWNBR) pada Kasus Malaria di Indonesia," *Jurnal Litbang Edusaintech*, vol. 2, no. 2, pp. 149–164, 2021.

[17] D. J. Briggs *et al.*, "Mapping urban air pollution using gis: A regression-based approach," *International Journal of Geographical Information Science*, vol. 11, no. 7, pp. 699–718, 1997.

[18] S. D. Oluwajana, P. Y. Park, and T. Cavalho, "Macro-level collision prediction using geographically weighted negative binomial regression," *Journal of Transportation Safety and Security*, vol. 14, no. 7, pp. 1085–1120, 2022.

[19] Sulistyo, R. D. Sagala, D. Asmoro, S. N. Rahma, B. Alisjahbana, and R. C. Koesoemadinata, *Annual Report National TB Program 2022*. Unicef Indonesia, 2022.

Exploring Association of Household Conditions and Community Behavior in Flood Events in Banjarbaru Using Apriori Method

Rifqi Aulya Rahman(1), Yuana Sukmawaty(2), Selvi Annisa(3)

[1,2,3]Statistics Study Program, Faculty of Mathematics and Natural Sciences,
Lambung Mangkurat University

Achmad Yani Street Km. 36, Banjarbaru, 70714, South Kalimantan, Indonesia
e-mail: rrahman@ulm.ac.id(1), yuana_s@ulm.ac.id(2) , selvi.annisa@ulm.ac.id(3)

ABSTRAK

Banjir di Banjarbaru sebagian besar telah dikaji dari perspektif penyebab alami dan fisik, seperti curah hujan dan topografi wilayah. Sementara itu, hubungan antara kondisi rumah tangga dan perilaku masyarakat saat banjir jarang dieksplorasi secara kuantitatif. Penelitian ini mengisi kesenjangan tersebut dengan menerapkan algoritma Apriori pada data kuesioner dari rumah tangga terdampak banjir untuk menemukan aturan asosiasi. Hasil penelitian menunjukkan bahwa gangguan pada mata pencaharian selama banjir cenderung diikuti dengan penurunan pendapatan, sementara banjir yang berlangsung lebih dari satu hari umumnya memicu evakuasi anggota keluarga dan mendorong pemerintah menyediakan tempat penampungan sementara. Aturan-aturan ini mengimplikasikan bahwa kebijakan mitigasi banjir sebaiknya memprioritaskan sistem peringatan dini, penyediaan tempat penampungan sejak awal, serta bantuan ekonomi yang guna meningkatkan ketahanan rumah tangga terdampak.
Kata kunci: Apriori; Aturan Asosiasi; Banjir; Eksplorasi

ABSTRACT

Floods in Banjarbaru have mostly been studied from the perspective of natural and physical causes, such as rainfall and the region's topography. Meanwhile, the association between household conditions and community behavior during floods is rarely explored quantitatively. This research aims to fill that gap by applying the Apriori algorithm to questionnaire data from flood-affected households to find association rules. The study found that disruptions in livelihoods during floods tend to be followed by a decrease in income, while floods lasting more than one day generally trigger the evacuation of family members and prompt the government to provide temporary shelters. These key rules imply that flood mitigation policies should prioritize early warning systems, pre-positioning of shelter facilities, and targeted economic assistance to enhance the resilience of affected households.
Keywords: *Apriori; Association Rules; Flood; Exploration*

INTRODUCTION

Banjarbaru City is one of the areas in South Kalimantan Province that is prone to flooding [1], with more than 50% of its area categorized as vulnerable or somewhat vulnerable to floods based on vulnerability zoning [2]. The high intensity of rainfall based on climatological analysis [3], [4], changes in land use, and relatively flat topographical conditions are the main triggers for waterlogging at various points in the city [5]. As a result, floods have a significant impact on various aspects of life, ranging from infrastructure damage to health and socio-economic disruptions [6].

Besides natural factors and the physical structure of the city, the condition of the residential environment and community behavior also increase vulnerability to flooding. Areas such as Cempaka District experience more severe impacts due to the characteristics of slum residential environments and poor drainage systems [7]. Unplanned settlements along riverbanks often lack proper drainage systems, and the community tends to build without considering water conservation and green open spaces [8]. The habit of throwing garbage into drainage channels and covering the ground with cement drastically reduces water absorption, accelerating the occurrence of flooding [9]. In areas like Kemuning, floods are even considered a routine occurrence every rainy season by the residents, indicating low preparedness and effectiveness of early warning systems [10]. As a result, this area has become a top priority for flood management [11].

Most of the previous flood research in Banjarbaru still uses an analytical approach that focuses on the causes of floods. There are rarely any studies that specifically apply an exploratory approach in the context of flood-prone settlements, where this approach has the advantages in identifying hidden relationships, especially those related to the condition of the home environment, the behavior of riverside communities, and the impact of floods. Exploratory concepts like Apriori are effective in uncovering hidden patterns that are not detected by conventional analysis. The choice of association rule mining over conventional methods like logistic regression or correlation matrices is due to its ability to analyze complex, multidimensional categorical data without assumptions of linearity or predictor independence [12].

Apriori works by identifying combinations of items that frequently appear in the dataset through an iterative process, starting from single items to more complex combinations, filtering itemsets based on a specified minimum threshold [13]. Association rules are widely used to analyze transaction data, but this method is also relevant in processing questionnaire data. Several studies have shown that association rules can reveal interesting patterns in respondents' responses that are complex and multidimensional. For example, this method is effective in extracting relationships between questionnaire variables using the Apriori algorithm [14].

This research is a preliminary study that uses the Apriori algorithm to identify correlation patterns between household environmental conditions and community behavior in flood-prone areas of Banjarbaru City, specifically in the Cempaka and Banjarbaru Selatan sub-districts, based on questionnaire data. This research is based on the question: What association patterns emerge between household environmental conditions and community behavior during flood events in Banjarbaru City? To answer this question, the objective of this study is to discover association rules between variables in the data and to understand how these relationships reflect the environmental conditions and community behavior in flood-prone areas. The results are expected to offer new insights into community flood vulnerability and support more effective, participatory mitigation strategies.

METHOD

Data were obtained from a survey using questionnaires, involving 102 households in flood-affected areas along riverbanks in Banjarbaru City, South Kalimantan Province, conducted in August 2023 across Banjarbaru Selatan and Cempaka sub-districts. The questionnaire includes the house's coordinates and 26 topics of questions covering the condition of the affected houses, preventive measures before the flood, the situation during the flood, and the post-flood condition, with mutually exclusive answer categories for each question. The data were analyzed without respondent removal, outlier identification, or normalization, assuming all data were valid, relevant, and appropriate for the categorical algorithm used. Minor missing data were retained, as in transaction structures some items may not appear in a transaction. Response bias was minimized through interviews and direct household observations. The data is transformed into a transaction format, as illustrated in Figure 1. Each row in the data is represented as a single transaction that reflects the unique characteristics of a house, namely the combination of categories from the 26 questions. The transformation defines answers as items, with each house's category combination forming a transaction. This resulted in 102 transactions and 71 distinct item types.

House	Question 1	Question 2	…	Question 26
1	Category	Category	…	Category
2	Category	Category	…	Category
…	…	…	…	…
102	Category	Category	…	Category

⬇

ID	Transaction
1	{Category, Category, …, Category}
2	{Category, Category, …, Category}
…	…
102	{Category, Category, …, Category}

Figure 1. Transformation of Questionnaire Data to Transactions

The analysis was conducted using R version 4.4.2 with the association rule library called arules version 1.7-9 [15]. The process begins with exploring the profile data of houses affected by the flood. The coordinate points are visualized in the form of a map to show the location and distribution of houses, while profile variables such as length of residence, number of family members, and water height are displayed through graphs to illustrate the proportions and relationships of characteristics between variables. Next, association rule mining is applied to mine patterns among these characteristics. Association rules are a data mining technique for finding relationships between items in transaction data. While often used in retail, they can also analyze questionnaire data by transforming responses into transaction form [12]. Association rules describe relationships in if-then form ("⇒"), meaning if a transaction contains a certain itemset (antecedent), it is likely to contain another itemset (consequent). The mining process has two stages: identifying frequent itemsets and generating rules from them. The classic algorithm for this is Apriori [13], which is described as follows and illustrated in Figure 2.

1. Preparing input of transactions D.

2. Setting the support (supp) and confidence (conf) thresholds. Support that is too high results in few rules, whereas support that is too low results in too many rules. High confidence indicates strong rules, while low confidence results in less meaningful rules [16]. The rules generated by this research are divided into two types:

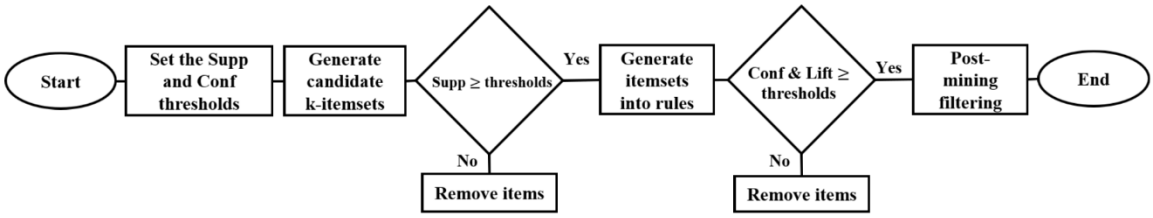


Figure 2. Apriori process

   a). General rules by setting high support parameters (minsup 60%) and strong confidence (minconf 80%). The 60% minimum support for general rules refers to thresholds commonly used in studies [17], with confidence based on [18].
   b). Unique rules by setting low support parameters (minsup 10%) and strong confidence (minimum 90%). The minimum 10% support follows [19], with confidence set at the upper bound of [18] range to ensure stronger reliability for rare associations.

3. Finding frequent 1-itemset. Keep the items when $supp \geq minsup$ which is calculated as

$$Supp(A) = \frac{Number\ of\ transactions\ containing\ A}{Total\ number\ of\ transactions}. \quad (1)$$

   Iteratively generate candidate k-itemsets from frequent (k−1)-itemsets and repeat until no new frequent itemsets remain.

4. Form rules from each frequent itemset consisting of ≥ 2 items. Keep the rules when $conf \geq minconf$ which is calculated as

$$Conf\ (A \Rightarrow B) = \frac{Supp\ (A \cup B)}{Supp\ (A)}. \quad (2)$$

5. Evaluate each rule by also calculating the lift ratio. Keep the rules when lift ≥ 1, where

$$Lift\ (A \Rightarrow B) = \frac{Conf\ (A \Rightarrow B)}{Supp\ (B)}. \quad (3)$$

   If $Lift > 1$, then A and B tend to appear together more often than expected if they are independent, indicating a positive relationship.

6. The previous steps are applied to each type of rule (general and unique).Then, the remaining rules for each type are filtered by removing those that are redundant, tautological, or do not provide additional meaning, and selecting rules with unique consequents. Leverage as

$$Lev(A \Rightarrow B) = Supp\ (A \cup B) - Supp\ (A) \times Supp\ (B) \quad (4)$$

   and conviction as

$$Conv(A \Rightarrow B) = \frac{1 - Supp\ (B)}{1 - Conf\ (A \Rightarrow B)} \quad (5)$$

   are calculated as evaluations of quality and reliability.

RESULT AND DISCUSSION

A total of 102 houses were affected by the flood, including 53 houses in the Banjarbaru Selatan District and 49 in the Cempaka District. The red dots in Figure 3 indicate the affected houses in both sub-districts. In Banjarbaru Selatan, 49 houses were affected by the overflow of the Kemuning River, while four others were damaged due to flooding in low-lying areas caused by heavy rain, impacting a riverbank area approximately 2 km long. The affected houses in Banjarbaru Selatan are concentrated along the river and near the main road network, as shown in the inset map. In Cempaka, there were 44 incidents due to rain and river overflow along Mistar Cokrokusumo Street, and five others were affected due to poor water absorption in Gunung Kupang Village. The distribution of the affected houses in Cempaka, as seen in the inset map, follows the course of the river and the main roads, particularly in areas with limited drainage capacity.

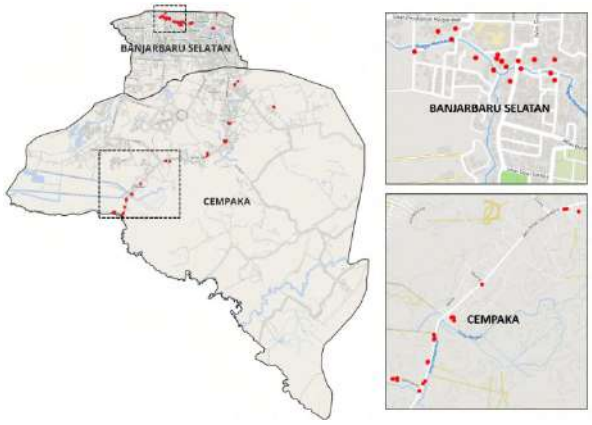


Figure 3. Distribution of flood-affected houses

Figure 4 shows that houses inhabited for more than 25 years are most affected by floods, especially houses with 1-3 and 4-5 family members. A similar pattern is observed in houses aged 15-25 years, dominated by families of 4-5 members. Meanwhile, houses inhabited for less than 15 years tend to be less affected and are generally inhabited by families with a maximum of 5 members. These findings indicate that the flood-affected areas are dominated by older residential neighborhoods with medium family density.

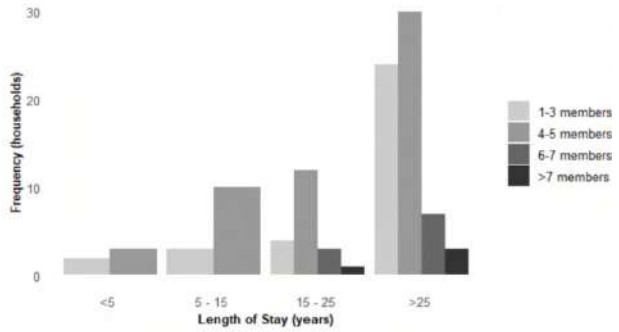


Figure 4. Distribution of household members by length of stay

About 80 events occur within a short duration, specifically less than one day. This is evident in the mosaic plot in Figure 5. In floods lasting less than one day, the most frequently recorded water height falls within the 0–0.5 meter category, with a proportion of 31.4%. Meanwhile, in floods lasting more than one day, the highest proportion is recorded in the water height category of more than 1 meter, at 16.7%. This condition primarily occurs in the Bangkal Village area, Cempaka District, particularly around Jalan Mistar Cokrokusumo, where nearly half of the surveyed houses are located. These findings indicate a relatively high risk of flooding in the area, both during short-duration and long-duration floods, which could potentially inundate houses and disrupt transportation activities in the vicinity. Looking at the frequency variable in the second mosaic plot, floods with a frequency of more than five times tend to occur with higher water levels. In this category, the highest proportion is recorded at water levels of more than 1 meter, which is 28.2%. Areas with water levels of more than 1 meter are spread across various points but are dominated in Bangkal Village, Cempaka District. These findings indicate that areas more frequently affected by floods tend to experience more severe flooding in terms of water height. This suggests a particular vulnerability in the Bangkal Village area to recurring major floods, as well as an increased potential for environmental damage, infrastructure damage, and threats to public safety.

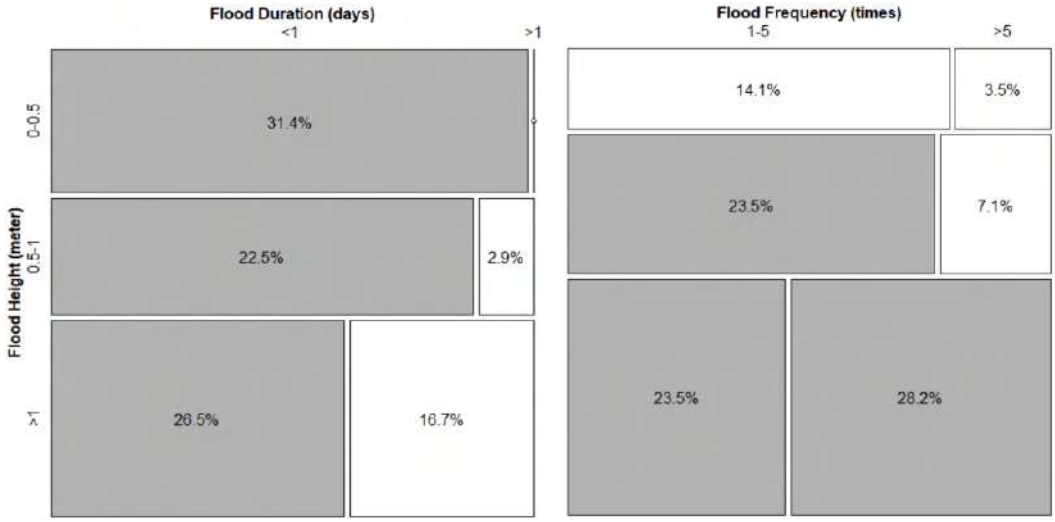


Figure 5. Distribution of proportions of flood height, duration, and frequency

The data dimension is in the form of a matrix with 102 transactions and 71 items. Figure 6 displays the transaction-item matrix, which has a density of 36% (relatively dense). The number of response items per transaction varies between 24 and 26 items, with an average of 25.56 items across all transactions. The range indicates that the majority of households responded to each question, showing their engagement in the survey.

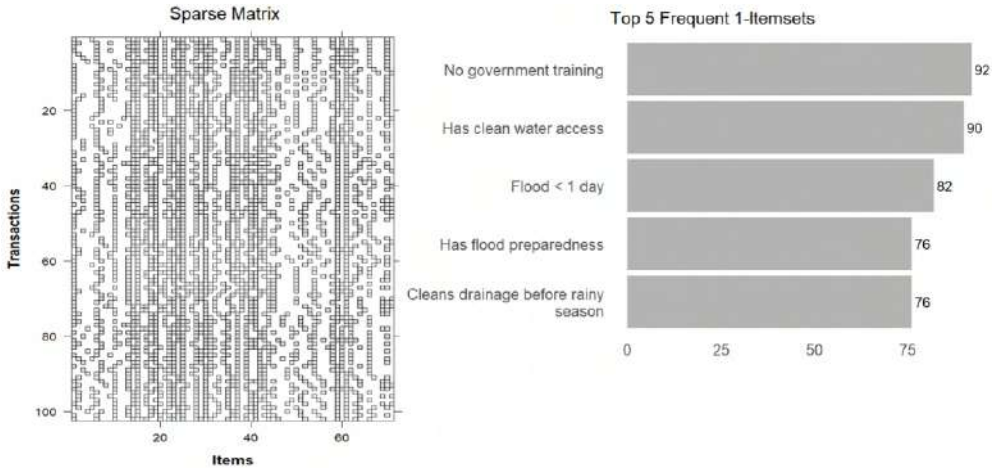


Figure 6. Sparse matrix and Top 5 Frequent 1-itemsets

Frequent itemsets in Figure 6 show an overview of the individual items that appear. Items with the highest frequency found, such as "No emergency response training provided by the government," "Houses facilitated with clean water," and "Flood duration less than a day," have high occurrences in over 80 houses. These items reflect common issues experienced by most households in facing floods. The next frequency search identifies frequent item pairs (2-itemsets) based on the general rule parameters (min support 60%). Unique rule itemsets are not shown due to limited results. Based on Table 1, the three itemsets with the highest frequency are {Houses facilitated with clean water, No government training for emergency response}, {Houses facilitated with clean water; Flood duration less than a day}, and {No government training for emergency response, Flood duration less than a day}. These three pairs of conditions appeared in more than 70 houses and were often found together in the observed houses.

Most houses have access to clean water and are located in areas that only experience short-duration floods, but there is still a lack of disaster preparedness training from the government. The generalization of these findings indicates that although basic infrastructure such as access to clean water is adequate and most affected areas only experience flooding for a short duration, the aspect of community preparedness for disasters remains very weak due to the lack of government intervention in the form of emergency response training. The policy relevance that can be drawn is the need for the government to integrate disaster preparedness training programs into flood mitigation strategies, especially in the focus areas of Kemuning Village, which are physically quite prepared but still socially vulnerable, so that the community's response to floods can be faster, more effective, and minimize further risks for the local community.

Table 1. Top 3 Frequent 2-Itemset

Itemset	Frequency
{House has access to clean water, There is no government-provided training for emergency response}	81
{House has access to clean water; Flood duration is less than one day}	75
{There is no government-provided training for emergency response, Flood duration is less than one day}	74

Association rules were formed from frequent 2-itemsets with a maximum rule length of two in if-then form, divided into general rules that reflect common community behavior patterns and unique rules according to their thresholds. The rules were then filtered by selecting those with a lift value greater than one. Redundant rules, such as reverse rules, subsets, or tautological rules, were removed during the post-mining filtering stage to improve clarity and relevance, with a focus on rules with unique consequents. Leverage and conviction values were calculated and reviewed as additional indicators of rule quality and reliability. The results are presented in Table 2.

The first general rule is supported by a support of 65.6% and a confidence of 98.5%. This indicates that the majority of households in the data experience this condition, and livelihood disruptions are almost always accompanied by a decrease in income. The lift value of 1.34, leverage of 0.16, and conviction of 18.00 indicate a strong and reasonable positive relationship, with a low probability of occurring by chance. The rule is meaningful but may have inflated confidence due to the consequent's dominance. Next, the second general rule shows that when there is drainage channel cleaning activity before the rainy season, the flood duration is generally less than one day. This rule is supported by a support value of 64.7% and a confidence value of 86.8%. This indicates that most households in the data experience shorter flood durations when there is drainage channel cleaning activity. The lift value of 1.08, leverage of 0.04, and conviction of 1.49 indicate a positive relationship, not as strong as the first rule, and it could be coincidental. This means that some instances where flooding lasts less than a day may not be entirely due to drainage cleaning, but rather there are other factors at play. Furthermore, the high confidence may partly result from the high prevalence of short flood duration in the dataset, so the potential inflation of confidence should be considered when interpreting this rule.

Table 2. General and Unique Rules

General Rules	**Supp**	**Conf**	**Lift**	**Lev**	**Conv**
Livelihood is disrupted during floods ⇒ Income decreases during floods*	65.6%	98.5%	1.34	0.16	18.00
There are drainage cleaning activities before the rainy season ⇒ Flood duration is less than one day*	64.7%	86.8%	1.08	0.04	1.49
Unique Rules					
Flood duration is more than one day ⇒ Family members evacuate during the disaster*	19.6%	100%	2.21	0.10	Inf
Flood duration is more than one day ⇒ Local government provides temporary shelter for residents*	18.6%	95%	2.15	0.09	11.17

*Significant association between questions ($p < 0.01$, Fisher's exact test)

The decline in community income during floods is the result of various interconnected conditions. When floods occur, the main economic activities especially trade along the banks of the Kemuning River or Jalan Mistar Cokrokusumo are disrupted, often leading to loss of income. In areas such as Cempaka District, the risk of income loss sharply increases when floods occur amidst the high density of low-income residents [1], consistent with previous findings on the vulnerability of low-income urban communities during floods. This risk is exacerbated by the lack of emergency savings in many households, as revealed in the survey. The policy implications of

these findings emphasize the importance of supporting the economic sustainability of communities, such as emergency cash assistance, access to capital during disasters, and strengthening micro-enterprises. Next, the second general rule shows a positive relationship, although not strong. Field facts in the Cempaka area indicate that poor drainage is one of the causes of flooding, as also noted in earlier studies. The condition is worsened by blockages, river narrowing, and sedimentation due to infrequent drainage maintenance [9]. Routine channel cleaning is carried out as a preventive measure before the rainy season, but this is not enough to reduce the duration of flooding. The construction of supporting infrastructure, such as spillways, river normalization, and channel capacity enhancement, is necessary to accommodate extreme rainfall [9]. Moreover, measures like socializing flood-prone warning areas are needed to improve community flood preparedness. [11].

The first unique rule is supported by a confidence value of 100% and a lift of 2.21, indicating a very strong positive relationship between the duration of the flood and the family's decision to evacuate. A leverage value of 0.10 and infinite conviction further strengthen the evidence that there are no cases in the data where flooding for more than one day is not followed by family evacuation, making this relationship very strong and not coincidental. Meanwhile, the second unique rule is supported by a support value of 18.6% and a confidence of 95%, which means that in most cases, the government does provide temporary housing assistance when prolonged flooding occurs. The lift and conviction indicate a significant positive, although not as strong as the first unique rule.

The first unique rule is supported by the fact that floods often recur every year in certain areas, particularly on Jalan Mistar Cokrokusumo, which is prone to inundation due to heavy rainfall [9]. As a result, the floods last a long time and fill the entire house, forcing residents to evacuate themselves to avoid health risks, property damage, and difficulties in accessing basic needs [10], as documented in local disaster reports. The chosen evacuation sites are usually family homes or mosques. Residents transport their belongings to evacuation sites until the flood recedes, then return to clean their homes. Subsequently, the second rule highlights the active role of the government in providing temporary shelters so that residents can take refuge until the flood recedes, reflecting standard disaster response practices. Flood socialization conducted by the government also urges residents to remain on alert, as evidence of their readiness to evacuate if the flood persists for an extended period [10]. Collaboration between the community and the government in disaster mitigation is crucial to ensure the availability of emergency facilities that safeguard the safety and health of residents during the disaster.

CONCLUSION

This study found association patterns, such as livelihood disruptions followed by decreased income, and floods lasting more than one day that triggered family evacuations and the provision of temporary shelters by the government. These findings highlight the need to strengthen early warning systems, emergency response training, and economic support for affected communities. Nevertheless, this study has limitations, particularly in the association rules produced, where the dominance of certain categories could obscure the true relationships. Future analyses should apply or algorithms methods to minimize this effect to produce more accurate and policy-relevant results.

REFERENCE

[1] BNPB, "Dokumen Kajian Risiko Bencana Nasional Provinsi Kalimantan Selatan 2022 -

2026," 2021. [Online]. Available: https://inarisk.bnpb.go.id/pdf/Kalimantan Selatan/Dokumen KRB Prov. Kalimantan Selatan_final draft.pdf.

[2] E. Setiawan, F. Yusran, F. Razie, and R. Mustika, "Zonasi Tingkat Kerentanan Banjir di Kota Banjarbaru Kalimantan Selatan," *EnviroScienteae*, vol. 11, pp. 136–142, 2015.

[3] E. Sofia and M. Amalia, "Analisis Karakteristik Curah Hujan di Kota Banjarbaru Berdasarkan Data Stasiun Klimatologi Banjarbaru," *J. Teknol. Berkelanjutan*, vol. 10, no. 01, pp. 36–41, Apr. 2021, doi: 10.20527/jtb.v10i01.192.

[4] Y. Farida *et al.*, "Modeling the Flood Disaster in South Kalimantan Using Geographically Weighted Regression and Mixed Geographically Weighted Regression," *ITM Web Conf.*, vol. 58, p. 04004, Jan. 2024, doi: 10.1051/itmconf/20245804004.

[5] R. N. Adi and E. Savitri, "The effect of land cover changes on the 2021 flood in the Barito watershed, South Kalimantan," in *IOP Conference Series: Earth and Environmental Science*, Nov. 2022, vol. 1109, no. 1, p. 012017, doi: 10.1088/1755-1315/1109/1/012017.

[6] N. Wahyuningsih, M. Ruslan, and B. Badaruddin, "Analisis Penyebab Banjir di Kecamatan Cempaka Kota Banjarbaru Provinsi Kalimantan Selatan," *J. Sylva Sci.*, vol. 5, no. 6, p. 944, Dec. 2022, doi: 10.20527/jss.v5i6.7138.

[7] Novitasari, Nurfansyah, H. Kurdi, and E. Nashrullah, "Peningkatan Kapasitas Masyarakat Dalam Kesiapsiagaan Bencana Banjir Di Kawasan Kumuh Cempaka," in *Pro Sejahtera (Prosiding Seminar Nasional Pengabdian kepada Masyarakat)*, 2023, vol. 5, pp. 137–144.

[8] N. Novitasari and H. Kurdi, "Flood Mitigation in Banjar Regency, South Kalimantan, Indonesia in 2021: Between Hydro-meteorological factor and Anthropogenic factor," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 999, no. 1, p. 012010, Mar. 2022, doi: 10.1088/1755-1315/999/1/012010.

[9] I. Subri and A. Mursadin, "Study of Flood Management Due to Rainfall on Drainage Network System in Residential Area (Case Study : Cempaka Sub District , Banjarbaru City , South Borneo)," *Am. J. Eng. Res.*, vol. 10, no. 7, pp. 153–160, 2021.

[10] M. Kurniati, Periyadi, Ariansyah, E. Febriana, and H. Rahmadina, "Penyebarluasan Informasi Peringatan Dini Bencana Banjir di Kantor Kelurahan Kemuning Banjarbaru," *MENGABDI J. Has. Kegiat. Bersama Masy.*, vol. 2, no. 2, pp. 20–26, Apr. 2024, doi: 10.61132/mengabdi.v2i2.492.

[11] R. Riduan, Jamiyaturrasyidah, R. Khair, and C. Abdi, "Sosialiasi Lokasi Peta Rawan Genangan Kota Banjarbaru Menggunakan Teknologi Pengindraan Jauh di PUPR Banjarbaru," *ILUNG J. Pengabdi. Inov. Lahan Basah Unggul*, vol. 1, no. 3, pp. 143–156, 2022, doi: doi.org/10.20527/ilung.v1i3.

[12] F. Peng, Y. Sun, Z. Chen, and J. Gao, "An Improved Apriori Algorithm for Association Rule Mining in Employability Analysis," *Teh. Vjesn. - Tech. Gaz.*, vol. 30, no. 5, pp. 1435–1442, Oct. 2023, doi: 10.17559/TV-20230327000481.

[13] V. Srinadh, "Evaluation of Apriori, FP growth and Eclat association rule mining algorithms," *Int. J. Health Sci. (Qassim).*, vol. 6, no. March, pp. 7475–7485, Apr. 2022, doi: 10.53730/ijhs.v6nS2.6729.

[14] K. Gulzar, M. Ayoob Memon, S. M. Mohsin, S. Aslam, S. M. A. Akber, and M. A. Nadeem, "An Efficient Healthcare Data Mining Approach Using Apriori Algorithm: A Case Study of Eye Disorders in Young Adults," *Information*, vol. 14, no. 4, p. 203, Mar. 2023, doi:

10.3390/info14040203.

[15] M. Ledmi, M. E. H. Souidi, M. Hahsler, A. Ledmi, and C. K. Mohamed, "Mining association rules for classification using frequent generator itemsets in arules package," *Int. J. Data Mining, Model. Manag.*, vol. 15, no. 2, pp. 203–221, 2023, doi: 10.1504/IJDMMM.2023.131399.

[16] E. Hikmawati, N. U. Maulidevi, and K. Surendro, "Minimum threshold determination method based on dataset characteristics in association rule mining," *J. Big Data*, vol. 8, no. 1, p. 146, Dec. 2021, doi: 10.1186/s40537-021-00538-3.

[17] Y. Zakur and L. Flaih, "Apriori Algorithm and Hybrid Apriori Algorithm in the Data Mining: A Comprehensive Review," *E3S Web Conf.*, vol. 448, p. 02021, Nov. 2023, doi: 10.1051/e3sconf/202344802021.

[18] N. Alangari and R. Alturki, "Association Rule Mining in Higher Education: A Case Study of Computer Science Students," in *EAI/Springer Innovations in Communication and Computing*, 2020, pp. 311–328.

[19] S. Darrab, D. Broneske, and G. Saake, "MaxRI: A method for discovering maximal rare itemsets," in *2021 4th International Conference on Data Science and Information Technology*, Jul. 2021, pp. 334–341, doi: 10.1145/3478905.3478972.

Implementation of K-Means Cluster for Districts or Cities in West Java Province Based on Unemployment Indicators

Alifa Nur Oktaviani [(1)], Atika Nurani Ambarwati [(2)]

Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang.

Jl. Prof. Dr. Hamka Km 01 No. 17 Tambakaji Ngaliyan
e-mail: alifaoktaviani28@gmail.com[(1)], atika.nurani@gmail.com[(2)]

ABSTRAK

Tingginya disparitas tingkat pengangguran terbuka antarwilayah di Provinsi Jawa Barat menjadi isu yang memerlukan pemetaan berdasarkan karakteristik sosial ekonomi daerah. Tujuan dari studi ini adalah untuk mengelompokkan kabupaten/kota di Jawa Barat berdasarkan tingkat pengangguran terbuka dan faktor-faktor yang memengaruhinya dengan menggunakan metode K-Means Cluster. Data yang digunakan merupakan data sekunder tahun 2023 yang diperoleh dari Badan Pusat Statistik. Uji asumsi dilakukan menggunakan KMO dan Bartlett's Test untuk memastikan kecukupan sampel dan kelayakan struktur data. Hasil analisis menunjukkan bahwa wilayah di Jawa Barat terbagi menjadi tiga cluster. Cluster 1 terdiri dari daerah dengan pembangunan tinggi namun tingkat pengangguran juga tinggi, seperti Kota Bandung dan Kabupaten Bekasi. Cluster 2 mencakup wilayah dengan kondisi sosial ekonomi menengah, seperti Kota Depok dan Kabupaten Bandung. Cluster 3 terdiri dari wilayah dengan tingkat pembangunan rendah, IPM rendah, dan kemiskinan tinggi, seperti Ciamis, Garut, dan Pangandaran. Temuan ini menunjukkan adanya ketimpangan yang signifikan antarwilayah dan dapat menjadi dasar bagi penyusunan kebijakan penanggulangan pengangguran yang lebih terarah dan berbasis wilayah.
Kata kunci: *K-Means*; Pengangguran Terbuka; Jawa Barat

ABSTRACT

The high disparity of open unemployment rate among regions in West Java Province is an issue that requires mapping based on regional socioeconomic characteristics. The purpose of this study is to group districts/cities in West Java based on the open unemployment rate and its influencing factors using the K-Means Cluster method. The data used is secondary data for the year 2023 obtained from the Central Bureau of Statistics. Assumption test was conducted using KMO and Bartlett's Test to ensure sample adequacy and feasibility of data structure. The results of the analysis show that the regions in West Java are divided into three clusters. Cluster 1 consists of regions with high development but also high unemployment rates, such as Bandung City and Bekasi Regency. Cluster 2 includes regions with medium socioeconomic conditions, such as Depok City and Bandung Regency. Cluster 3 consists of regions with low development, low HDI, and high poverty, such as Ciamis, Garut, and Pangandaran. These findings indicate the existence of significant inequality among regions and can serve as a basis for the formulation of more targeted and region-based unemployment reduction policies.
Keywords: *K-Means; Open Unemployment; West Java*

INTRODUCTION

Indonesia is among the nations that possess a significant labor market potential, which could serve as a catalyst for economic growth. Nevertheless, the challenges related to population and workforce in developing nations pose a hindrance to national development, as elevated unemployment levels impact the success of economic advancement in Indonesia [1]. Based on information from the Badan Pusat Statistika (2011), open unemployment refers to those in the workforce who are either without a job or actively seeking employment. This category comprises individuals who are searching for jobs, starting a business, those who have given up on finding work due to a belief that job opportunities are nonexistent, and people who have secured employment but have yet to begin their roles (unemployed). Consequently, the open unemployment rate serves as a measure that reflects the count of individuals who are not employed [2]. Unemployment is not only a social problem but also an economic problem, because unemployment can cause problems in the form of changes in the economic growth of developing countries such as Indonesia [3]. According to the BPS report from 2023, Indonesia's open unemployment rate stands at 5. 32%, translating to 7. 86 million individuals. While this percentage might seem low, it still warrants concern due to its significant impact on various facets of life in Indonesia, including social and economic dimensions. The root cause of unemployment can be traced to wage structures that do not align with the balance between labor demand and supply, leading to a decline in the mobility of the workforce [4].

Most of the population classified as open unemployment comes from West Java Province. In addition, the unemployment rate in West Java Province has not met the regional target stated in 2019-2023 Regional Medium-Term Development Plan. The success of a region's economic development can be hampered by a high unemployment rate, because unemployment is closely related to other economic variables, so it is a very important parameter [2].

In this research, we aim to group regencies or cities within West Java Province by utilizing the unemployment rate as a key indicator through K-Means method. A previous study by Nurul Nurjanah et al., entitled "Implementation of K-Means Clustering to Group Unemployment Rates," which resulted in the formation of three clusters, it was proven that the K-Means method can be a useful tool for identifying trends in data related to unemployment issues [5]. K-Means is a straightforward algorithm for grouping data that operates in an unsupervised manner, meaning it does not require labeled inputs [6]. We anticipate that this approach will lead to more effective outcomes in decreasing the unemployment figures in West Java. The K-Means algorithm is widely recognized for its utility in clustering analysis in studies, primarily due to its ability to deliver optimal clusters with rapid convergence [7]. Clustering is a method of analyzing data aimed at categorizing or forecasting the value of a target variable. It seeks to partition the dataset in a way that items within the same cluster are more similar to one another than to those in separate clusters [8], [9]. In overcoming the problem of unemployment, local governments can determine districts or cities that require special attention through grouping or clustering so as to create appropriate program adjustments in the future [10].

METHOD

In this research, the approach used K-Means Clustering. The information utilized in this research consists of secondary data related to unemployment metrics in the West Java Province. This data has been sourced from the BPS publication for West Java from the year 2023.

Research Variabels

The factors examined in this research consist of predictor variables (X). These predictor variables represent elements that may influence Open Unemployment in Central Java. Here are the variables identified:

Table 1. Research Variables

Variables	Descriptions
X1	Open Unemployment Rate
X2	Human Development Index
X3	Percentage of Poor Population
X4	GRDP
X5	Labor Force Participation Rate
X6	Minimum Wage

Data Processing

The process for categorizing regencies or cities in West Java Province according to the Unemployment Indicator includes these actions:

1. Descriptive analysis related to unemployment in West Java Province in 2023
2. Perform Assumption Test for cluster analysis, specifically:
   a. Evaluation of sample representativeness

   The Kaiser-Meyer-Olkin (KMO) test is commonly employed to determine whether a sample accurately reflects the population.This test shows how suitable the sample is, with values ranging from 0 to 1. A KMO score between 0. 5 and 1 means the sample effectively represents the overall population [11].

$$KMO = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (1)$$

Description:

r_{ij} = correlation coefficient between variable *i* and variable *j*

a_{ij} = partial correlation coefficient between variable *i* and variable *j*

p = total number of variables.

   b. Multicollinearity Examination

   Multicollinearity shows a strong connection among independent variables. In cluster analysis, it's important that variables are free from multicollinearity. To find multicollinearity, one can look at the correlation matrix. If the correlation coefficient between the variables is between 0. 8 and 1. 0, it indicates multicollinearity exists [12]

$$r_{(x,r)} = \frac{n(\sum_{i=1}^{n} X_1 Y_i) - (\sum_{i=1}^{n} Y_i)}{\sqrt{n(\sum_{i=1}^{n} X_i^2) - (\sum_{i=1}^{n} X_i)^2}} \quad (2)$$

Description:

$r =$ the correlation coefficient between variables X and Y

$n =$ number of research observations

3. Determination of k Value through the Elbow Method

The Elbow method is used to find the best clustering results by looking at the percentage of cluster members that form a right angle at a certain point. This point shows the best SSE value. The text includes the formula to calculate the SSE value [13]:

$$SSE = \sum_{i=1}^{n} \sum_{x_i \epsilon S_k} ||x_i - C_k||^2 \quad (3)$$

Description:

n : cluster

x_i : *i*-th data

S_k : *k*-th cluster element

C_k : average center in the *k*-th cluster/centroid

4. Clustering using K-Means Cluster

K-Means is a method for grouping data based on distance, using only numerical traits [14]. Its goal is to sort data into clusters based on a chosen k value, which decides how many groups there will be. Data points with similar features are placed in the same cluster. The main steps to follow when using the K-Means algorithm include:

a. Determining the cluster centroid value

b. Determining the Euclidean distance for every object in relation to the central point [15]. In calculating the Euclidean distance, the following equation can be used:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (4)$$

Descpription:

$d(x,y)$ = Euclidean distance

$x =$ cluster center data

$y =$ data on the attribute

$x_i =$ data at the cluster center to i

$y_i =$ data on every *i*-th

c. Classify objects based on their proximity to the center.

5. Interpretation of clustering results and cluster mapping

6. Make conclusions and recommendations

RESULT AND DISCUSSION

Descriptive Analytics

Before conducting a cluster analysis, it is necessary to examine the state of unemployment in West Java in 2023 to understand its characteristics.

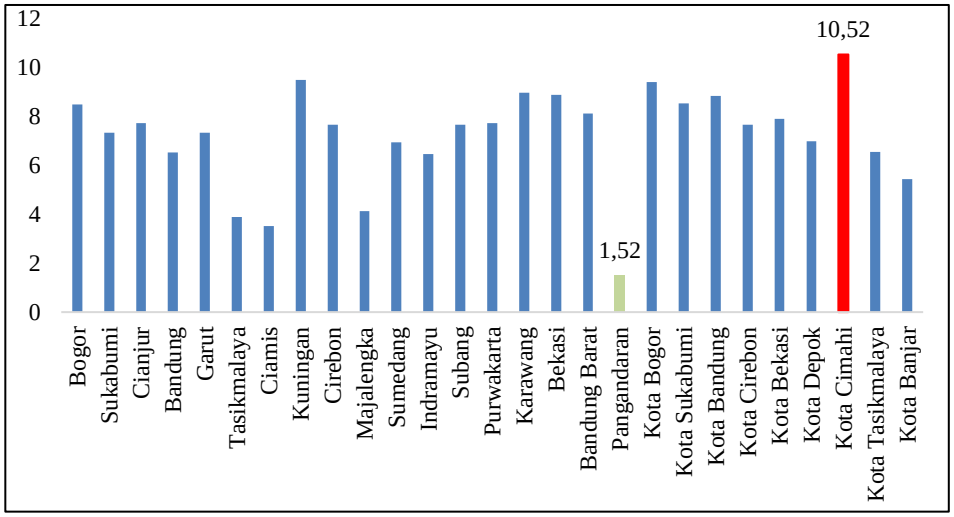


Figure 1. Open Unemployment Rate in West Java Province, 2023

The graph of the open unemployment rate in West Java Province in 2023 shows clear variations between districts and cities. Bekasi City recorded the highest open unemployment rate, possibly due to the high rate of urbanization and the mismatch between the available workforce and vacancies. Meanwhile, Pangandaran Regency shows the lowest unemployment rate, reflecting more stable labor market conditions. Other regions are at an intermediate level, illustrating the imbalance in labor distribution. This suggests the need for more focused employment policies in accordance with the character of each region, particularly to increase employment opportunities in urban areas.

K-Means Cluster Analysis Assumption

1. The sample represent the population

Table 2. KMO Value

KMO and Bartlett's Test	
Kaiser-Meyer-Olkin	0.701

Based on the table, the KMO value is greater than 0.5 and close to 1. Therefore, it can be concluded that the sample represents the population and can be analyzed using cluster analysis.

2. Multicollinearity test

Table 3. Matrix Correlation

Variabel	X1	X2	X3	X4	X5	X6
X1	1	0.408	-0.25	0.365	-0.553	0.547
X2	0.408	1	-0.774	0.29	-0.383	0.55
X3	-0.25	-0.774	1	-0.433	0.177	-0.634
X4	0.365	0.29	-0.433	1	-0.283	0.662

X5	-0.553	-0.383	0.177	-0.283	1	-0.404
X6	0.547	0.55	-0.634	0.663	-0.404	1

Based on the results of the correlation matrix above, it can be concluded that the data is free of multicollinearity. The correlation coefficient value between the variables in the matrix is less than 0.8, so the analysis can proceed to the next stage.

Using Elbow Method to Determine the Number of Clusters

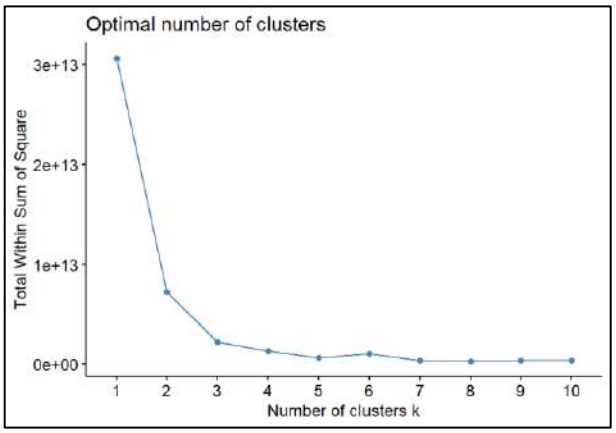


Figure 2. Elbow Method Plot

Based on Figure 2, it can be seen that the most optimal number of clusters is 3 as seen from the graph that shows a right-angled line. After the value of k is known, then the clustering stage can be carried out using the K-Means cluster method.

Clustering with the K-means Cluster

After figuring out how many clusters there are, the subsequent action is to find the centroid value for every cluster. The centroid value is derived from averaging the values in each cluster, beginning with a random selection for the initial cluster center value. The clustering procedure ends when the new initial cluster center value matches the previous one. Below is the initial center value that was calculated:

Table 4. Initial Cluster Centers

Initial Cluster Centers			
	Cluster		
	1	2	3
TPT	8.87	9.39	1.52
IPM	75.76	77.85	69.38
PPM	4.93	6.67	8.98
PDRB	99.00	20.78	11.68
TPAK	65.00	64.81	80.15
UM	97.92	83.97	10.57

Table 4 represents the initial phase of the data grouping procedure. Prior to conducting iterative tests, this information will be utilized to create three separate clusters.

Table 5. Iteration

Iteration History[a]			
	Change in Cluster Centers		
Iteration	1	2	3
1	23.483	17.898	18.243
2	0.000	2.305	1.855
3	0.000	0.000	0.000

Table 5 illustrates the iteration process in cluster grouping based on the initial table, where there are two iterations performed. In the first and second iterations, the resulting centeroids did not show significant differences. However, in the third iteration, the centeroids that appear show significant results. Therefore, all clusters have been formed, and the iteration process is stopped at the third iteration with a minimum distance of 71.451.

Table 6. Final Cluster Centers

Final Cluster Centers			
	Cluster		
	1	2	3
X1	8.78	7.90	6.14
X2	75.73	75.05	71.09
X3	6.01	6.91	9.82
X4	78.66	24.53	17.31
X5	64.90	66.67	67.83
X6	86.24	64.78	19.95

Table 5 shows the final results of the clustering process which consists of three clusters for each variable. The following is the interpretation of the K-Means cluster results above:

Cluster 1 : Cluster 1 shows the profile of a region with a relatively high open unemployment rate (8.78%) and a high Human Development Index (75.73). Regions in this cluster have the lowest percentage of people living in poverty (6.01%) and the highest Gross Regional Domestic Product (78.66) among the three clusters. Cluster 1's labor force participation rate is intermediate (64.9%), but this region has the highest minimum wage (86.24%). These characteristics suggest that Cluster 1 may represent an area with a strong economy and employment challenges, or an area with a dominant formal economy sector that offers high wages but has a certain level of unemployment.

Cluster 2 : Cluster 2 has a slightly lower open unemployment rate of 7. 90% and a Human Development Index of 75. 05 compared to Cluster 1. However, poverty levels rose above Cluster 1's 6. 91%, and the gross regional domestic product dropped by 24. 53%. The labor force participation rate in Cluster 2 is slightly higher than the 66. 67% of Cluster 1, and the minimum wage is moderate at 64. 78%. Overall,

Cluster 2 shows average socioeconomic conditions with moderate economic growth and increased poverty issues.

Cluster 3 : Cluster 3 has the lowest open unemployment rate (6.14%) and the lowest Human Development Index (71.09) of the three clusters. This cluster has the highest percentage of poor people (9.82%) and the lowest Gross Regional Domestic Product (17.31%). Cluster 3 has the highest labor force participation rate (67.83%), but the minimum wage is very low (19.95%). Overall, Cluster 3 indicates areas with high labor participation rates but low job quality, low wages, and high poverty rates.

Table 7. Districts/Cities in Each Cluster

Cluster 1	**Cluster 2**	**Cluster 3**
Kab. Bekasi	Bandung Barat	Ciamis
Kab.Bogor	Kab. Bandung	Cianjur
Karawang	Kab. Sukabumi	Kab. Cirebon
Kota Bandung	Kota Bekasi	Garut
	Kota Bogor	Indramayu
	Kota Cimahi	Kab. Tasikmalaya
	Kota Depok	Kota Banjar
	Purwakarta	Kota Cirebon
	Subang	Kota Sukabumi
	Sumedang	Kota Tasikmalaya
		Kuningan
		Majalengka
		Pangandaran

Table 7 shows the results of the K-means clustering analysis of the open unemployment rate in West Java. The regions are divided into three groups based on influencing factors. The first group consists of regions in the "good" category. These regions have high development indicators, such as HDI, GRDP, and minimum wage. However, the unemployment rate is still high due to urbanization pressure. Regions in this group include Bekasi Regency, Bogor Regency, Karawang, and Bandung City. The second group is in the medium category. These regions have fairly balanced socioeconomic conditions and are close to urban areas. Examples include Kota Depok, Kota Cimahi, and Kab. Bandung. The third group is in the low category and covers rural areas, such as Ciamis, Garut, and Pangandaran. These regions have low HDI and GRDP, and their level of development lags far behind the other two groups. This division illustrates regional inequality and is important for formulating targeted policies.

Table 8. Distance Between Clusters

Distance between final Cluster			
Cluster	1	2	3
1		58.273	90.606
2	58.273		45.714
3	90.606	45.714	

The space separating centroids, commonly known as centroid distance, measures how far apart the centroids of different clusters are from each other. Clustering is considered effective when this distance measurement is sufficiently large. A greater value indicates a larger gap between the clusters, which makes the distinctions between one cluster and another more apparent.

Table 9. Total Number Of Cluster Members

Number of Cases in each Cluster		
Cluster	1	4.000
	2	10.000
	3	13.000
Valid		27.000
Missing		0.000

This table shows the number of cases in each cluster from the clustering analysis. Cluster 1 has four cases, Cluster 2 has ten cases, and Cluster 3 has thirteen cases, making it the largest. There are 27 valid cases in total with no missing data. These numbers demonstrate that Cluster 3 is the largest, followed by Cluster 2; Cluster 1 is the smallest. This data is important for understanding the representation of each cluster in the sample and the relative significance of each group.

CONCLUSION

Analysis using the K-Means Cluster method divides West Java Province into three groups based on open unemployment rates and socioeconomic indicators. The first group consists of developed regions such as Bekasi Regency, Bogor Regency, Karawang, and Bandung City, which show high development despite having high open unemployment due to urbanization and lack of employment. The second group is medium-sized regions such as West Bandung, Bandung Regency, and other cities, which have stable socioeconomic conditions with unemployment and labor force participation in the medium category. The third group includes lagging regions such as Ciamis, Cianjur, and others, which show low development, high poverty rates, and low unemployment that may be influenced by the informal sector. These results show significant regional differences, so different policies are needed for each group to make unemployment reduction more effective and in line with regional characteristics.

REFERENCE

[1] A. A. Putrie and R. Sanjaya, "Pengelompokan Kabupaten/Kota Berdasarkan Indikator Tingkat Pengangguran Menggunakan Algoritma K-Means Clustering (Studi Kasus: Provinsi Jawa Barat)," vol. 2, no. 2, 2021, [Online]. Available: https://jabar.bps.go.id

[2] R. Ardian, U. Sultan, A. Tirtayasa, M. Syahputra, and D. Dermawan, "Pengaruh Pertumbuhan Ekonomi Terhadap Tingkat Pengangguran Terbuka Di Indonesia," vol. 1, no. 3, 2022.

[3] L. Luk, A. Mufida, and M. S. Nasir, "Analisis Dinamis Tingkat Pengangguran di Indonesia," 2021. [Online]. Available: https://economics.pubmedia.id/index.php/jmsd

[4] S. Pasuria and N. Triwahyuningtyas, "Pengaruh Angkatan Kerja, Pendidikan, Upah Minimum, Dan Produk Domestik Bruto Terhadap Pengangguran Di Indonesia," *Sibatik*

Journal: Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi, dan Pendidikan, vol. 1, no. 6, pp. 795–808, Apr. 2022, doi: 10.54443/sibatik.v1i6.94.

[5] N. Nurjanah, N. Suarna, W. Prihartono, and M. Kec Kesambi Kota Cirebon, "Implementasi K-Means Clustering Untuk Mengelompokan Tingkat Pengangguran," 2024.

[6] D. R. Ningsih, "Pengelompokan Produksi Daging Sapi Menurut Provinsi di Indonesia Tahun 2017-2022 dengan Menggunakan Metode K-Means," *ESTIMASI: Journal of Statistics and Its Application*, pp. 113–125, Jan. 2024, doi: 10.20956/ejsa.v5i1.26988.

[7] Muharni Sita dan Sigit Andriyanto, "Penerapan Metode K-Means Clustering Pada Data Tingkat Pengangguran Terbuka Tahun 2016-2018 Dan 2019-2021," *Penerapan Metode K-Means Clustering Pada Data Tingkat Pengangguran Terbuka Tahun 2016-2018 Dan 2019-2021*, vol. 22, pp. 90–99, 2022.

[8] F. A. Tanjung, A. P. Windarto, M. Fauzan, M. P. Studi, S. Informasi, and S. Tunas Bangsa, "Penerapan Metode K-Means Pada Pengelompokkan Pengangguran Di Indonesia," vol. 6, pp. 61–74, [Online]. Available: https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

[9] M. Athoillah, I. Irawan, M., and M. Imah, Elly, "Study Comparison of SVM-, K-NN- and Backpropagation-Based Classifier for Image Retrieval," *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 2015.

[10] J. Veronika, "Analisis Tingkat Pengangguran Di Kota Palopo Menggunakan Metode K-Means," 2024.

[11] D. R. Ningrat, D. Asih, I. Maruddani, and T. Wuryandari, "Analisis Cluster Dengan Algoritma K-Means Dan Fuzzy C-Means Clustering Untuk Pengelompokan Data Obligasi Korporasi," *Jurnal Gaussian*, vol. 5, no. 4, pp. 641–650, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[12] M. W. Talakua, Y. A. Lesnussa, and M. Y. Matdoan, "Analisis Klaster untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Pendidikan dengan Menggunakan Metode Ward," *Jurnal Statistika dan Aplikasinya*, vol. 5, no. 1, 2021.

[13] M. Desdianti, N. N. Debataraja, and E. Sulistianingsih, "Analisis K-Means Clustering Dengan Metode Elbow Pada Pengelompokan Tingkat Pengangguran Di Kalimantan Barat," 2024.

[14] M. Rais, R. Goejantoro, and S. Prangga, "Optimalisasi K-Means Cluster dengan Principal Component Analysis Optimization of K-Means Cluster with Principal Component Analysis on the Grouping of Districts/Cities on the Island of Kalimantan Based on Unemployment Rate Indicator," 2021.

[15] C. Auditiyah, "Pengelompokkan Daerah Rawan Demam Berdarah (DBD) di Jawa Timur Menggunakan Metode K-Means," *ESTIMASI: Journal of Statistics and Its Application*, pp. 205–215, Jul. 2024, doi: 10.20956/ejsa.v5i2.27091.

J Statistika Vol. 18, No. 1, (2025), Hal. 887 - 896 jurnal.unipasby.ac.id

Bonus-Malus Premium for Third Party Liability Insurance with Poisson-Lindley Distribution Claim Frequency and Exponential-Inverse Gamma Distribution Claim Severity

Bilqis Nur Rizkia [(1)], Aceng Komarudin Mutaqin [(2)]

[1,2]Department oStatistics, Faculty of Mathematics and Natural Science, Universitas Islam Bandung

Jln. Ranggagading No.8, Tamansari, Kecamatan Bandung Wetan, Kota Bandung, Jawa Barat
E-mail: bilqisnr76@gmail.com[(1)], aceng.k.mutaqin@gmail.com[(2)]

ABSTRACT

Asuransi adalah bentuk gotong royong yang memberikan perlindungan terhadap risiko tak terduga. Dengan asuransi, seseorang merasa lebih tenang menghadapi kemungkinan buruk, baik bagi dirinya maupun hartanya. Seiring meningkatnya jumlah kendaraan bermotor di Indonesia, risiko kecelakaan juga naik, sehingga asuransi kendaraan bermotor, khususnya asuransi tanggung jawab pihak ketiga (*Third Party Liability* (TPL), menjadi penting. Untuk meningkatkan keadilan, perusahaan asuransi menerapkan sistem premi berdasarkan riwayat klaim, salah satunya sistem bonus malus. Dalam penelitian ini dibahas perhitungan premi pada sistem bonus malus untuk asuransi TPL, dengan asumsi frekuensi klaim berdistribusi Poisson-Lindley dan besar klaim berdistribusi eksponensial-invers gamma. Data yang digunakan adalah data sekunder dari PT. XYZ pada tahun underwriting 2019 untuk pemegang polis kategori dua. Hasil analisis menunjukkan bahwa distribusi tersebut sesuai dengan data, dan sistem bonus malus yang optimal memberikan premi murni awal sebesar Rp22.970 untuk pemegang polis baru. Premi tahun berikutnya disesuaikan berdasarkan klaim: meningkat jika ada klaim, dan menurun jika tidak ada klaim.
Kata kunci: Bonus-Malus; Asuransi *Third Party Liability*; Distribusi Poisson-Lindley dan Eksponensial-Invers Gamma

ABSTRACT

Insurance is a form of mutual cooperation that provides protection against unforeseen risks. With insurance, individuals can feel more secure about potential future losses, whether related to themselves or their property. As the number of motor vehicles in Indonesia increases, so does the risk of traffic accidents, making motor vehicle insurance particularly Third Party Liability (TPL) insurance increasingly important. To enhance fairness, insurance companies implement premium systems based on claim history, one of which is the bonus-malus system. This study discusses premium calculation in a bonus-malus system for TPL insurance, assuming that claim frequency follows a Poisson-Lindley distribution and claim severity follows an exponential-inverse gamma distribution. The data used are secondary data obtained from PT. XYZ for the 2019 underwriting year, focusing on policyholders in category two. The analysis results indicate that the selected distributions fit the data well. The optimal bonus-malus system determines that the initial pure premium to be paid by new policyholders is Rp22,970. Premiums in subsequent years are adjusted based on claim activity: increasing if a claim is made and decreasing if no claim occurs.
Keywords: *Bonus-Malus; Third Party Liability Insurance; Poisson-Lindley and Exponential-Inverse Gamma Distributions*

INTRODUCTION

Indonesia, with the number of motor vehicles increasing every year, faces major challenges in managing transportation and highway safety. Data from the Badan Pusat Statistik (BPS) noted that from 2018 to 2023, the number of motorcycles increased from 106 million to 132 million units, up 26 million units in six years. Cars also increased from 14 million to 18 million units [1]. This growth not only causes congestion, but also increases the potential for financial losses due to accidents. In this case, there are still many road users, including third parties who are harmed in accidents, do not have adequate financial protection.

Motor vehicle insurance is an important solution. There are two main types of coverage in motor vehicle insurance in Indonesia, namely Comprehensive and Total Loss Only (TLO), where the insured must choose one of them. In addition, additional protections such as Third Party Liability (TPL) can be added as needed. TPL Insurance provides protection against claims for compensation from third parties who suffer losses due to accidents caused by policyholders. As a measure to improve protection for all road users, the Indonesian government has announced a plan to implement mandatory TPL insurance for all motor vehicles starting January 2025 [2].

One of the important things in motor vehicle insurance is to determine the amount of premium that must be paid by the policyholder [3]. One method to determine the amount of premium is the bonus-malus system. This system provides incentives in the form of premium reductions for drivers who do not file claims (bonus) and imposes higher premiums for those who often file claims (malus). The system is designed to encourage safer and more responsible driving behavior. The simplest bonus-malus system is based on the frequency of claims filed by policyholders. The system is considered unfair because insured who submits a claim with a small value is subject to the same premium as a claim with a large value. The solution to improve the fairness of the bonus-malus system is to consider two components, namely the frequency of claims and the severity of claims.

Various previous studies have examined the bonus-malus system in motor vehicle insurance using a different distribution approach. Eygenio and Qoyyimi [4] designed a generalized bonus-malus system with the frequency of Negative Binomial distributed claims and the magnitude of Pareto distributed claims on motor vehicle insurance in the 2014 motor vehicle insurance policy data of the insurance company Jasindo. Adisti and Mutaqin [5] applied pure premium calculation to the bonus-malus system where the frequency of Negative Binomial distributed claims and the amount of Weibull distributed claims on category 8 motor vehicle insurance data in region 3 in Indonesia. Sevina and Purwadi [6] applied pure premium calculation to the bonus-malus system where the frequency of Geometric distributed claims and the amount of Weibull distributed claims on the secondary data of motor vehicle insurance in Indonesia. In this study, the application of premium calculation in the bonus-malus system will be carried out on motor vehicle insurance in Indonesia for Third Party Liability (TPL) coverage, with the assumption that the claim frequency follows a mixture Poisson-Lindley distribution and the claim severity follows a mixture Exponential-Inverse Gamma distribution. The Poisson-Lindley distribution is selected because it offers greater flexibility than the standard Poisson distribution in modeling overdispersed count data, a common feature in insurance claim frequencies. It also provides a better fit when the variance of the data exceeds the mean, which is often observed in real insurance datasets.

METHOD

The data used in this study is secondary data taken from the recording of PT. XYZ in the 2019 underwriting year, includes the frequency of claims and severity claims from policyholders of motor vehicle insurance products covered comprehensive expansion of Third Party Liability (TPL) category two in Indonesia.

Poisson Distribution

A discrete random variable K is said to be distributed in Poisson with a parameter $\lambda > 0$, and has the following probability mass function:

$$P(k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{ for } k = 0,1,2, \ldots \quad (1)$$

Poisson-Lindley Mixture Distribution

The Poisson-Lindley mixture distribution is formed from the Poisson distribution with the parameter λ following the Lindley distribution with the parameter $\delta > 0$. The probability density function of the Lindley distribution is:

$$\pi(\lambda) = \frac{\delta}{\delta+1}(\lambda+1)e^{-\delta\lambda}, \text{ for } \lambda > 0, \delta > 0 \quad (2)$$

The probability function of the Poisson-Lindley mixture distribution is as follows:

$$p_k = P(K = k) = \int_0^\infty P(k|\lambda)\pi(\lambda)d\lambda$$

$$p_k = \frac{\delta^2(k+\delta+2)}{(\delta+1)^{k+3}}; k = 0,1,2, \ldots, \delta > 0 \quad (3)$$

The parameters of the Poisson-Lindley mixture distribution can be estimated using the Maximum Likelihood Estimation (MLE) method. The log-likelihood function of the Poisson-Lindley mixture distribution for random samples $k = k_1, k_2, \ldots, k_n$ is:

$$l(\delta) = 2n \ln \delta + \sum_{i=1}^n \ln(k_i + \delta + 2) - \ln(\delta+1) \sum_{i=1}^n (k_i + 3) \quad (4)$$

The first derivative of the log-likelihood function against a parameter δ equalized to zero is obtained as follows:

$$\frac{d}{d\delta} = \frac{2n}{\delta} + \sum_{i=1}^n \frac{1}{(k_i+\delta+2)} - \frac{1}{\delta+1}\sum_{i=1}^n (k_i + 3) = 0 \quad (5)$$

There is no analytical solution to obtain parameter estimates δ, therefore a numerical method is used, namely the Newton-Raphson iteration method, which requires a second derivative of the log-likelihood function for the parameter δ in the calculation process.

$$\frac{d^2}{d\delta^2} = -\frac{2n}{\delta^2} - \sum_{i=1}^n \frac{1}{(k_i+\delta+2)^2} + \frac{1}{(\delta+1)^2}\sum_{i=1}^n (k_i + 3) \quad (6)$$

The Newton-Raphson iteration equation is:

$$\hat{\delta}_{h+1} = \hat{\delta}_h - \frac{\frac{2n}{\hat{\delta}} + \sum_{i=1}^n \frac{1}{(k_i+\hat{\delta}+2)} - \frac{1}{\hat{\delta}+1}\sum_{i=1}^n (k_i+3)}{-\frac{2n}{\hat{\delta}^2} - \sum_{i=1}^n \frac{1}{(k_i+\hat{\delta}+2)^2} + \frac{1}{(\hat{\delta}+1)^2}\sum_{i=1}^n (k_i+3)}; \; h = 0,1,2, \ldots \quad (7)$$

The iteration process is stopped if $\left|\hat{\delta}_{h+1} - \hat{\delta}_h\right| < \varepsilon$, with the value $\varepsilon = 1 \times 10^{-6}$. The initial value used in the iteration process is obtained from the estimation of the Lindley distribution parameter using the moment method which is [7]:

$$\hat{\delta}_0 = \frac{-(\bar{k}-1)+\sqrt{(\bar{k}-1)^2+8\bar{k}}}{2\bar{k}}, \text{ for } \bar{k} > 0 \quad (8)$$

where:

$$\bar{k} = \frac{\sum_{i=1}^{n} k_i}{n} \quad (9)$$

Exponential Distribution

A continuous random variable X is said to follow an exponential distribution with the parameter θ if it has the following density function [8]:

$$f(x|\theta) = \frac{e^{-x/\theta}}{\theta}, \text{ for } x > 0, \theta > 0 \quad (10)$$

The cumulative distribution function of the exponential distribution with the parameter θ:

$$F(x|\theta) = 1 - e^{-x/\theta}, \text{ for } x > 0, \theta > 0 \quad (11)$$

Exponential-Inverse Gamma Mixture Distribution

The exponential distribution of the inverse gamma mixture is formed from an exponential distribution with a parameter θ following an inverse gamma distribution with the parameter $\alpha > 0$ and $\beta > 0$. The probability density function of the inverse gamma distribution is [8]:

$$F\pi(\theta) = \frac{\left(\frac{\beta}{\theta}\right)^{\alpha} e^{-\beta/\theta}}{\theta\Gamma(\alpha)}, \text{ for } \theta > 0,\ \alpha > 1,\ \beta > 0 \quad (12)$$

The probability function of the exponential–inverse gamma mixture distribution is as follows:

$$f(x) = \int_0^{\infty} f(x|\theta)\pi(\theta)d\theta$$

$$f(x) = \frac{\alpha\beta^{\alpha}}{(x+\beta)^{\alpha+1}}, \text{ for } x > 0,\ \alpha > 0,\ \beta > 0 \quad (13)$$

The cumulative distribution function of the exponential–inverse gamma mixture distribution is:

$$F(x) = 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha}, \text{ for } x > 0,\ \alpha > 0,\ \beta > 0 \quad (14)$$

The parameters of the exponential–inverse gamma mixture distribution can be estimated using the Maximum Likelihood Estimation (MLE) method. The log-likelihood function of the exponential–inverse gamma mixture distribution for random samples $x = x_1, x_2, \ldots, x_n$ is:

$$l(\alpha,\beta) = n\ln\alpha + n\alpha\ln(\beta) - (\alpha+1)\sum_{i=1}^{n}\ln(x_i+\beta) \quad (15)$$

The first derivative of the log-likelihood function to the parameter α and β which is equaled to zero is obtained as follows:

$$\frac{dl(\alpha,\beta)}{d\alpha} = \frac{n}{\alpha} + n\ln(\beta) - \sum_{i=1}^{n}\ln(x_i+\beta) = 0 \quad (16)$$

$$\frac{dl(\alpha,\beta)}{d\beta} = n\frac{\alpha}{\beta} - (\alpha+1)\sum_{i=1}^{n}\frac{1}{x_i+\beta} = 0 \quad (17)$$

There is no analytical solution to obtain parameter estimates α and β therefore a numerical method is used, namely the Newton-Raphson iteration method, which requires a second derivative of the log-likelihood function to the parameter α and β in the calculation process.

$$\frac{d^2l(\alpha,\beta)}{d\alpha^2} = -\frac{n}{\alpha^2} \quad (18)$$

$$\frac{d^2 l(\alpha,\beta)}{d\beta^2} = -n\frac{\alpha}{\beta^2} + (\alpha+1)\sum_{i=1}^{n}\frac{1}{(x_i+\beta)^2} \quad (19)$$

$$\frac{d^2 l(\alpha,\beta)}{d\alpha d\beta} = \frac{n}{\beta} - \sum_{i=1}^{n}\frac{1}{x_i+\beta} \quad (20)$$

Suppose $\gamma = (\alpha,\beta)T$ it is a parameter vector. At the $(h+1)$ iteration, the updated parameter can be obtained:

$$\gamma^{(h+1)} = \gamma^{(h)} - \left[\frac{d^2 l(\gamma)}{d\gamma d\gamma^T}\Big|_{\gamma=\gamma^{(h)}}\right]^{-1}\left[\frac{dl(\gamma)}{d\gamma}\Big|_{\gamma=\gamma^{(h)}}\right], h = 0,1,2,.. \quad (21)$$

where,

$$\frac{dl(\gamma)}{d\gamma}\Big|_{\gamma=\gamma^{(h)}} = \begin{pmatrix} \frac{dl(\alpha,\beta)}{d\alpha}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} \\ \frac{dl(\alpha,\beta)}{d\beta}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} \end{pmatrix} \quad (22)$$

and

$$\frac{d^2 l(\gamma)}{d\gamma d\gamma^T}\Big|_{\gamma=\gamma^{(h)}} = \begin{pmatrix} \frac{d^2 l(\alpha,\beta)}{d\alpha^2}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} & \frac{d^2 l(\alpha,\beta)}{d\alpha d\beta}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} \\ \frac{d^2 l(\alpha,\beta)}{d\alpha d\beta}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} & \frac{d^2 l(\alpha,\beta)}{d\beta^2}\Big|_{\alpha=\alpha^{(h)};\,\beta=\beta^{(h)}} \end{pmatrix} \quad (23)$$

The iteration process is stopped when $\left|\gamma^{(h+1)} - \gamma^{(h)}\right| < \varepsilon$, with the value $\varepsilon = 1 \times 10^{-6}$. The initial values used in the iteration process can be obtained using the estimation of the parameters of the exponential-inverse gamma mixture distribution using the moment method, as follows:

$$\hat{\alpha}_0 = \frac{-2s^2}{(\bar{x}^2 - s^2)} \quad (24)$$

$$\hat{\beta}_0 = \bar{x}(\hat{\alpha}_0 - 1) \quad (25)$$

Goodness-of-Fit Testing for the Poisson-Lindley Mixture Distribution

The following are the steps in conducting a Poisson-Lindley mixture distribution compatibility test analysis of claim frequency data using the Chi-Squared test:

1. Formulating a test hypothesis.
   H_0 : Claim frequency data comes from a Poisson-Lindley mixture distributed population.
   H_1 : Claim frequency data does not come from a Poisson-Lindley mixture distributed population.
2. Estimating the parameters of the Poisson-Lindley mixture distribution using the Newton-Raphson iteration numerical method.
3. Calculates the probability for each claim frequency category, p_k, for k = 0.1.2,... based on Equation (3).
4. Calculate the expected value for each claim frequency category.
5. Calculate the statistical value of the Chi-Square test using the equation:

$$\chi^2 = \sum_{i=1}^{m}\frac{(O_i - E_i)^2}{E_i} \quad (26)$$

   Information:

E_i : The number of observations expected in the ith category based on the distribution function

O_i : Number of observations observed in the ith category

6. Decide whether the null hypothesis is accepted or rejected. The null hypothesis is rejected if $\chi^2 \geq \chi^2_{(m-p-1)(1-\alpha)}$. According to Howell [9], the chi-square test can be invalid if the value of $E_i < 5$, but it can still be used if the sample total is large enough and the null hypothesis is not rejected.

Goodness-of-Fit Testing for the Exponential-Inverse Gamma Mixture Distribution

The following are the steps in conducting a match test of the exponential–inverse gamma mixture distribution match to claim severity data using the Kolmogorov-Smirnov test:

1. Formulating a test hypothesis.

   H_0 : Claim severity data to comes from a population of exponential mix-inverse gamma.

   H_1 : Claim severity data does not come from a population with an exponential-inverse gamma mixture.
2. Estimating the parameters of the exponential-inverse gamma mixture distribution parameters using the Newton-Raphson iteration numerical method.
3. Sorting claim severity data on motor vehicle claims from smallest to largest value.
4. Perform calculations based on the empirical cumulative distribution function on the data for each observation using the equation:

$$F_n(x_i) = \frac{\text{the number of observations } \leq x_i}{n} \tag{27}$$

5. Calculate the value of the cumulative distribution function of the exponential-inverse gamma mixture using Equation (14).
6. Calculating the statistical value of the Kolmogorov-Smirnov test using the equation:

$$D = \max_{1 \leq i \leq n} |F_n(x_i) - F^*(x_i)| \tag{28}$$

7. Decide whether the null hypothesis is accepted or rejected. The null hypothesis is accepted if the test statistic D is smaller than the critical value.

Premium Calculation with the Optimal Bonus-Malus System

The following are the steps in calculating pure premium using the optimal bonus-malus system:

1. Estimate the average number of claims for the year (t + 1) using the equation:

$$\hat{\lambda}_{t+1,N} = \frac{(N+1)(N+2+t+\hat{\delta})}{(t+\hat{\delta})(N+1+t+\hat{\delta})} \tag{29}$$

2. Estimate the average claim severity for the year (t + 1) using the equation:

$$\hat{\theta}_{t+1,N} = \frac{\sum_{i=1}^{N} x_i + \hat{\beta}}{N+\hat{\alpha}-1} \tag{30}$$

3. Calculate the estimated motor vehicle insurance premium for the year ($t = 0$ and $N = 0$) or for new policyholders, using the following equation:

$$\widehat{\text{premium}}_{1,0} = \frac{(2+\hat{\delta})}{(\hat{\delta})(1+\hat{\delta})} \cdot \frac{\hat{\beta}}{\hat{\alpha}-1} \tag{31}$$

4. Calculate the estimated motor vehicle insurance premium for the year (t + 1) for the claim frequency of 0 to 4 times and the period of the year of coverage from the 1st to the 5th year using the equation:

$$\widehat{\text{premium}}_{t+1,N} = \frac{(N+1)(N+2+t+\hat{\delta})}{(t+\hat{\delta})(N+1+t+\hat{\delta})} \cdot \frac{\sum_{i=1}^{N} x_i + \hat{\beta}}{N+\hat{\alpha}-1} \quad (32)$$

RESULT AND DISCUSSION

Goodness-of-Fit Test for the Poisson-Lindley Mixture Distribution

The test hypothesis is formulated as follows:

H_0: The claim frequency data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year comes from a population that follows a Poisson-Lindley mixture distribution.

H_1: The claim frequency data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year does not come from a population that follows a Poisson-Lindley mixture distribution.

Estimation of the parameters of the Poisson-Lindley mixture distribution was sought using the Newton-Raphson iteration method with the help of RStudio software version 4.2.3. The initial value of the estimated Poisson-Lindley distribution parameter is calculated based on Equation (8), with the average value of the data as follows:

$$\bar{k} = \frac{\sum_{i=1}^{n} k_i}{n} = \frac{226}{56488} = 0.004$$

Thus, the initial value for the estimation of the parameters of the Poisson-Lindley mixture distribution is obtained:

$$\hat{\delta}_0 = \frac{-(\bar{k}-1)+\sqrt{(\bar{k}-1)^2+8\bar{k}}}{2\bar{k}} = \frac{-(0{,}004-1)+\sqrt{(0.004-1)^2+8(0.004)}}{2(0.004)} = 250.9925$$

Based on the results of calculations using the Newton-Raphson iteration method, in the 3rd iteration, an estimate of the Poisson-Lindley mixture distribution parameters for category 2 motor vehicle insurance claim frequency data of . This value can be used to calculate the estimated value of the probability for each category of motor vehicle insurance claim frequency using Equation (3). $\hat{\delta} = 250.939$. Table 1. presents the values required to calculate the statistical value of the test. The final result of calculating the chi-squared statistical value based on Equation (26) is found in column (5) of the last line, which is 0.0129.

Table 1. The values required in the statistical calculation of the test.

Claim frequency (k)	Number of Policies (O_i)	Claim Frequency Probability (p_k)	Expected Claim Frequency (E_i)	$\frac{(O_i - E_i)^2}{E_i}$
(1)	(2)	(3)	(4)	(5)
0	56263	0.9960	56262.901	1.76×10^{7}
1	224	0.0039	224.2024	0.0002
≥2	1	1.5816×10^{-5}	0.8934	0.0127
Sum	56488	1	56488	0.0129

With a significance level of 5%, the chi-squared quantile value with a degrees of freedom $1 = (3 - 1 - 1)$ is 3.8414 [9]. Because the statistical value of the test is smaller than the quantile value, the null hypothesis is accepted and it is concluded that the claim frequency data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year, comes from a population that follows a Poisson-Lindley mixture distribution.

Goodness-of-Fit Test for the Exponential-Inverse Gamma Mixture Distribution

The hypotheses used in this test are formulated as follows:

H_0: The claim severity data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year comes from a population that follows a exponential-inverse gamma mixture distribution.

H_1: The claim severity data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year does not come from a population that follows a exponential-inverse gamma mixture distribution. Estimation of the parameters of the exponential-inverse gamma mixture distribution is sought using the Newton-Raphson iteration method with the help of software Rstudio version 4.2.3. The initial value of the estimated parameter of the exponential-inverse gamma distribution parameter calculated based on Equation (24) for $\hat{\alpha}_0$ and Equation (25) for $\hat{\beta}_0$, with the mean values and variances of the claim severity data which are respectively $\bar{x} = 5760171.668$ and $s^2 = 5.1307 \times 10^{13}$. Thus, the initial value is obtained for the estimation of the parameters of the exponential-inverse gamma mixture distribution:

$$\hat{\alpha}_0 = \frac{-2s^2}{(\bar{x}^2 - s^2)} = \frac{-2(5.1307 \times 10^{13})}{((5.760.171{,}668)^2 - 5.1307 \times 10^{13})} = 5.66$$

Substitute the value in Equation (25), to obtain:$\hat{\alpha}_0$

$$\hat{\beta}_0 = \bar{x}(\hat{\alpha}_0 - 1) = 5760171.668(5.66 - 1) = 26843584.47$$

Based on the results of the calculation using the Newton-Raphson iteration method, in the 7th iteration, an estimate of the parameters of the mixture distribution was obtained exponential-inverse gamma for the claim severity data category 2 motor vehicle insurance claims are equal to $\hat{\alpha} = 6.4909$ and $\hat{\beta} = 31524867$. Table 2. presents the values needed to calculate the statistics of the Kolmogorov-Smirnov test. Based on the results of the calculation in Table 2., then the statistics of the Kolmogorov-Smirnov test can be calculated using Equation (28), namely:

$$D = \max_{1 \le i \le n} |F_n(x_i) - F^*(x_i)| = 0.0625.$$

With a significance level of 5%, the critical value is 0.0905 [10]. Since the statistical value of the Kolmogorov-Smirnov test is smaller than its critical value, the null hypothesis is accepted and it is concluded that the claim severity data of motor vehicle insurance with comprehensive coverage extended by TPL category 2 for the 2019 underwriting year comes from a population that follows a exponential-inverse gamma mixture distribution.

Table 2. Calculation Results for the Kolmogorov-Smirnov Test Claim Severity Data

i	x_i	$F_n(x_i)$	$F^*(x_i)$	$\lvert F_n(x_i) - F^*(x_i)\rvert$
(1)	(2)	(3)	(4)	(5)
1	300000	0.0044	0.0596	0.0552

2	325000	0.0088	0.0644	0.0556
3	385000	0.0133	0.0758	0.0625
4	386000	0.0221	0.0760	0.0583
5	386000	0.0221	0.0760	0.0538
:	:	:	:	:
226	50000000	1	0.9979	0.0021

Optimal Bonus-Malus System Premium Calculation

The determination of the optimal bonus-malus system premium is obtained using two components, namely the frequency of claims and the amount of claims. It was found that the claim frequency follows a Poisson–Lindley mixture distribution with an estimated parameter value of $\hat{\delta} = 250.939$, and the claim severity follows an exponential–inverse gamma mixture distribution with estimated parameter values of $\hat{\alpha} = 6.4909$ and $\hat{\beta} = 31524867$. For new policyholders ($t = 0, N = 0$):

$$\widehat{\text{prem}}_{1,0} = \frac{(2+\hat{\delta})}{(\hat{\delta})(1+\hat{\delta})} \cdot \frac{\hat{\beta}}{\hat{\alpha}-1} = \frac{(2+250.939)}{(250.939)(1+250,939)} \cdot \frac{31524867}{6,4909-1} = 22970$$

This value indicate that the contribution to the pure premium for the TPL extension that must be paid by new policyholders in the first year is Rp22,970. For policyholders who did not file a claim in the first year ($t = 1, N = 0$):

$$\widehat{\text{prem}}_{2,0} = \frac{(N+1)(N+2+t+\hat{\delta})}{(t+\hat{\delta})(N+1+t+\hat{\delta})} \cdot \frac{\sum_{i=1}^{N} x_i + \hat{\beta}}{N+\hat{\alpha}-1} = \frac{(0+1)(0+2+1+250.939)}{(1+250.939)(0+1+1+250.939)} \cdot \frac{0+31524867}{0+6.4909-1} = 22878$$

This value indicate that the contribution to the pure premium for the TPL extension that must be paid by policyholders who do not file a claim in the first year is Rp22,878. For policyholders who make a claim in the first year once ($t = 1, N = 1, x_1 = 6000000$):

$$\widehat{\text{prem}}_{2,1} = \frac{(1+1)(1+2+1+250.939)}{(1+250.939)(1+1+1+250.939)} \cdot \frac{6000000+31524867}{1+6.4909-1} = 46074$$

This value indicate that the contribution to the pure premium for the TPL extension that must be paid by policyholders who submit a claim in the first year once with a claim amount of Rp6,000,000 is Rp46,074. For policyholders who make a claim in the first year twice ($t = 1, N = 2, x_1 = 6000000, x_2 = 300000$):

$$\widehat{\text{prem}}_{2,2} = \frac{(2+1)(2+2+1+250.939)}{(1+250.939)(2+1+1+250.939)} \cdot \frac{(6000000+300000)+31524867)}{2+6.4909-1} = 60362$$

Table 3. presents the amount of premiums that must be paid by policyholders, starting from the new policyholder to the amount of premiums for the following years. The severity of the claim is for example: the first claim is Rp6,000,000, the second claim is Rp300,000, the third claim is Rp10,000,000, and the fourth claim is Rp50,000,000.

Table 3. Premium Value Based on Claim Frequency and Claim Amount (Rupiah)

Year (t)	Number of Claims (N)				
	0	1	2	3	4
0	22,970				
1	22,878	46,074	60,362	89,775	205,353
2	22,788	45,891	60,123	89,419	204,538
3	22,697	45,709	59,885	89,065	203,730

4	22,608	45,529	59,649	88,715	202,928
5	22,519	45,351	59,416	88,367	202,132

CONCLUSION

This study applies the calculation of pure premiums within the bonus-malus system using a Poisson–Lindley mixture distribution for claim frequency and an exponential–inverse gamma mixture distribution for claim severity, based on motor vehicle insurance data with comprehensive coverage extended by TPL category 2 in Indonesia. The results indicate that the Poisson–Lindley distribution fits the claim frequency data well, and the exponential–inverse gamma distribution fits the claim severity data appropriately. The estimated pure premium for new policyholders in the first year is Rp22,970. The premium amounts in subsequent years are determined by the frequency and severity of the claims submitted. Premiums increase if claims occur and decrease if no claims are made. It is recommended that insurance companies offering TPL coverage consider implementing the bonus-malus system in premium calculations. Future research may explore alternative distributions for claim frequency and severity to evaluate more optimal approaches within the bonus-malus system.

REFERENCES

[1] GoodStats, "Perkembangan jumlah kendaraan bermotor Indonesia, sepeda motor terbanyak," *GoodStats.id*, 2024. [Online]. Available: https://data.goodstats.id/statistic/perkembangan-jumlah-kendaraan-bermotor-indonesia-sepeda-motor-terbanyak-KC4lR. [Accessed: Feb. 6, 2025].

[2] Kompas.id, "Menyoal rencana wajib asuransi kendaraan bermotor di Indonesia," *Kompas*, Jul. 26, 2024. [Online]. Available: https://www.kompas.id/baca/riset/2024/07/26/menyoal-rencana-wajib-asuransi-kendaraan-bermotor-di-indonesia. [Accessed: Feb. 6, 2025].

[3] Y.-L. Grize, "Applications of statistics in the field of general insurance: An overview," *Int. Stat. Rev.*, vol. 83, pp. 135–159, 2015, doi: 10.1111/insr.12066.

[4] F. Eygenio and D. T. Qoyyimi, "Sistem bonus malus tergeneralisasi dengan frekuensi klaim berdistribusi binomial negatif dan besar klaim berdistribusi Pareto," Undergraduate thesis, Dept. Stat., Univ. Gadjah Mada, Yogyakarta, 2019.

[5] R. I. Adisti and A. K. Mutaqin, "Perhitungan premi murni pada sistem bonus malus untuk frekuensi klaim berdistribusi binomial negatif dan besar klaim berdistribusi Weibull pada data asuransi kendaraan bermotor di Indonesia," *J. Gaussian*, vol. 10, no. 2, pp. 170–179, 2021, doi: 10.14710/j.gauss.v10i2.30084.

[6] G. R. Sevina and J. Purwadi, "Bonus malus system for motorized vehicle insurance using geometric distributions and Weibull distributions," *J. Fundam. Math. Appl.*, vol. 6, no. 1, pp. 71–79, 2023, doi: 10.14710/jfma.v6i1.16505.

[7] M. E. Ghitany, B. Atieh, and S. Nadarajah, "Lindley distribution and its application," *Math. Comput. Simul.*, vol. 78, no. 4, pp. 493–506, 2008, doi: 10.1016/j.matcom.2007.06.007.

[8] S. Klugman, H. H. Panjer, and G. E. Willmot, *Loss Models: From Data to Decisions*, 4th ed. Hoboken, NJ: Wiley, 2019.

[9] Sudjana, *Metode Statistika*, 6th ed. Bandung: Tarsito, 2005.

[10] S. Nugroho, *Statistika Nonparametrika*. Yogyakarta: Graha Ilmu, 2008.

Evaluating Patient Satisfaction in Surabaya Public Health Centers Using an Integrated IPA–Kano Framework

Salman Alfarizi Pradana Andikaputra(1), Sumartono(2), Nuril Huda(3)

[1,2,3] Universitas Dr. Soetomo Surabaya

Jalan Semolowaru 84, surabaya 60118, East Java, Indonesia
E-mail: Salman.alfarizi@unitomo.ac.id(1), sumartono@unitomo.ac.id(2),
nurilhudamohammad9@gmail.com(3)

ABSTRAK

Pelayanan kesehatan yang berkualitas di Puskesmas sangat penting untuk meningkatkan kepuasan dan kepercayaan masyarakat, terutama di kota besar seperti Surabaya. Penelitian ini bertujuan mengidentifikasi dan memprioritaskan atribut layanan menggunakan Importance Performance Analysis (IPA), mengklasifikasikannya dengan metode Kano, serta mengintegrasikan keduanya untuk merumuskan strategi peningkatan layanan yang komprehensif. Metode survei dengan kuesioner digunakan pada 85 pasien Puskesmas Surabaya, kemudian dianalisis melalui IPA dan Kano. Hasil IPA menunjukkan terrdapat 4 variabel berada di kuadran prioritas rendah. Keempat variabel tersebut adalah kecepatan proses pendaftaran, sikap petugas dalam memberikan pelayanan, kondisi ruang tunggu, dan ketersediaan area parkir yang memadai. sementara metode Kano mengklasifikasikan seluruh atribut sebagai One Dimensional (O), artinya peningkatan kualitas atribut akan sebanding dengan peningkatan kepuasan pasien. Hasil perhitungan dengan metode Kano didapatkan bahwa semua variabel yang tertera merupakan aspek One Dimensional (O), menunjukkan Peningkatan kepuasan konsumen akan sebanding jika atribut ini ditingkatkan. Sebaliknya, kepuasan akan berkurang jika kinerja atribut ini menurun. Integrasi Metode IPA dan Metode Kano ada beberapa variabel yang harus di tingkatkan yakni kecepatan proses pendaftaran, sikap petugas dalam memberikan pelayanan, kondisi ruang tunggu, dan ketersediaan area parkir yang memadai. Temuan ini menggarisbawahi pentingnya fokus pada aspek-aspek tersebut untuk meningkatkan kualitas pelayanan dan kepuasan pasien secara keseluruhan.
Kata kunci: Pelayanan Puskesmas; *Importance Performance Analysis* (IPA); Metode Kano; Integrasi Metode Kano dan IPA

ABSTRACT

Quality healthcare services at community health centers (Puskesmas) are crucial for enhancing public satisfaction and trust, particularly in large cities like Surabaya. This study aims to identify and prioritize service attributes using Importance Performance Analysis (IPA), classify them using the Kano method, and integrate both approaches to formulate a comprehensive service improvement strategy. A survey method with questionnaires was administered to 85 patients at Puskesmas Surabaya, and the data were analyzed using IPA and Kano methods. The IPA results show that four variables fall into the low-priority quadrant: registration speed, the attitude of the staff in providing service, waiting room conditions, and the availability of adequate parking spaces. Meanwhile, the Kano method classified all attributes as One Dimensional (O), meaning that improving the quality of these attributes will proportionally enhance patient satisfaction. Conversely, satisfaction will decrease if the performance of these attributes declines. The integration of IPA and Kano methods highlights the need to improve the speed of registration, staff service attitudes, waiting room conditions, and parking availability. These findings emphasize the importance of focusing on these aspects to improve overall service quality and patient satisfaction.
Keywords: *Public Health Center services; Importance Performance Analysis (IPA); Kano Method; IPA Method and Kano Method integration*

INTRODUCTION

In developing countries like Indonesia, healthcare services are a crucial indicator for improving the quality of life for the population [1]. As a metropolitan city with a continuously growing population, Surabaya faces challenges in maintaining the quality of healthcare services, especially through public facilities like community health centers (Puskesmas) [2]. Puskesmas located in urban centers play a strategic role in supporting public health [3]. Therefore, measuring patient satisfaction with the facilities and services provided is essential [4].

Health issues are one of the factors affecting quality of life, reflected in the fulfillment of basic human needs [5]. Improvements in healthcare are critical, as this sector is closely linked to development, particularly in terms of human resource development [6]. Population growth is a major cause of environmental problems. Overpopulation can lead to slums, poverty, environmental pollution, and the depletion of natural resources [7].

Importance Performance Analysis (IPA) is a multi-attribute method used to measure the relationship between consumer perceptions (in this study, Puskesmas patients) and the priorities for improving the quality of products or services, also known as quadrant analysis [8]. Importance Performance Analysis (IPA) combines the measurement of importance and satisfaction factors in a two-dimensional graph, making data easier to understand and providing practical recommendations [9]. The IPA graph is divided into four quadrants: Quadrant I "maintain good performance," Quadrant II "priority for improvement," Quadrant III "low priority," and Quadrant IV "possibly excessive" [10].

The Kano Model, developed by Noriaki Kano, categorizes product or service attributes based on their ability to meet customer needs and provide satisfaction [11]. A study at Arifin Hospital in Pekanbaru found that attributes that increased patient satisfaction included the doctor's accuracy in diagnosis, the availability of medication, the speed of examination services, and the tidiness of patient rooms, with satisfaction increases of 74%-87% when met. Conversely, attributes that decreased satisfaction included the speed of examination, the accuracy of diagnosis, staff performance in providing information, and services that did not differentiate patient status, with satisfaction decreases of 80%-85% when unmet.

Both IPA and Kano methods are commonly used to evaluate customer satisfaction, but each has its limitations when used separately [12]. IPA only measures two dimensions: importance and performance, without considering emotional impacts on customers, while Kano focuses more on customer preferences and expectations, without identifying priorities for improvement based on performance. Therefore, integrating both methods is necessary to provide a more comprehensive picture of service quality [13]. This study aims to identify and prioritize service attributes using IPA, classify them with the Kano model, and integrate both approaches to formulate a comprehensive service improvement strategy. The hypothesis is that integrating IPA and Kano will provide deeper insights into improving healthcare service quality.

METHOD

This study adopts a quantitative approach using a survey method. Primary data were collected through the distribution of questionnaires and interviews with patients at Puskesmas Surabaya. The sampling technique used was convenience sampling, where the questionnaire was randomly given to

patients who were present and willing to participate as respondents, with a total sample of 85 respondents.

The research instrument consists of demographic variables and patient satisfaction variables. Demographic variables include respondent origin, gender, age, marital status, occupation, and income level. Meanwhile, patient satisfaction variables were measured using an interval scale ranging from 0 (strongly disagree) to 4 (strongly agree), with columns for expectations and actual experiences based on the five SERVQUAL dimensions, which are:

1. Reliability – including friendliness, clarity of information, and service speed.
2. Responsiveness – covering the responsiveness and promptness of staff in providing service.
3. Assurance – relating to comfort, security, and staff competence.
4. Empathy – including the attention and care of staff towards patients.
5. Tangibles – covering physical facilities such as cleanliness, staff appearance, and parking area.

Data were analyzed using the Importance Performance Analysis (IPA) method to identify priority improvements, and the Kano Model to classify service attributes based on their impact on patient satisfaction [14].

To analyze the facilities at Puskesmas Surabaya that need improvement and maintenance, Importance Performance Analysis (IPA) is used with the following steps:

1. Calculate the average score for the "Actual" (X) and Expectation (Y) columns to obtain ($\bar{X}$) and ($\bar{Y}$)
2. Calculate the averages ($\bar{X}$) and ($\bar{Y}$) to obtain the boundary values ($\bar{\bar{X}}$) and ($\bar{\bar{Y}}$).
3. Create a plot on a Cartesian diagram as shown in Figure 2.1 using SPSS 17.0 software.
4. Interpret the Cartesian diagram based on the results by analyzing which variables fall into Quadrants I, II, III, and IV, and draw conclusions.

To analyze customer satisfaction with the facilities at Puskesmas, the Kano Method is used with the following steps:

1. Categorize each attribute (A, M, O, I, R, Q) from each functional and dysfunctional question using formulas in Microsoft Excel, with a value range as indicated in the table below.

Tabel 1. Score from Questionnaire Score to Value Canoe Method

Questionnaire Value Range	Value of the Kano Method
0 – 0,79	5
0,8 – 1,59	4
1,6 – 2,39	3
2,4 – 3,19	2
3,2 - 4	1

2. Determine which category (A, M, O, I, R, Q) appears most frequently, so that the category reflects each attribute.
3. Integrated each attributes of Kano and Importance Performance Analysis (IPA) method, so the result can be considered more focused for improvement in the future.

Seri (Kategori Kano)	Kepentingan	Kinerja	Kuadran IPA	Prioritas Strategi	
				Peningkatan Kinerja	Pertahankan Kinerja
M (Must Be)	Tinggi	Tinggi	I	-	1
	Tinggi	Rendah	II	1	-
	Rendah	Rendah	III	2	-
	Rendah	Tinggi	IV	-	2
O (One Dimensional)	Tinggi	Tinggi	I	-	3
	Tinggi	Rendah	II	3	-
	Rendah	Rendah	III	4	-
	Rendah	Tinggi	IV	-	4
A (Attractive)	Tinggi	Tinggi	I	-	5
	Tinggi	Rendah	II	5	-
	Rendah	Rendah	III	6	-
	Rendah	Tinggi	IV	-	6

(sumber: Kuo *et al.*, 2011)

Figure 1. Integration of IPA and Kano Model

RESULT AND DISCUSSION

Importance Performance Analysis (IPA)

Importance Performance Analysis (IPA) is used to compare consumer perceptions regarding the importance of service quality (Importance) with the level of service quality provided (Performance), using the five service quality dimensions that have been developed. IPA analysis on the dimensions of reliability, responsiveness, assurance, empathy, and tangibles was conducted to assess patient satisfaction at Puskesmas Surabaya concerning the variables related to the reliability dimension as follows.

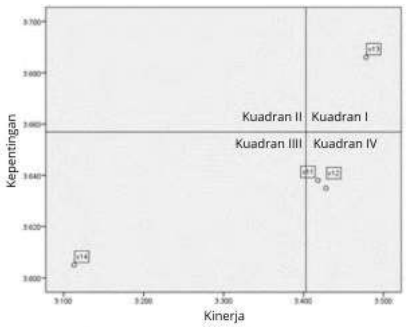


Figure 2. IPA Results for the Reliability Dimension

In Figure 2, the variable that falls into Quadrant I, which indicates high-quality service and needs to be maintained, is the staff's ability to clearly explain the usage of medications ($X_{1.3}$) In Quadrant III, which indicates low priority, aspects of service that receive less attention from patients and have lower service quality include the speed of the registration process ($X_{1.4}$). The staff's friendliness and politeness ($X_{1.1}$), and patient satisfaction with the information provided by pharmacy staff ($X_{1.2}$) fall into Quadrant IV, meaning that although these service elements are considered less important, the service provided is still satisfactory.

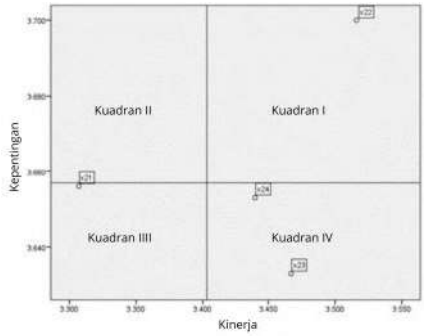


Figure 3. IPA Results for the Responsiveness Dimension

In Figure 3, the variable that falls into Quadrant I, indicating high-quality service that should be maintained, is the staff's ability to clearly explain how to use medications ($X_{2.2}$). In Quadrant III, which represents low priority, the service aspects that receive less attention from patients and have lower quality include the speed of the registration process ($X_{2.1}$). In Quadrant IV, which reflects less important service elements, the staff's friendly response to patient suggestions ($X_{2.3}$), and the staff's priority on patient comfort ($X_{2.4}$) are perceived as satisfactory despite being considered less important.

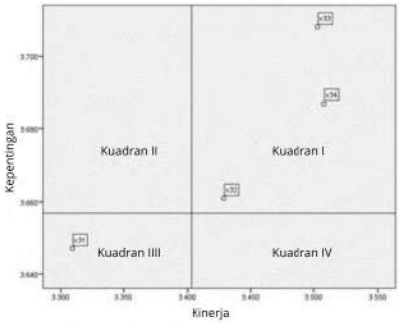


Figure 4. IPA Results for the Assurance Dimension

Based on Figure 4, the variables that fall into Quadrant I and represent excellent service, which should be maintained, include security while at the Puskesmas ($X_{3.2}$), cleanliness of the patient examination rooms ($X_{3.3}$), and staff's broad knowledge about Puskesmas operations and patient diseases ($X_{3.4}$). In Quadrant III, the low-priority variable is the comfort of the waiting area ($X_{3.1}$)

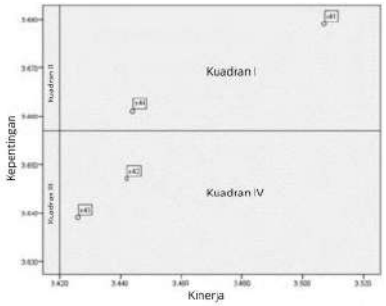


Figure 5. IPA Results for the Empathy Dimension

Based on Figure 5, the variables that fall into Quadrant I and indicate superior service that should be maintained include the staff's patience ($X_{4.1}$) and the staff's high empathy towards patients ($X_{4.4}$). In Quadrant IV, considered less important but still satisfactory, the availability of a good practice schedule ($X_{4.2}$), and a platform for patients to express complaints ($X_{4.3}$) fall into this category.

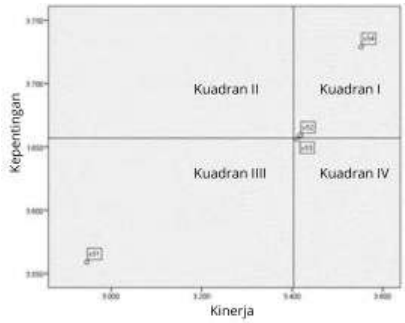


Figure 6. IPA Results for the Tangibles Dimension

In Figure 6, the variables that fall into Quadrant I, indicating excellent service that should be maintained, include cleanliness at the Puskesmas ($X_{5.2}$) and staff appearance being clean and neat ($X_{5.4}$) In Quadrant III, the low-priority variable is the availability of sufficient parking space ($X_{5.1}$). Cleanliness and comfort of the restrooms ($X_{5.3}$) are categorized in Quadrant IV, meaning they are considered less important, but the service provided remains satisfactory.

Analisis Metode Kano

In 1980, Noriaki Kano developed a diagram to categorize product or service attributes based on their ability to provide satisfaction to customers or service users. The results of the Kano Method calculations are as follows:

a. Results of the Kano Method Test for the Reliability Dimension are as follows.

Table 2. Kano Method Calculation for the Reliability Dimension

Variable	Explanation	Kano Method Calculation	Kano Method Categories
$X_{1.1}$	Hospitality and courtesy of officers	*Must-be (*M) =17 *One Dimensional (*O)=67 *Indifferent (*I)=1	*One Dimensional(*O)
$X_{1.2}$	Submission of overall pharmaceutical information by officers	*Must-be (*M) =13 *One Dimensional (*O)=70 *Indifferent (*I)=2	*One Dimensional(*O)
$X_{1.3}$	Officers conveying how to use drugs to be consumed	*Must-be (*M) =10 *One Dimensional (*O)=71 *Indifferent (*I)=2 *Atrractive* (A)=2	*One Dimensional(*O)

$X_{1.4}$	Speed of registration process	*Must-be (*M) =9 *One Dimensional (*O)=53 *Indifferent (*I)=19 *Atrractive* (A)=3 *Reverse* (R)=1	*One Dimensional(*O)

From the results of the Kano method calculation on 85 respondents, it was found that all variables in the Reliability Dimension are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

b. Results of the Kano Method Test for the Responsiveness Dimension are as follows.

Table 3. Kano Method Calculation for the Responsiveness Dimension

Variable	Explanation	Kano Method Calculation	Kano Method Categories
$X_{2.1}$	Attitudes of officers in providing services	*Must-be (*M) =18 *One Dimensional (*O)=62 *Indifferent (*I)=5	*One Dimensional(*O)
$X_{2.2}$	Attitudes of officers in responding to patients' complaints	*Must-be (*M) =9 *One Dimensional (*O)=73 *Indifferent (*I)=1 *Atrractive* (A)=2	*One Dimensional(*O)
$X_{2.3}$	Attitudes of officers in responding to patients' suggestions	*Must-be (*M) =15 *One Dimensional (*O)=66 *Indifferent (*I)=2 *Atrractive* (A)=2	*One Dimensional(*O)
$X_{2.4}$	Officers prioritize patients' comfort	*Must-be (*M) =14 *One Dimensional (*O)=68 *Indifferent (*I)=3	*One Dimensional(*O)

From the results of the Kano method calculation on 85 respondents, it was found that all variables in the Responsiveness Dimension are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

c. Results of the Kano Method Test for the Assurance Dimension are as follows.

Table 4. Kano Method Calculation for the Assurance Dimension

Variable	Explanation	Kano Method Calculation	Kano Method Categories
$X_{3.1}$	Waiting room conditions	*Must-be (*M) =16 *One Dimensional (*O)=61 *Indifferent (*I)=7 *Atrractive* (A)=1	*One Dimensional(*O)

$X_{3.2}$	Safety at the puskesmas	*Must-be (*M) =14 *One Dimensional (*O)=68 *Indifferent (*I)=3	*One Dimensional(*O)
$X_{3.3}$	Checking room conditions	*Must-be (*M) =12 *One Dimensional (*O)=72 *Indifferent (*I)=1	*One Dimensional(*O)
$X_{3.4}$	Officer knowledge of puskesmas	*Must-be (*M) =10 *One Dimensional (*O)=72 *Indifferent (*I)=2 *Reverse* (R)=1	*One Dimensional(*O)

From the results of the Kano method calculation on 85 respondents, it was found that all variables in the Assurance Dimension are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

d. Results of the Kano Method Test for the Empathy Dimension are as follows.

Table 5. Kano Method Calculation for the Empathy Dimension

Variable	Explanation	Kano Method Calculation	Kano Method Categories
$X_{4.1}$	Officer patience	*Must-be (*M) =7 *One Dimensional (*O)=75 *Indifferent (*I)=3	*One Dimensional(*O)
$X_{4.2}$	Availability of treatment schedules	*Must-be (*M) =19 *One Dimensional (*O)=64 *Indifferent (*I)=1 *Atrractive* (A)=1	*One Dimensional(*O)
$X_{4.3}$	The availability of a medium for conveying the patient's complaints	*Must-be (*M) =11 *One Dimensional (*O)=69 *Indifferent (*I)=5	*One Dimensional(*O)
$X_{4.4}$	The officer gives sympathy to the patient	*Must-be (*M) =13 *One Dimensional (*O)=68 *Indifferent (*I)=1 *Reverse* (R)=1	*One Dimensional(*O)

From the results of the Kano method calculation on 85 respondents, it was found that all variables in the Empathy Dimension are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

e. Results of the Kano Method Test for the Tangibles Dimension are as follows.

Table 6. Kano Method Calculation for the Tangibles Dimension

Variable	Explanation	Kano Method Calculation	Kano Method Categories
$X_{5.1}$	Availability of adequate parking area	*Must-be (*M) =29 *One Dimensional (*O)=41 *Indifferent (*I)=13 *Atrractive* (A)=1 *Reverse* (R)=1	*One Dimensional(*O)
$X_{5.2}$	Cleanliness in the Puskesmas area is good	*Must-be (*M) =15 *One Dimensional (*O)=69 *Indifferent (*I)=1	*One Dimensional(*O)
$X_{5.3}$	The cleanliness and comfort of the toilet is good	*Must-be (*M) =17 *One Dimensional (*O)=64 *Indifferent (*I)=2 *Atrractive* (A)=2	*One Dimensional(*O)
$X_{5.4}$	The officer is clean and neat	*Must-be (*M) =6 *One Dimensional (*O)=78 *Indifferent (*I)=1	*One Dimensional(*O)

From the results of the Kano method calculation on 85 respondents, it was found that all variables in the Tangibles Dimension are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

Integration of the Importance Performance Analysis (IPA) Method with the Kano Method

The IPA-Kano integration is applied to address the shortcomings of each method. This model formulates priority development strategies for each service indicator. The results of the IPA-Kano integration are presented in Table 7.

Table 7. IPA-Kano Method Integration

Variable	Information	Result Of Kano Method	Kuadran IPA	Priority Strategy
$X_{1.1}$	Hospitality and courtesy of officers	O	4	Keep It Up
$X_{1.2}$	Submission of overall pharmaceutical information by officers	O	4	Keep It Up
$X_{1.3}$	Officers conveying how to use drugs to be consumed	O	1	Keep It Up
$X_{1.4}$	Speed of registration process	O	3	Improve

$X_{2.1}$	Attitudes of officers in providing services	O	3	Improve
$X_{2.2}$	Attitudes of officers in responding to patients' complaints	O	1	Keep It Up
$X_{2.3}$	Attitudes of officers in responding to patients' suggestions	O	4	Keep It Up
$X_{2.4}$	Officers prioritize patients' comfort	O	4	Keep It Up
$X_{3.1}$	Waiting room conditions	O	3	Improve
$X_{3.2}$	Safety at the puskesmas	O	1	Keep It Up
$X_{3.3}$	Checking room conditions	O	1	Keep It Up
$X_{3.4}$	Officer knowledge of puskesmas	O	1	Keep It Up
$X_{4.1}$	Officer patience	O	1	Keep It Up
$X_{4.2}$	Availability of treatment schedules	O	4	Keep It Up
$X_{4.3}$	The availability of a medium for conveying the patient's complaints,	O	4	Keep It Up
$X_{4.4}$	The officer gives sympathy to the patient	O	1	Keep It Up
$X_{5.1}$	Availability of adequate parking area	O	3	Improve
$X_{5.2}$	Cleanliness in the Puskesmas area is good	O	1	Keep It Up
$X_{5.3}$	The cleanliness and comfort of the toilet is good	O	4	Keep It Up
$X_{5.4}$	The officer is clean and neat	O	1	Keep It Up

Based on the results of the integration of the results of IPA and the Kano Method in Table 7, the following are some of the attributes based on strategic priorities that need to be improved in this study:

a. The speed of registration process
b. The officers' attitude while providing services
c. Waiting room's condition
d. The availability of adequate parking area

The speed of the registration process, quick, accurate, and friendly service from staff, as well as a comfortable waiting area, play a significant role in enhancing customer satisfaction, especially in the healthcare sector. Studies have shown that efficient service and friendly interactions improve

the positive perception of service quality. Additionally, a comfortable waiting area can reduce stress and increase patient comfort. However, the lack of adequate parking space can reduce patient comfort and time efficiency, ultimately affecting their overall experience and satisfaction with the provided service.

CONCLUSION

Based on a study of 85 patients at Puskesmas Surabaya using Importance Performance Analysis (IPA), no aspects of service were found in the priority quadrant (Quadrant II). This indicates that nearly all patients were satisfied with the aspects addressed in the questionnaire. However, several aspects were identified as low-priority, including the speed of the registration process, the speed, accuracy, and friendliness of staff (doctors and nurses), the comfort of the waiting area, and the availability of adequate parking.

The results of the Kano method calculation revealed that all listed variables are classified as One Dimensional (O), indicating that an increase in customer satisfaction will correspond to an improvement in these attributes. Conversely, satisfaction will decrease if the performance of these attributes declines.

The integration of the IPA and Kano methods showed that although most service elements are satisfactory, several aspects require further attention and improvement, including the speed of the registration process, the attitude of staff in providing service, the condition of the waiting area, and the availability of sufficient parking space. These findings underscore the importance of focusing on these aspects to enhance overall service quality and patient satisfaction.

This study is limited to Puskesmas Surabaya with a small sample size, so future research should expand the sample to include other Puskesmas and consider external factors that may influence patient satisfaction, such as policies or socioeconomic conditions, to gain a more comprehensive understanding.

REFERENCE

[1] D. A. Hafizh, “Inovasi Pelayanan Publik; Studi Deskriptif tentang Penerapan Layanan e-Health dalam meningkatkan Kualitas Pelayanan Kesehatan di Puskesmas Pucangsewu Kota Surabaya,” *Kebijak. Dan Manaj. Publik*, vol. 4, no. 3, 2016.

[2] A. Rofiq, “Partisipasi masyarakat dalam keberhasilan pengembangan program posyandu lansia di Puskesmas Jagir Surabaya.” Universitas Airlangga, 2018.

[3] L. MUTIARA, “Pengaruh Self Efficacy Dan Religiusitas Terhadap Organizational Citizenship Behavior (Ocb) Pada Perawat Di Rumah Sakit Jiwa Tampan Provinsi Riau.” Fakultas Psikologi, 2023.

[4] V. Sesrianty, R. Machmud, and F. Yeni, “Analisa kepuasan pasien terhadap mutu pelayanan keperawatan,” *J. Kesehat. Perintis*, vol. 6, no. 2, pp. 116–126, 2019.

[5] I. Y. Kiling and B. N. Kiling-Bunga, “Pengukuran dan faktor kualitas hidup pada orang usia lanjut,” *J. Heal. Behav. Sci.*, vol. 1, no. 3, pp. 149–165, 2019.

[6] A. A. Saputro and N. N. Synthiawati, “Efektifitas Whatsapps Group Pada Pembelajaran Jarak Jauh Mata Kuliah Manajemen Olahraga Selama Covid-19,” *STAND J. Sport. Teach. Dev.*, vol. 2, no. 1, pp. 20–25, 2021.

[7] M. Ayu, S. Hadi, Y. Utomo, Y. B. Pramana, and H. Suntoko, “Implementing QFD for

improving service quality and social fund management," *WAKTU J. Tek. UNIPA*, vol. 23, no. 1, pp. 16–22, 2025.

[8] P. N. Farida, A. Kurniawan, and D. Amelia, "Analisis Tingkat Kepuasan Masyarakat Terhadap Pelayanan BPJS Kesehatan Cabang Utama Surabaya Dengan Metode Customer Satisfaction Indeks dan Importance Performance Analysis," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 16, no. 1, pp. 462–473, 2023.

[9] H. Winarno and T. Absor, "Analisis Kualitas Pelayanan Dengan Metode Service Quality (Servqual) Dan Importance Performance Analysis (Ipa) Pada Pt. Media Purna Engineering," *J. Manaj. Ind. dan Logistik*, vol. 1, no. 2, pp. 146–160, 2017.

[10] B. S. Santoso, M. F. Anwar, and S. Hermawati, "Analisis Kualitas Website Menggunakan Metode Webqual Dan Importance-Performance Analysis (IPA) Pada Situs Kaskus," *no. Sept.*, 2015.

[11] A. D. W. I. H. Hendriawan, "Peningkatan Kualitas Kerupuk Rambak Dengan Metode Servqual Dan Kano Di Umkm Ciptarasa Desa Tiremenggal, Dukun, Gresik." Universitas PGRI Adibuana Surabaya, 2020.

[12] P. Prihono and R. Migrihani, "Pengaruh Perilaku Konsumen Terhadap Keputusan Pembelian Sepeda Motor Dengan Menggunakan Metode Kano," *Waktu J. Tek. UNIPA*, vol. 16, no. 1, pp. 49–61, 2018.

[13] N. B. Puspitasari, H. Suliantoro, and L. Kusumawardhani, "Analisis kualitas pelayanan dengan menggunakan integrasi Importance Performance Analysis (IPA) dan model Kano (studi kasus di PT. Perusahaan Air minum Lyonnaise Jaya Jakarta)," *J@ ti Undip J. Tek. Ind.*, vol. 5, no. 3, pp. 185–198, 2010.

[14] S. Maulidiyah, "Analisis Dan Perbaikan Kualitas Pelayanan Pusat Layanan Terpadu (Plt) Di Perguruan Tinggi XYZ Dengan Metode Servqual Dan Kano Model p. 6," *SI. Inst. Teknol. Sepuluh Nop.*, 2021.

Application of ARIMAX-LSTM Model in Forecasting the Price of Broiler Chicken in Central Java

Divayanti Febri Sakina [(1)], Trimono [(2)], Amri Muhaimin [(3)]

[1,2,3]Universitas Pembangunan "Veteran" Jawa Timur, Indonesia

Jalan Rungkut Madya, Surabaya 60294, East Java, Indonesia
e-mail: divayantifebrisakina20@gmail.com[(1)], trimono.stat@upnjatim.ac.id[(2)]
amri.muhaimin.stat@upnjatim.ac.id[(3)]

ABSTRAK

Perekonomian Jawa Tengah tumbuh 4,98% pada 2023 dengan sektor perdagangan sebagai penggerak utama, termasuk komoditas daging ayam ras yang produksinya meningkat dari 621.718,06 ton (2021) menjadi 791.997,10 ton (2023.) Namun, harga komoditas ini mengalami fluktuasi yang cukup besar, terutama dipengaruhi oleh faktor-faktor eksternal seperti permintaan yang meningkat selama periode hari libur nasional dan harga produk substitusi seperti telur ayam dan daging sapi yang dapat mempengaruhi daya beli daging ayam ras. Data harga daging ayam, harga telur ayam, dan harga daging sapi diperoleh dari *website* resmi PIHPS (Pusat Informasi Harga Pangan Strategis), sedangkan data pekan sebelum libur diperoleh menggunakan pustaka Python "*holidays*". Penelitian ini mengembangkan model Hybrid ARIMAX-LSTM untuk memprediksi harga daging ayam secara lebih akurat. Model ARIMAX digunakan untuk menangkap pola linier dari harga ayam telur dengan mempertimbangkan variabel eksternal (harga telur, daging sapi, dan pekan libur nasional) sementara LSTM menangkap pola non-linier residual yang tidak dapat dijelaskan oleh model ARIMAX. Hasilnya menunjukkan bahwa model Hybrid menghasilkan MAPE 1,19%, lebih akurat dibandingkan ARIMAX tunggal (MAPE 1,38%). Prediksi harga Januari 2025 berkisar Rp35.300 – Rp35.900/kg, menunjukkan stabilitas tanpa fluktuasi ekstrem. Penelitian ini memberikan solusi prediktif yang dapat digunakan oleh pemerintah dan pelaku usaha dalam pengendalian harga serta stabilisasi pasar.
Kata kunci: ARIMAX, Harga Daging Ayam Ras, Hybrid, LSTM., Prediksi Harga

ABSTRACT

Central Java's economy grew 4.98% in 2023 with the trade sector as the main driver, including the broiler chicken meat commodity whose production increased from 621,718.06 tons (2021) to 791,997.10 tons (2023). However, the price of this commodity experiences considerable fluctuations, mainly influenced by external factors such as increased demand during the national holiday period and the price of substitute products such as chicken eggs and beef that can affect the purchasing power of broiler chicken meat. Data on chicken meat prices, chicken egg prices, and beef prices were obtained from the official website of PIHPS (Strategic Food Price Information Center), while data for the week before the holiday was obtained using the Python library "holidays". This research develops a Hybrid ARIMAX-LSTM model to predict chicken meat prices more accurately. The ARIMAX model is used to capture the linear pattern of chicken egg prices by considering external variables (egg prices, beef, and national holidays), while the LSTM captures non-linear residual patterns that cannot be explained by the ARIMAX model. The results show that the Hybrid model produces a MAPE of 1.19%, which is more accurate than the single ARIMAX (MAPE 1.38%). The predicted January 2025 price ranges from IDR 35,300 - IDR 35,900/kg, showing stability without extreme fluctuations. This research provides a predictive solution that can be used by the government and businesses in price control and market stabilization.
Keywords: *ARIMAX, Broiler Chicken Price, Hybrid, LSTM, Price Prediction*

INTRODUCTION

Central Java's economy grew by 4.98% in 2023 with GRDP reaching Rp1,696.79 trillion, where trade became the main sector due to its role as a distribution center for goods and services [1]. In this case, broiler chicken meat is one of the livestock commodities traded at fluctuating prices and is the most widely produced livestock. On Badan Pusat Statistik website, in 2021 the province of Central Java produced 621,718.06 tons of broiler chicken meat, then continued to increase to 791,997.10 tons in 2023 exceeding the province's needs by 446,700.34 tons, thus experiencing a surplus between supply and demand [2], [3]. On the website of Pusat Informasi Harga Pangan Strategis Nasional (PIHPS), the average price of broiler chicken meat in Central Java province before fasting was Rp.35,802/kg, but during the fasting month until before Eid the price of broiler chicken meat rose to Rp.38,810/kg [4]. Then on April 16, 2024 the price of chicken meat jumped to Rp.40,200/kg and decreased to Rp.35,700/kg on May 01, 2024 [4]. With the imbalance between the amount of production and the need for broiler chicken meat, making it one of the commodities that often experience price fluctuations. Fluctuating prices will cause the value of profit or loss to fluctuate and become inconsistent [5].

Fluctuations in the price of broiler chicken meat are influenced by the demand for broiler chicken meat which tends to increase ahead of holidays or national holidays, the uneven supply of broiler chicken meat in various markets, and price comparisons of substitute products such as the price of eggs and beef [6], [7]. When the price of chicken meat rises high enough, some consumers will switch to other animal products as a more affordable alternative. Inappropriate price control can trigger wider economic impacts, so price prediction is important to help manage the risk of price fluctuations in the market. Methods such as ARIMA (Autoregressive Integrated Moving Average) are often used for short-term predictions by looking at past trends, but have the disadvantage of ignoring external variables as supporting factors [8]. Based on this, the ARIMA model continues to evolve into ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables) to improve accuracy, but is still limited to linear patterns [9]. Therefore, a method that can capture linear and non-linear patterns is needed, namely using a Hybrid model [10].

For a model that can capture non-linear patterns well is the LSTM (Long Short Term Memory) model, which is a recurrent neural network specifically designed to understand complex patterns in sequential data [11]. LSTM has three gates (input, output, forget) that filter relevant information, making it more adaptive to dynamic fluctuations than linear models [12]. Therefore, the Hybrid ARIMAX-LSTM approach was developed by reprocessing the residual results from the linear model into a non-linear model to improve the prediction accuracy over the base model [13]. The main advantage of this model lies in the ability of LSTM to recognize residual patterns that are not captured by ARIMAX, while utilizing exogenous variables. Until now, there are still few studies that focus on broiler price forecasting by considering external variables (prices of substitute products and the approach of national holidays). So this research aims to achieve better and more accurate accuracy in forecasting the price of broiler chicken by considering external variables with the Hybrid ARIMAX-LSTM model.

In previous research that compared ARIMA, ARIMAX, and Hybrid ARIMAX-LSTM models to predict the value of Indonesia's oil and gas imports by considering the factors of world crude oil prices, currency exchange rates, and inflation. The results show that the Hybrid ARIMAX-

LSTM (0,1,2) model with crude oil price and inflation variables gets the best evaluation with a MAPE value of 8.01% on test data and 9.58% on training data, compared to other models [14]. Based on the results of the above research, this research proposes a new approach in predicting the price of broiler chicken meat by considering external variables (prices of substitute products and ahead of national holidays) from January 2019 to December 2024 using the ARIMAX model combined (Hybrid) with the LSTM model which will be evaluated by Mean Absolute Percentage Error (MAPE). This research also forecasts the price of broiler chicken meat in January 2025 on weekdays (Monday to Friday) using known values of exogenous variables.

METHOD

In this research, the price data for broiler chicken meat, eggs, and beef are sourced from Pusat Informasi Harga Pangan Strategis (PIHPS) website [4]. Meanwhile, the week before public holidays variable was created using the Python holidays library, which is also used to identify the days leading up to public holidays. The data period of this research is from January 2019 to December 2024 on weekdays (Monday - Friday) with a total of 1566 rows. The following is the flow of analysis in this research:

a. The data that has been collected is combined and data preprocessing is carried out by overcoming missing values and checking outliers so that it becomes a form of time series data which is divided into 1556 rows of training data and 10 testing data
b. In the training data, stationarity testing is carried out using the Augmented Dickey Fuller (ADF) method followed by a differencing process on variables that are not stationary, then the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are carried out to determine the optimum lag until ARIMAX modeling is carried out using the best order.
c. Forecasting results from ARIMAX modeling on training data and test data, residual calculations are carried out and residual assumption tests are carried out using the Autocorrelation (White Noise) test with L-jung Box. If the assumption test is not met, proceed to the LSTM modeling process on the residual data.
d. The residual data is divided into training data and test data, and will be converted into a temporal format by forming time steps whose number is adjusted to the data division in ARIMAX modeling.
e. LSTM modeling that has been built on the training data is performed to produce new residual forecasts
f. Forecasting test data using the best LSTM model to produce new residual forecasts
g. The residual data forecasting results from the LSTM model are combined with the forecasting results from the ARIMAX model or called the Hybrid process
h. Forecasting for January 2025 with the Hybrid ARIMAX-LSTM model that has been built with known exogenous variable data.

For the workflow of this research is shown in Figure 1.

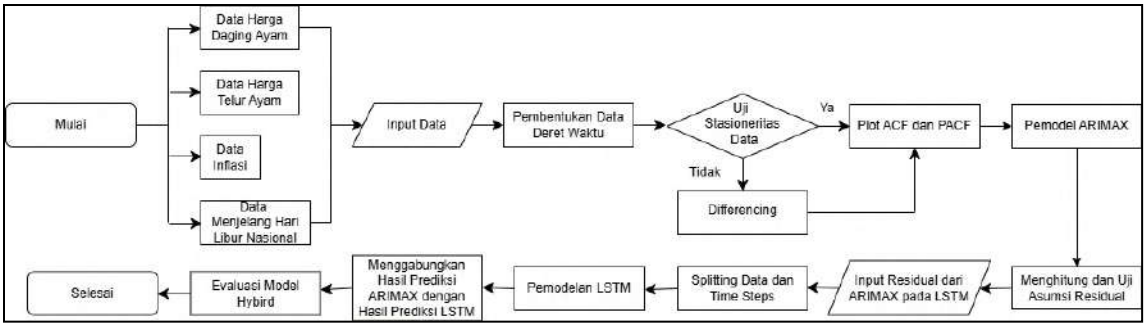


Figure 1. Research Workflow

Hybrid ARIMAX-LSTM

Hybrid method was first published by Zhang (2003), which combines ARIMA model with Neural Network to forecast time series data [15]. Time series data is composed of autocorrelation patterns that have linear and non-linear components. In this research, the linear component will be handled by the ARIMAX model, which is tasked with recognizing the linear pattern of the data as well as the influence of external variables. Meanwhile, the non-linear component will be handled by the LSTM model, which is able to capture complex patterns and non-linear relationships remaining in the residuals of the ARIMAX model. In general, time series data is expressed by equation (1).

$$y_t = L_t + N_t \quad (1)$$

Where L_t represents the linear component predicted by ARIMAX, and N_t represents the non-linear component predicted by LSTM from the residuals of the linear component. The final result of this Hybrid method is the sum of the linear and non-linear predictions, resulting in more accurate forecasting. The general form of the Hybrid method result is expressed in equation (2).

$$y'_t = L'_t + N'_t \quad (2)$$

Mean Absolute Percentage Error (MAPE)

To assess whether the model is accurate or not in forecasting chicken meat prices by considering external variables can be done using MAPE evaluation. MAPE is used to measure the error rate between the original data and the prediction results in the form of a percentage [16], [17]. The general form of MAPE is stated in equation (3).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{X_t - F_t}{X_t}\right| \times 100\% \quad (3)$$

If the MAPE percentage result is smaller or less than 10%, the prediction results are very accurate, whereas if the MAPE percentage result is above 50%, the prediction results are not accurate [18].

RESULT AND DISCUSSION

Initial Data Exploration

Before entering into the data preparation process, the data is divided into two parts: training data or in sample data 1,556 rows (period January 01, 2019 to December 21, 2024), and test data or out sample data 10 rows (period December 22, 2024 to December 31, 2024). This division is done sequentially based on time (time-series split), so the test data is out-of-sample forecasting data

that is never seen by the model during the training process. Furthermore, missing value and outlier checks were carried out on the training data and test data. From the check, in the training data there were 48 missing data in the columns "Price_Meat_Chicken", 'Price_Egg', and "Price_Meat_Cow", while in the test data there were no missing values. The missing data is resolved by interpolating the data based on time [19][20], where the missing data is filled in by estimating the values before and after based on time. Then, for the results of checking outliers in the training data, there are 14 outlier data in the "Price_Meat_Chicken" column, while in the test data there is 1 outlier data in the 'Price_Eggs' column and there are 4 outlier data in the "Price_Cow Meat" column. The outlier data is not addressed because the price data describes real events as a result of external factors [14]. Figure 2 shows the distribution of chicken meat prices in the training data. The graph highlights several periods of significant price increases, such as in 2020 when the price rose from Rp 28,000/kg to Rp 42,000/kg. The graph also shows a random pattern and no seasonal pattern.

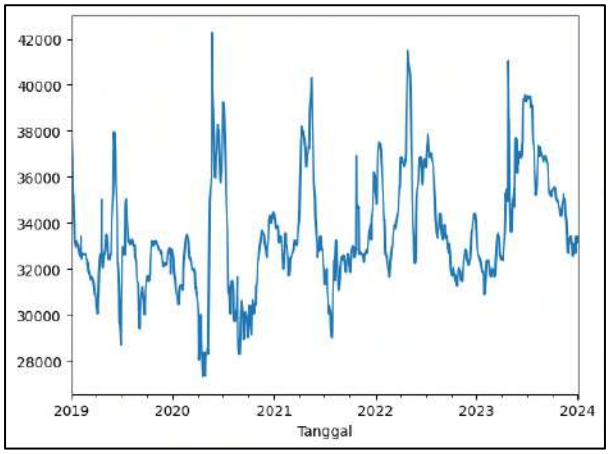


Figure 2. Distribution of Breast Chicken Price Data

In the training data, stationarity testing is carried out, as a condition for forecasting with the ARIMAX model in the "Chicken_Meat_Price" column. To test stationarity in mean or trend using the ADF (Augmented Dickey-Fuller) test method. For the ADF method with the hypothesis:

H_0 : Chicken Meat Price Data is Not Stationary

H_1 : Chicken Meat Price Data is Stationary

Significance value: α= 5%

Test criteria: H_0 is rejected if p-value < 5%

The results of stationarity testing with the ADF test on each variable are shown in Table 1 as follows:

Table 1. ADF Test Results of Chicken Meat Price

Variable	Test Statistic	p-value	Result
Chicken_Meat_Price	-5.449	0.000003	Stationary

Based on the stationarity test results in Table 1, it shows that the "Chicken_Meat_Price" column is stationary, so that the variable does not need to be transformed or differenced and can proceed to the next process. for the value of D (d) = 0.

Furthermore, identifying the order of the model based on the analysis of ACF and PACF plots on the training data to assist in determining the optimum lag and finding a combination of orders as ARIMAX parameters. To display ACF and PACF plots, the Python library statsmodels.graphics.tsaplots can be used. The results of the ACF plot on the variable "Price_Daging_Ayam" show an exponential decline pattern, which indicates there is no significant relationship between lags, resulting in the value of MA(q) = 0. Meanwhile, the results of the PACF plot on the same variable show a cut-off after the 1st, 2nd, and 3rd lags, so the possible AR(p) values are 1, 2, or 3. The ACF and PACF plots of the variable "Chicken_Meat_Price" are shown in Figure 3 for the order results of the plot can be a benchmark in determining the order combination, but to produce the best ARIMAX model that is significant in all parameters, another combination experiment can be done by trying other order combinations.

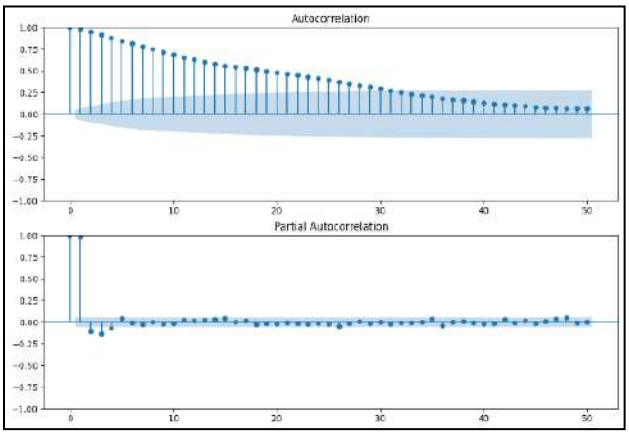


Figure 3. ACF and PACF Plot of Chicken Meat Price

ARIMAX Modeling

In the training data and test data, division is made into target variables (y) and external variables (x). The target variable is the price of chicken meat and the external variables are the price of eggs, the price of beef, and the week before the holiday. So that this process will produce "y_train", "x_train", 'y_test', and "x_test" to forecast the price of broiler chicken based on external variables. The ARIMAX model is then built to predict the price of broiler chicken meat in "y_train" by considering the variables in "x_train", using the best order combination, which is [0,0,3]. This combination means that the model does not use an autoregressive (AR) component, does not perform differencing because the data is stationary, and only involves three lags in the moving average (MA) to account for 3 lag errors to correct noise. For ARIMAX modeling results, it is expressed in equation (4).

$$Y_t = 0.1395(Harga\_Telur_t) + 0.2487(Harga\_Daging\_Sapi_t) + 490.47(Pekan\_Sebelum\_Libur_t) + e_t + 1.3760 \times e_{t-1} + 1.2143 \times e_{t-2} + 0.6255 \times e_{t-3} \quad (4)$$

The results of the ARIMAX modeling analysis are shown in

Table ***2*** as follows:

Table 2. ARIMAX Modeling Results

Variable	Coefficient	p-value	Result
Harga_Telur	0.1395	0.000	Significant
Harga_Daging_Sapi	0.2487	0.000	Significant
Pekan_Sebelum_Libur	490.47	0.000	Significant
MA (1)	1.3760	0.000	Significant
MA (2)	1.2143	0.000	Significant
MA (3)	0.6255	0.000	Significant

Based on the results of the ARIMAX analysis on the model parameters shown in Table 2, the variables of egg prices, beef prices, and the week before the holiday $p - value < 0.5$, indicating that the three variables are statistically significant and have a direct effect on the demand for broiler chicken meat that causes an increase or decrease in prices. Then, a Rp1 increase in the price of eggs and beef is estimated to increase the price of chicken meat by Rp0.14 and Rp0.25, respectively. Meanwhile, the week before the national holiday had the biggest impact, with chicken meat prices increasing by around Rp490.47 compared to a normal day.

Residual Assumption Test

After the forecasting results of the ARIMAX model are obtained, Ljung-Box testing is then carried out on the ARIMAX residuals to determine whether the data meets white noise (no autocorrelation) or not (has autocorrelation). The L-jung Box test uses the hypothesis:

H_0 : Residuals meets white noise

H_1 : Residuals does not meet white noise

Significance value: α= 5%

Test criteria: H_0 is rejected if p-value < 5%

The results of the white noise test with the L-jung Box test on the residual data are shown in Table 3 as follows:

Table 3. Residual Assumption Test Results

Test L-jung Box	p-value	Result
167.74	0.000	Residuals does not meet white noise

Based on Table 3 L-Jung Box testing shows the test statistic value is very large and the p-value is much smaller than the significant value, then reject H_0. The conclusion of the test shows strong evidence that the residuals are not random or still have a pattern. Since the residual data has not met the white noise assumption test, the residual data will be further processed into LSTM modeling to handle non-linear patterns.

LSTM Modeling

Before entering into LSTM modeling, preprocessing is carried out by converting residual data into 2 dimensions. Furthermore, the residual data is normalized / scaling the residual data using the MinMaxScaler method. Then, the residual data will be divided into 2 parts, namely, training data and test data with data division of 1,546 and 10 data. The time steps in this research are 5, because chicken meat prices have complex patterns and are influenced by historical movements within a certain period of time and can reflect relevant trends and fluctuations. Furthermore, the

data will be converted into 3 dimensions (number of samples, number of time steps in each sample, and number of features) to adjust the format of the LSTM model.

The basic LSTM model used in this research has 3 LSTM layers, 3 Dropout layers, and 2 Dense layers. In the model training process, EarlyStopping is added to stop training automatically if there is no improvement in the validation loss value after several consecutive epochs, so as to prevent overfitting and speed up training time. The following LSTM model structure is shown in Table 4 which is used to train and test data.

Table 4. Structure Model LSTM

Model LSTM	
LSTM	128, Return Sequences =True
Dropout	0,2
LSTM	64, Return Sequences =True
Dropout	0,1
LSTM	32, Return Sequences =False
Dropout	0,1
Dense	16, Activation='RELU'
Dense	1
Optimizer	Adam
Epoch	7
Batch Size	32

Hybrid ARIMAX-LSTM Modeling

The forecasting results on the test data are then combined with the forecasting results from the ARIMAX model, forming a hybrid approach. This process is known as Hybrid ARIMAX-LSTM modeling. Evaluation of the Hybrid forecasting results was carried out using the MAPE metric, which resulted in a value of 1.19%. When compared to the MAPE value of the ARIMAX (1.38%) model separately, the Hybrid model shows better performance. This shows that the Hybrid approach is able to improve accuracy in forecasting the price of broiler chicken meat in Central Java. Figure 4 shows a comparison of the original data, ARIMAX prediction results, and ARIMAX-LSTM prediction results.

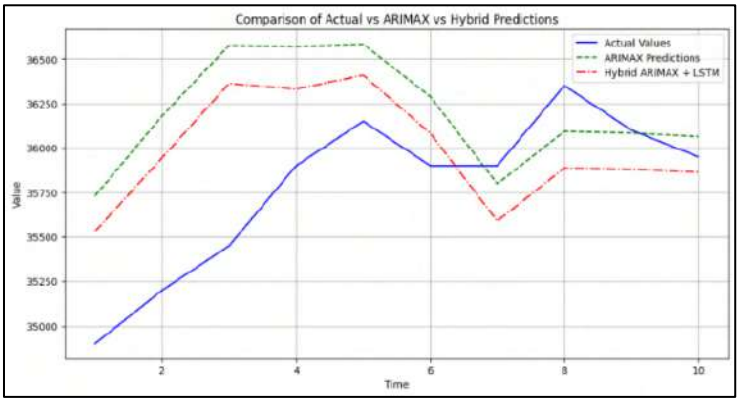


Figure 4. Comparison Chart of Original Data, ARIMAX, ARIMAX-LSTM

Forecasting January 2025

Furthermore, from the ARIMAX and LSTM models that have been built and the Hybrid ARIMAX -LSTM process is carried out, to forecast the price of broiler eggs for January 2025. This model utilizes historical egg price data from January 2019 to December 2024, as well as known exogenous variables in January 2025. The resulting price prediction for January 2025 shows price variations that are still within a reasonable range, without unfounded extreme spikes, indicating that the model can produce stable, accurate and reliable output. Forecast results for January 2025 are shown in Table 5.

Table 5. Forecasting Results for January 2025

Tanggal	Harga_ Telur	Harga_ Daging_Sapi	Pekan_Sebelum_ Libur	Prediksi_Harga_ Daging_Ayam
01/01/2025	30100	127400	1	35.372
02/01/2025	30100	128050	0	35.518
03/01/2025	29950	128150	0	35.966
06/01/2025	29950	128150	0	35.928
07/01/2025	29000	128200	0	35.760
08/01/2025	28800	128100	0	35.685
09/01/2025	28600	128200	0	35.692
10/01/2025	28400	128200	0	35.665
13/01/2025	27550	128250	0	35.551
14/01/2025	27350	128250	0	35.515
15/01/2025	27200	128250	0	35.492
16/01/2025	27050	128250	0	35.472
17/01/2025	26850	128250	0	35.442
20/01/2025	26350	128250	1	35.861
21/01/2025	26200	128250	1	35.839
22/01/2025	25900	128250	1	35.797
23/01/2025	25750	128250	1	35.776
24/01/2025	25700	128250	1	35.768
27/01/2025	25500	127300	1	35.504
28/01/2025	25500	127300	1	35.503
29/01/2025	25500	127300	1	35.503
30/01/2025	26050	128250	0	35.326
31/01/2025	25950	128250	0	35.311

CONCLUSION

In the research of predicting the price of broiler chicken meat in Central Java by considering external variables in the period January 2019 to December 2024 on weekdays using Hybrid ARIMAX-LSTM modeling, it can be concluded that

1) The ARIMAX (0,0,3) model can be used to capture linear patterns by considering exogenous variables (egg prices, beef, and holidays) well, but the residual assumption test results show

that the residual results do not meet white noise. Therefore, a hybrid process is performed with another model

2) The LSTM model is used to capture non-linear patterns of ARIMAX residuals that can improve prediction accuracy.
3) Hybrid ARIMAX-LSTM results provide better accuracy (MAPE 1.19%) than a single ARIMAX model (MAPE 1.38%).
4) The Hybrid model successfully predicts the price of broiler chicken meat influenced by external variables for January 2025 with a range of Rp35,300 - Rp35,900/kg, showing stability without extreme spikes.

Thus, the development of the Hybrid ARIMAX-LSTM model is proven to be able to provide potential price estimates to be used as a basis for decision making by market players and policy makers in dealing with future food price dynamics.

REFERENCE

[1] Badan Pusat Statistik, "Perekonomian Jawa Tengah Tahun 2023 Mengalami Pertumbuhan Sebesar 4,98 persen - Berita," jateng.bps.go.id. Accessed: Feb. 13, 2025. [Online]. Available: https://jateng.bps.go.id/id/news/2024/02/05/642/perekonomian-jawa-tengah-tahun-2023-mengalami-pertumbuhan-sebesar-4-98-persen.html

[2] Badan Pusat Statistik Indonesia, "Produksi Daging Ayam Ras Pedaging menurut Provinsi - Tabel Statistik - Badan Pusat Statistik Indonesia." Accessed: Feb. 12, 2025. [Online]. Available: https://www.bps.go.id/id/statistics-table/2/NDg4IzI=/produksi-daging-ayam-ras-pedaging-menurut-provinsi.html

[3] Badan Pusat Statistik Indonesia, "Peternakan Dalam Angka 2023 - Badan Pusat Statistik Indonesia." Accessed: Feb. 11, 2025. [Online]. Available: https://www.bps.go.id/id/publication/2023/12/22/5927b06e1dcde219f76cec59/peternakan-dalam-angka-2023.html

[4] PIHPS Nasional, "Berdasarkan Daerah - PIHPS." Accessed: Feb. 13, 2025. [Online]. Available: https://www.bi.go.id/hargapangan/TabelHarga/PasarTradisionalDaerah

[5] T. Trimono, I. G. S. M. Diyasa, K. M. Hindrayani, and M. Idhom, "Model ARIMA-ARCH/GARCH dan Ensemble ARIMA-ARCH/GARCH untuk Prediksi Kerugian pada Harga Komoditas Pertanian," *Prosiding Seminar Nasional Sains Data*, vol. 1, no. 1, pp. 1–11, Oktober 2021, doi: https://doi.org/10.33005/senada.v1i01.11.

[6] D. W. Lestari and S. K. Dini, "Forecasting The Price Of Shallots And Red Chilies Using The ARIMAX Model," *EKSAKTA: Journal of Sciences and Data Analysis*, vol. 5, no. 1, pp. 42–49, 2024.

[7] L. H. Hasibuan, D. M. Putri, M. Jannah, and S. Musthofa, "Analisis Metode Single Exponential Smoothing Dan Metode Regresi Linear Untuk Prediksi Harga Daging Ayam Ras," *Math Educa Journal*, vol. 6, no. 2, Art. no. 2, Oct. 2022, doi: 10.15548/mej.v6i2.3872.

[8] T. M. P. Wiryawanto, Z. Hawani, and M. A. Ramadhani, "Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 16, no. 1, Art. no. 1, Jul. 2023, doi: 10.36456/jstat.vol16.no1.a6633.

[9] R. Arianti, S. Sahriman, and L. P. Talangko, "Model ARIMA dengan Variabel Eksogen dan GARCH pada Data Kurs Rupiah," *ESTIMASI: Journal of Statistics and Its Application*, pp. 41–48, 2022.

[10] A. T. Damaliana, K. M. Hindrayani, and T. M. Fahrudin, "Hybrid Holt Winter-Prophet method to forecast the num-ber of foreign tourist arrivals through Bali's Ngurah Rai Airport," *IJDASEA (International Journal of Data Science, Engineering, and Analytics)*, vol. 3, no. 2, Art. no. 2, 2023, doi: 10.33005/ijdasea.v3i2.8.

[11] P. N. Yulisa, M. A. Haris, and P. R. Arum, "Peramalan Nilai Ekspor Migas di Indonesia dengan Model Long Short Term Memory (LSTM) dan Gated Recurrent Unit (GRU)," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 16, no. 1, Art. no. 1, Jul. 2023, doi: 10.36456/jstat.vol16.no1.a6121.

[12] P. A. Riyantoko, T. M. Fahrudin, D. A. Prasetya, T. Trimono, and T. D. Timur, "Analisis Sentimen Sederhana Menggunakan Algoritma LSTM dan BERT untuk Klasifikasi Data Spam dan Non-Spam," *Prosiding Seminar Nasional Sains Data*, vol. 2, no. 1, Art. no. 1, Dec. 2022, doi: 10.33005/senada.v2i1.53.

[13] M. R. Susila, M. Jamil, and B. H. Santoso, "Akurasi Model Hybrid ARIMA-Artificial Neural Network dengan Model Non Hybrid pada Peramalan Peredaran Uang Elektronik di Indonesia," *Jambura Journal of Mathematics*, vol. 5, no. 1, Art. no. 1, Jan. 2023, doi: 10.34312/jjom.v5i1.14889.

[14] A. S. R. Zega, A. K. Hidayat, N. T. Jannah, and F. Kartiasih, "Selecting The Best Model For Forecasting Indonesia's Oil And Gas Import Value Using Arimax And Arimax-Lstm," *Dynamic Management Journal*, vol. 8, no. 4, pp. 912–941, 2024, doi: http://dx.doi.org/10.31000/dmj.v8i4.

[15] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

[16] K. M. Hindrayani, I. G. S. Mas Diyasa, P. A. Riyantoko, and T. M. Fahrudin, "Studi Literatur Mengenai Prediksi Harga Saham Menggunakan Machine Learning," *santika*, vol. 1, pp. 71–75, Nov. 2020, doi: 10.33005/santika.v1i0.20.

[17] A. Muhaimin and T. Trimono, "Stock Price Modeling with Geometric Brownian Motion and Value with Risk PT Ciputra Development TBK," *Nusantara Science and Technology Proceedings*, pp. 177–186, May 2023, doi: 10.11594/nstp.2023.3329.

[18] T. Trimono, A. Sonhaji, and U. Mukhaiyar, "Forecasting Farmer Exchange Rate In Central Java Province Using Vector Integrated Moving Average," *Media Statistika*, vol. 13, no. 2, pp. 182–193, Dec. 2020, doi: 10.14710/medstat.13.2.182-193.

[19] R. K. Putri and M. Athoillah, "Identifikasi Berita Hoax Terkait Virus Corona Menggunakan Long Short-Term Memory," in Seminar Nasional Hasil Riset dan Pengabdian, 2022, pp. 506–513.

[20] B. Dissanayake, O. Hemachandra, N. Lakshitha, D. Haputhanthri, and A. Wijayasiri, "A Comparison of ARIMAX, VAR and LSTM on Multivariate Short-Term Traffic Volume Forecasting," *Conference of open innovations association, FRUCT*, pp. 564–570.

Rehabilitation and Law Enforcement as Optimal Controls in a Mathematical Model of Social Behavior

Nailul Izzati [(1)], Wahyuni Ningsih [(2)]

[1]Universitas Hasyim Asy'ari, [2]Politeknik Negeri Malang

[1]Jl. Irian Jaya No. 55 Tebuireng, Diwek, Jombang,
[2]Jl. Soekarno-Hatta No. 9 Jatimulyo, Lowokwaru, Malang
e-mail: nailulizzati@unhasy.ac.id[(1)], wahyuni_04@polinema.ac.id[(2)]

ABSTRAK

Perilaku sosial merupakan hasil interaksi antarindividu yang dapat membentuk kecenderungan menuju perilaku positif maupun menyimpang. Berdasarkan sudut pandang ini, dikembangkan suatu model matematika perilaku sosial yang membagi populasi menjadi kelompok kriminal dan non-kriminal. Penelitian-penelitian sebelumnya umumnya hanya mempertimbangkan penegakan hukum, seperti penangkapan dan pemenjaraan, sebagai strategi penanganan terhadap perilaku menyimpang, tanpa memasukkan aspek rehabilitasi. Penelitian ini mengkaji penerapan kendali optimal dalam model matematika perilaku sosial guna meminimalkan jumlah individu dalam kelompok kriminal, dengan rehabilitasi dan penegakan hukum sebagai variabel kendali. Permasalahan kendali optimal diselesaikan menggunakan Prinsip Minimum Pontryagin, sedangkan simulasi numerik dilakukan dengan metode Forward-Backward Sweep. Hasil simulasi menunjukkan bahwa kombinasi strategi rehabilitasi dan penegakan hukum secara signifikan mampu menurunkan populasi kriminal dalam model.
Kata kunci: penegakan hukum; perilaku sosial; rehabilitasi

ABSTRACT

Social behavior is the result of interactions between individuals, which can lead to tendencies toward either positive or deviant behavior. From this perspective, a mathematical model of social behavior is developed, dividing the population into criminal and non-criminal groups. Previous studies generally considered only law enforcement strategies—such as arrest and imprisonment—as responses to deviant behavior, without incorporating the aspect of rehabilitation. This study examines the application of optimal control in a mathematical model of social behavior to minimize the number of individuals in the criminal group, using rehabilitation and law enforcement as control variables. The optimal control problem is solved using Pontryagin's Minimum Principle, and numerical simulations are performed using the Forward-Backward Sweep Method. The simulation results show that a combination of rehabilitation and law enforcement strategies can significantly reduce the criminal population in the model.
Keywords: *Law enforcement; Rehabilitation; Social behavior*

INTRODUCTION

Social behavior encompasses all forms of actions or responses performed by individuals in the context of interactions with other individuals or groups within society. This behavior reflects how a person acts, communicates, or responds to the social norms and values around them. Social behavior and criminal acts are closely related, as criminal acts often arise as deviations from the prevailing social norms within a society. A criminal act is a form of deviant behavior that violates formal laws (legislation). It is a type of social behavior, but one that is negative, as it harms others and disrupts social order. A simple scheme illustrating the relationship between social behavior and criminal acts is shown in Figure 1.

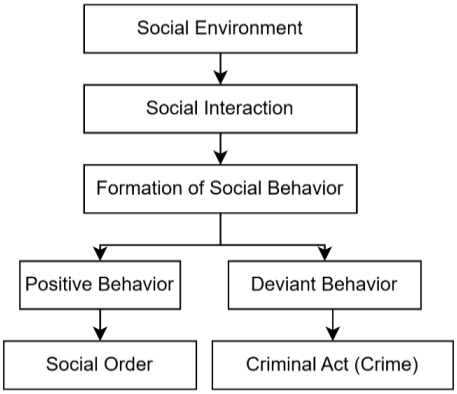


Figure 1. Diagram of the Relationship Between Social Behavior and Criminal Acts

One of the criminological theories from a social perspective, the Differential Association Theory, states that no behavior is inherited from one's parents. Patterns of criminal behavior are not passed down genetically but are learned through social interaction. Criminal behavior is acquired within groups through interaction and communication, and it is learned in those group settings [1]. Several studies in Indonesia that have discussed this theory include research on acts of terrorism [1], violence [2], cyberbullying [3], juvenile delinquency behavior such as “klitih” [4], cigarette use among teenagers [5], hacking cases [6], and drug abuse [7].

Social behavior related to criminal acts can be studied not only from a social science perspective. Using a mathematical approach, researchers have also examined social behavior, with mathematics serving as an alternative tool in combating crime. These studies include topics such as gang membership [8], financial crimes [9], armed groups [10], corruption [11], bullying [12], and drug abuse [13]. Some studies do not focus on a specific type of criminal act. For example, in the study by Abbas et al. [14], a population is divided into two compartments: criminal minded and non-criminal minded populations. In that study, law enforcement is applied as a means to reduce criminal behavior. This work was further developed in subsequent studies. Srivastav et al. [15] and Tripathi et al. [16] expanded the model by incorporating a Holling type II response function to describe interactions between criminal minded and non-criminal minded populations. Kumar & Abbas [17] and Kumar et al. [18] enhanced the model by considering age structure into the compartmental diagram. Izzati et al. [19] further extended the model by adding a religious population compartment.

In the study by Abbas et al. [14], only law enforcement was implemented as a strategy to reduce the criminal minded population—rehabilitation programs were not included. Law enforcement generally refers to the efforts of the state or legal authorities (such as the police, prosecutors, and courts) to ensure the enforcement of laws. This includes the arrest of offenders, judicial processes, sentencing, and execution of punishment (such as imprisonment or fines). On the other hand, rehabilitation is a more restorative approach that focuses on reintegrating offenders into society as law-abiding individuals. Rehabilitation is commonly applied to prisoners (e.g., vocational training, counseling), drug addicts (through medical and social rehabilitation), and juvenile offenders (via restorative justice approaches). In certain contexts, rehabilitation can be part of law enforcement, particularly in legal systems emphasizing corrective or restorative justice rather than purely punitive measures. For example, drug rehabilitation as an alternative to imprisonment, or rehabilitation within correctional institutions as part of prisoner reformation. However, not all forms of rehabilitation are considered law enforcement—for instance, purely medical rehabilitation (such as for accident victims) is not part of the legal system. In this study, rehabilitation programs are considered as an intervention for individuals who have committed criminal acts. From this foundation, the study explores an optimal control problem applied to a mathematical model of social behavior, where law enforcement and rehabilitation serve as the control variables.

METHOD

The steps carried out in this study are illustrated in the flowchart in Figure 2.

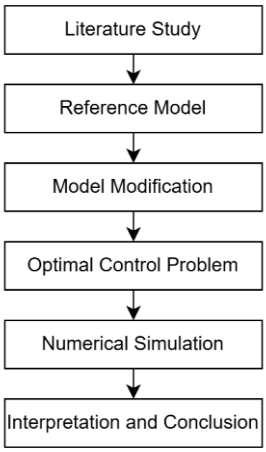


Figure 2. Research Flowchart

The first step is a literature review. In this phase, materials related to social behavior are collected, including both social science theories and mathematical models that researchers have previously discussed. The second step is to determine the model to be modified. This study builds upon the model developed by Abbas et al. [14], which models the social interaction between the criminal and non-criminal populations as follows:

$$\frac{dN_p}{dt} = \mu N_p \left(1 - \frac{N_p}{K}\right) - \alpha N_p C_p \quad (1)$$

$$\frac{dC_p}{dt} = -\gamma C_p + \alpha N_p C_p - l_c C_p \quad (2)$$

Where,

N_p : Non-criminal minded population.

C_p : Criminal minded population.

μ : Natural growth rate.

α : Interaction rate between non-criminal minded and criminal minded population.

γ : Natural death rate.

l_c : Law enforcement rate.

K : Carrying capacity.

The third step is model modification, which is carried out by incorporating law enforcement and rehabilitation into the model. Next, in the fourth step, the optimal control problem is applied. At this stage, Pontryagin's Minimum Principle is used to find an analytical solution. Then, in the following step, the Forward-Backward Sweep method is employed in simulations to solve the problem numerically. Finally, the results of the numerical simulations are interpreted, and conclusions are drawn.

RESULT AND DISCUSSION

This section discusses two types of optimal control problems. In the first optimal control problem, law enforcement (u_1) is applied as the control variable to minimize the criminal minded population. In the second problem, both law enforcement (u_1) and rehabilitation (u_2) are used as control variables to minimize the criminal minded population and the cost of rehabilitation.

Optimal Control Problem I

The mathematical model to be optimized is represented by the system of equations (3)–(4), with the objective function given by equation (5).

$$\frac{dN_p}{dt} = \mu N_p\left(1 - \frac{N_p}{K}\right) - \alpha N_p C_p \tag{3}$$

$$\frac{dC_p}{dt} = -\gamma C_p + \alpha N_p C_p - u_1 C_p \tag{4}$$

$$\min J = \int_0^T C_p\, dt \tag{5}$$

By applying Pontryagin's Minimum Principle, equations (6)–(12) are obtained as follows.

Hamiltonian:

$$H = C_p(t) + \lambda_1\left(\mu N_p\left(1 - \frac{N_p}{K}\right) - \alpha N_p C_p\right) + \lambda_2\left(-\gamma C_p + \alpha N_p C_p - u_1 C_p\right) \tag{6}$$

State and costate equations:

It is assumed that the optimal values of $u_1(t)$, $N_p(t), C_p(t)$ and $\lambda(t)$ are denoted by $u^*(t), N_p^*(t), C_p^*(t)$ and $\lambda^*(t)$, respectively.

$$\left(\frac{dN_p}{dt}\right)^* = \left(\frac{\partial H}{\partial \lambda_1}\right)_* = \mu N_p^*\left(1 - \frac{N_p^*}{K}\right) - \alpha N_p^* C_p^* \tag{7}$$

$$\left(\frac{dC_p}{dt}\right)^* = \left(\frac{\partial H}{\partial \lambda_2}\right)_* = -\gamma C_p^* + \alpha N_p^* C_p^* - u_1^* C_p^* \tag{8}$$

$$\left(\frac{d\lambda_1}{dt}\right)^* = -\left(\frac{\partial H}{\partial N_p}\right)_* = -\left(\lambda_1^*\left(\mu - \frac{2\mu N_p^*}{K} - \alpha C_p^*\right) + \lambda_2^* \alpha C_p\right) \tag{9}$$

$$\left(\frac{d\lambda_2}{dt}\right)^* = -\left(\frac{\partial H}{\partial C_p}\right)_* = -\left(1 - \alpha\lambda_1^* N_p^* + \lambda_2^*\left(-\gamma + \alpha N_p^* - u_1^*\right)\right) \quad (10)$$

Optimal condition:

The optimal control obtained in this problem is a bang-bang optimal control, which is expressed using the signum function

$$u_1^*(t) = \begin{cases} u_{1_{max}} & if\ Q^*(t) < 0 \\ u_1(t)\epsilon[u_{1_{min}}, u_{1_{max}}] & if\ Q^*(t) = 0 \\ u_{1_{min}} & if\ Q^*(t) > 0 \end{cases} \quad (11)$$

where $Q^*(t) = C_p^*\lambda_2^*$ [20].

Since the type of system in this study is a free-final time and free-final state system, the boundary conditions are

$$\lambda_1(T) = 0 \text{ and } \lambda_2(T) = 0. \quad (12)$$

Optimal Control Problem II

The mathematical model being optimized is given by the system of equations (13)-(14), and the objective function is defined by equation (15), where $A > 0$ represents the preference weight for reducing the intervention cost.

$$\frac{dN_p}{dt} = \mu N_p\left(1 - \frac{N_p}{K}\right) - \alpha N_p C_p \quad (13)$$

$$\frac{dC_p}{dt} = -\gamma C_p + \alpha N_p C_p - u_1 C_p - u_2 C_p \quad (14)$$

$$\min J = \int_0^T C_p + Au_2^2\, dt \quad (15)$$

By applying Pontryagin's Minimum Principle, equations (16)-(23) are obtained as follows.

Hamiltonian:

$$H = C_p + Au_2^2 + \lambda_1\left(\mu N_p\left(1 - \frac{N_p}{K}\right) - \alpha N_p C_p\right) + \lambda_2\left(-\gamma C_p + \alpha N_p C_p - u_1 C_p - u_2 C_p\right) \quad (16)$$

State and costate equations:

It is assumed that the optimal values of $u_1(t), u_2(t), N_p(t), C_p(t)$ and $\lambda(t)$ are denoted by $u^*(t), N_p^*(t), C_p^*(t)$, and $\lambda^*(t)$, respectively.

$$\left(\frac{dN_p}{dt}\right)^* = \left(\frac{\partial H}{\partial \lambda_1}\right)_* = \mu N_p^*\left(1 - \frac{N_p^*}{K}\right) - \alpha N_p^* C_p^* \quad (17)$$

$$\left(\frac{dC_p}{dt}\right)^* = \left(\frac{\partial H}{\partial \lambda_2}\right)_* = -\gamma C_p^* + \alpha N_p^* C_p^* - u_1^* C_p^* - u_2^* C_p^* \quad (18)$$

$$\left(\frac{d\lambda_1}{dt}\right)^* = -\left(\frac{\partial H}{\partial N_p}\right)_* = -\left(\lambda_1^*\left(\mu - \frac{2\mu N_p^*}{K} - \alpha C_p^*\right) + \lambda_2^*\alpha C_p\right) \quad (19)$$

$$\left(\frac{d\lambda_2}{dt}\right)^* = -\left(\frac{\partial H}{\partial C_p}\right)_* = -\left(1 - \alpha\lambda_1^* N_p^* + \lambda_2^*\left(-\gamma + \alpha N_p^* - u_1^* - u_2^*\right)\right) \quad (20)$$

Optimal conditions:

The optimal control $u_1^*(t)$ is of bang-bang type, meaning it only takes on extreme values—either the minimum or the maximum.

$$u_1^*(t) = \begin{cases} u_{1_{max}} & if\ Q^*(t) < 0 \\ u_{1_{min}} & if\ Q^*(t) > 0 \end{cases} \quad (21)$$

where $Q^*(t) = C_p^* \lambda_2^*$.

Whereas

$$u_2^*(t) = \frac{\lambda_2^* C_p^*}{2A} \tag{22}$$

where $0 \leq u_2^* \leq u_{2_{max}}$.

Since the type of system in this study is a free-final time and free-final state system, the boundary conditions are given by

$$\lambda_1(T) = 0 \text{ and } \lambda_2(T) = 0. \tag{23}$$

The system of equations obtained by applying Pontryagin's Minimum Principle is then implemented in a numerical simulation. The system of state and costate differential equations is solved using the Forward-Backward Sweep method based on the fourth-order Runge-Kutta scheme. The state equations are solved using the forward method, while the costate equations are solved using the backward method.

Numerical Simulations

The parameter values used in the numerical simulation are $\alpha = 0.5$, $\mu = 1.3$, $\gamma = 1.9$, dan $K = 6$, with the initial population conditions set as $N_p(0) = 1$, and $C_p(0) = 1$ [14]. The simulation results shown in Figure 3 compare the mathematical model without control and with optimal control in Problem I. In Figure 3, it is evident that optimal control can significantly reduce the criminal minded population. Without intervention, the number of criminal minded individuals decreases slightly from the initial value $C_p(0) = 1$ to $C_p(T) = 0.953$, which is due to the natural death rate within the criminal minded population. However, with optimal control—specifically the implementation of law enforcement—the final value decreases to $C_p(T) = 0.086$. The profile of optimal control, i.e. law enforcement, is shown in Figure 4. From Figure 4, it can be observed that law enforcement is applied at full intensity (100%) for a specific period of time until it successfully suppresses the criminal minded population C_p and brings the system to equilibrium.

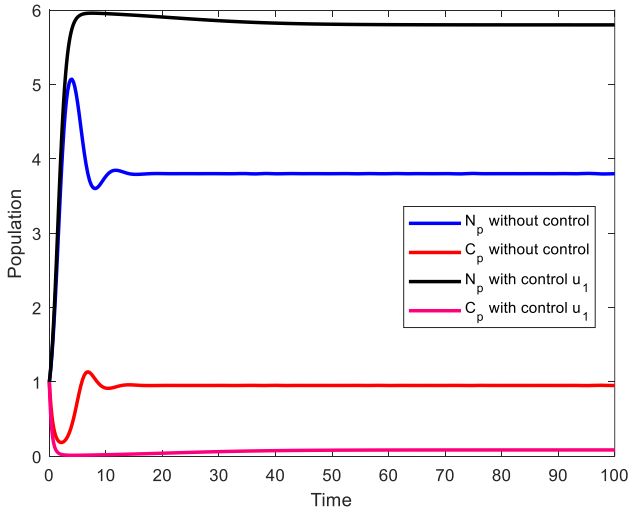


Figure 3. Number of criminal and non-criminal minded population over time in Problem I

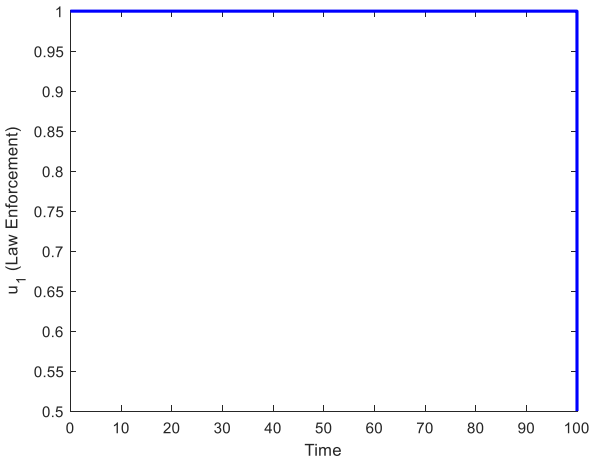


Figure 4. Law enforcement profile as optimal control on Problem I

Figure 5 presents the results of a comparison between the model without optimal control and the model with control in Optimal Control Problem II. In Figure 5, it is shown that the applied optimal control successfully reduces the criminal population significantly. By implementing law enforcement and rehabilitation on the criminal population, the number is reduced from $C_p(0) = 1$ to $C_p(T) = 0.012$. The profiles of the control variables are shown in Figures 6 and 7. In Figure 6, it can be seen that law enforcement is applied at full capacity (100%) up to a certain time period, until it effectively suppresses the criminal minded population C_p and brings it to equilibrium. In contrast, Figure 7 shows that rehabilitation is not applied at full intensity. This is due to the inclusion of rehabilitation cost minimization in the objective function. Meanwhile, the cost of law enforcement is not minimized in the model.

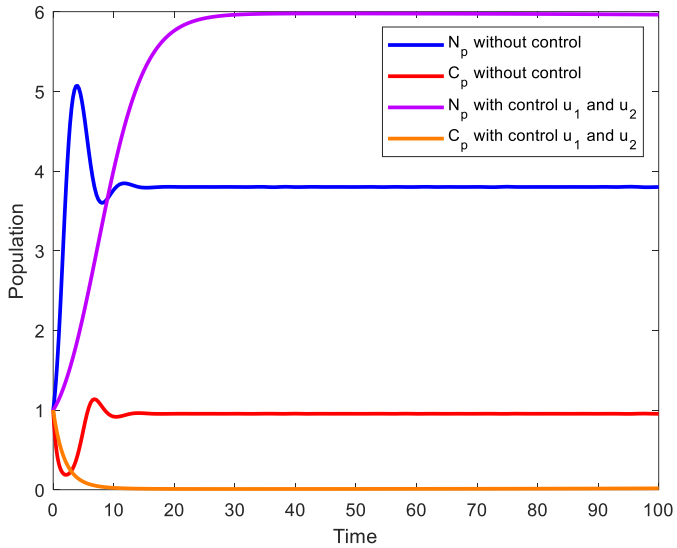


Figure 5. Number of criminal and non-criminal population over time in Problem II

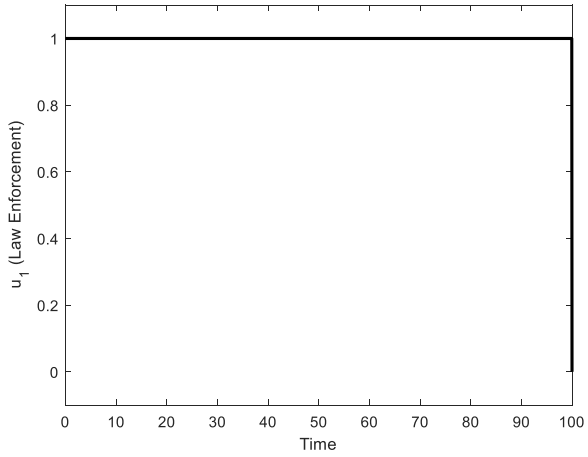


Figure 6. Law enforcement profile as optimal control on Problem II

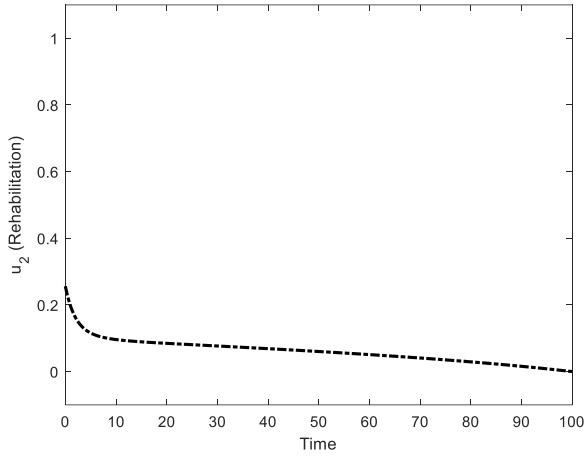


Figure 7. Rehabilitation profile as optimal control in Problem II

Table 1. Comparison of the Number of Non-Criminal Population and Criminal Population at Equilibrium Conditions

Number Population	Without Optimal Control	Optimal Control Problem I	Optimal Control Problem II
Non-Criminal minded	3.799	5.801	5.972
Criminal minded	0.953	0.086	0.012

The comparison of population size over time with and without optimal control is presented in Table 1. The results show that the number of criminal individuals under Optimal Control Problem II is smaller than in Problem I. This indicates that the application of both law enforcement and rehabilitation is more effective in minimizing the criminal population than law enforcement alone. This finding aligns with the statement in the introduction, which emphasizes that restorative measures—in this case, rehabilitation programs—are needed to address criminal behavior, alongside repressive measures, namely law enforcement.

CONCLUSION

Based on the results obtained, it can be concluded that the formulated optimal control problem can minimize the criminal population in the mathematical model of social behavior. This study only incorporated rehabilitation as a factor in addressing criminal acts. Future research may modify the model by considering other factors for intervention or by including additional state variables.

REFERENCE

[1] D. Guntara dan Budiman, "Tinjauan Kriminologi Terhadap Pelaku Tindak Pidana Terorisme di Indonesia dalam Perspektif Teori Differential Association," *Jurnal Justisi Hukum,* vol. 3, no. 1, pp. 106-119, 2018.

[2] F. Ramadhan dan I. Ridwan, "Analisa Kasus Kekerasan Di STPDN Ditinjau Dari Teori Differential Association Edwin H Sutherland Dalam Hukum Pidana Di Indonesia," *ATTAQWA: Jurnal Pendidikan Islam dan Anak Usia Dini,* vol. 1, no. 1, pp. 12-18, 2022.

[3] N. F. A. Idrus dan Y. Widowati, "Cyberbullying di Media Sosial dalam Pres Kriminologis dan Viktmologis," *Diversi Jurnal Hukum,* vol. 8, no. 2, pp. 217-241, 2022.

[4] C. J. Hisyam, A. Miftaqiyah, A. A. H. Putra, C. A. Kinanti, D. N. Ardani, N. P. Lubis dan S. Adila, "Analisis Pelaku Kenakalan Remaja "Klitih" dalam Perspektif Teori Asosiasi Diferensial Sutherland," *Harmoni : Jurnal Ilmu Komunikasi dan Sosial,* vol. 1, no. 4, pp. 81-89, 2023.

[5] C. Timur dan L. Nurhadiyanto, "Tinjauan Penggunaan Rokok Elektrik di Kalangan Remaja dalam Perspektif Teori Differential Association Theory," *Action Research Literate,* vol. 8, no. 8, pp. 2229-2233, 2024.

[6] B. Cahyadi, E. G. Gara, P. Pratama, G. Fitriadi, Arwansa dan D. S. Arian, "Hacker Anak Dalam Perspektif Teori Differential Association: Studi Kasus Peretasan Situs Pengadilan Negeri Kabupaten Konawe," *Ikraith-Humaniora,* vol. 8, no. 1, pp. 329-340, 2024.

[7] C. J. Hisyam, A. Aditya, D. Az-Zahra, D. Galih, F. Sholihah, M. Nafilata dan E. S. Sahla, "Keterasingan Sosial Sebagai Faktor Pemicu Penyalahgunaan Narkoba Di Kalangan Remaja," *Jurnal Ilmiah Research and Development Student,* vol. 3, no. 1, pp. 122-135, 2025.

[8] J. Sooknanan, B. Bhatt dan D. M. G. Comissiong, "A modified predator–prey model for the interaction of police and gangs," *Royal Society Open Science,* vol. 3, no. 160083, pp. 1-15, 2017.

[9] J. O. Akanni, F. O. Akinpelu, S. Olaniyi, A. Oladipo dan A. Ogunsola, "Modelling financial crime population dynamics: optimal control and cost-effectiveness analysis," *International Journal of Dynamics and Control,* vol. 8, p. 531–544, 2019.

[10] W. Nur dan Darmawati, "Mathematical Model of Armed Criminal Group with Pre-emitive and Repressive Intervention," *JOMTA Journal of Mathematics: Theory and Applications,* vol. 2, no. 2, pp. 27-32, 2020.

[11] A. Wahid, S. Toaha dan Kasbawati, "Kontrol Optimal Model Matematika Dinamika Korupsi dengan Pemberian Edukasi dan Kampanye, Perbaikan Sistem, dan Represif," *Proximal,* vol. 6, no. 1, pp. 53-68, 2022.

[12] H. A. Ashi, "Stability analysis of a simple mathematical model for school bullying," *AIMS Mathematics,* vol. 7, no. 4, p. 4936–4945, 2022.

[13] C. Alfiniyah, A. Puspitasari dan Fatmawati, "Mathematical Modelling of Drug Abuse Reduction Strategies taking into account the Treatment Type and Risks Level," *Jambura Journal of Biomathematics,* vol. 4, no. 1, pp. 23-30, 2023.

[14] S. Abbas, J. P. Tripathi dan A. A. Neha, "Dynamical analysis of a model of social behavior: Criminal vs non-criminal population," *Chaos, Solitons, and Fractals,* vol. 98, pp. 121-129, 2017.

[15] A. K. Srivastav, M. Gosh dan P. Chandra, "Modeling dynamics of the spread of crime in a society," *Stochastic Analysis and Applications,* vol. 37, no. 6, pp. 991-1011, 2019.

[16] J. P. Tripathi, S. Bugalia, K. Burdak dan S. Abbas, "Dynamical analysis and effects of law enforcement in a social interaction model," *Physica A,* vol. 567, p. 125725, 2021.

[17] M. Kumar dan S. Abbas, "Modelling and prevention of crime using age-structure and law enforcement," *Journal of Mathematical Analysis and Applications,* vol. 519, no. 2, p. 126849, 2023.

[18] M. Kumar, S. Dhama, F. A. Alqarni dan S. Abbas, "Mathematical analysis and optimal control of age-structured social interaction model with law enforcement," *Mathematical Methods in the Applied Sciences,* vol. 48, no. 3, pp. 3712-3725, 2024.

[19] N. Izzati, J. W. Leksono dan N. Yannuansa, "Dynamical Analysis of Mathematical Model of Social Behavior with Law Enforcement and Religious Approaches," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika,* vol. 17, no. 1, pp. 672-681, 2024.

[20] W. Ningsih, H. Purwaningsih dan R. El Maghfiroh, "Kendali Optimal Model Pertumbuhan Mikroalga dalam Chemostat," *Limits: Journal of Mathematics and Its Applications,* vol. 19, no. 2, pp. 229-240, 2022.

Rejecting Reduction: Clarifying the Concept of Deep Learning in Mathematics Teaching in the Era of Artificial Intelligence

Anis Munfarikhatin(1)*, Irmawaty Natsir (2)

[1,2] Universitas Musamus

Jl. Kamizaun Mopah Lama, Merauke, Papua Selatan

e-mail: munfarikhatin_fkip@unmus.ac.id(1), natsir_fkip@unmus.ac.id(2)

ABSTRAK

Artikel ini bertujuan untuk mengklarifikasi dan meluruskan pemahaman ganda mengenai istilah *deep learning* dalam konteks pendidikan matematika di era kecerdasan buatan. Sering kali, istilah ini direduksi menjadi sekadar teknologi berbasis AI yang mengandalkan internet, padahal dalam ranah pedagogi, deep learning merupakan pendekatan pembelajaran yang menekankan pemahaman konseptual yang mendalam, keterkaitan antargagasan, serta transfer pengetahuan ke situasi baru. Penelitian ini menggunakan pendekatan kajian konseptual berbasis analisis literatur, dengan sumber data sekunder dari artikel jurnal, buku, serta laporan kebijakan terbitan 2000–2024. Hasil kajian menunjukkan bahwa teknologi deep learning memiliki potensi untuk mendukung pembelajaran matematika melalui fitur seperti pengenalan tulisan tangan, sistem evaluasi otomatis, tutor cerdas, dan pembelajaran adaptif. Namun, penerapan teknologi ini juga menghadapi tantangan serius seperti keterbatasan data kontekstual, infrastruktur digital yang tidak merata, sifat model yang sulit diinterpretasi, serta isu etika dan privasi data. Di sisi lain, pendekatan pedagogis deep learning justru menempatkan guru sebagai aktor utama dalam mendesain pembelajaran bermakna. Oleh karena itu, integrasi antara teknologi dan pedagogi harus dilakukan secara kritis dan kontekstual. Inovasi pembelajaran di era AI harus tetap berpijak pada prinsip humanistik dan kesadaran akan realitas sosial-budaya siswa, khususnya dalam konteks Indonesia yang beragam.

Kata kunci: deep learning; pedagogi matematika; kecerdasan buatan; pembelajaran bermakna; pendidikan

ABSTRACT

This article aims to clarify and clarify the dual interpretations of the term' deep learning' in the context of mathematics education in the era of artificial intelligence. The term is often reduced to merely an AI-based technology that relies on the internet. In contrast, in the pedagogical domain, deep learning refers to a learning approach that emphasizes deep conceptual understanding, connections between ideas, and the transfer of knowledge to new situations. This study adopts a conceptual review approach based on literature analysis, using secondary sources such as journal articles, books, and policy reports published between 2000 and 2024. The findings show that deep learning technology holds potential to support mathematics learning through features such as handwriting recognition, automated evaluation systems, intelligent tutoring, and adaptive learning. However, the implementation of this technology also faces serious challenges, including limitations in contextual data availability, uneven digital infrastructure, the opaque nature of model interpretation, and issues of ethics and data privacy. On the other hand, the pedagogical approach to deep learning places the teacher as the main actor in designing meaningful learning experiences. Therefore, the integration of technology and pedagogy must be carried out critically and contextually. Educational innovations in the AI era must remain grounded in humanistic principles and an awareness of students' sociocultural realities—especially in the diverse context of Indonesia.

Keywords: *deep learning; mathematics pedagogy; artificial intelligence; meaningful learning; education*

INTRODUCTION

The development of artificial intelligence (AI) technology over the past decade has driven the emergence of various innovations in the field of education, including in mathematics learning. One of the most widely developed forms of AI implementation is *deep learning*, a machine learning approach that utilizes multi-layered artificial neural networks to recognize patterns, perform classifications, and generate data-based predictions [1][2]. In practice, deep learning has been applied in various technology-based educational systems, such as handwritten mathematical expression recognition [3][4], automated evaluation systems [5][6], and data-driven adaptive learning.

However, it is important to note that the term *deep learning* in the technological domain differs fundamentally from the concept of *deep learning* in pedagogy. In educational studies, *deep learning* refers to a learning process that emphasizes deep conceptual understanding, interconnectedness of ideas, critical thinking skills, and the transfer of knowledge to new contexts [7][8]. This process cannot be achieved merely through the automated delivery of materials or adaptive algorithms; it requires teacher involvement in designing meaningful, contextual, and reflective learning experiences. Therefore, even though AI technology holds potential to support the learning process, it does not automatically ensure the occurrence of pedagogical deep learning.

The pedagogical approach to deep learning increasingly demands the presence of teachers as the primary facilitators of learning. Teachers play a crucial role in fostering mathematical dialogue, interpreting meaning, and adapting instructional strategies to the needs and sociocultural backgrounds of students [9]. In the context of Indonesia—particularly in regions such as Papua, which face geographic, cultural, and infrastructural challenges—meaningful learning approaches must prioritize the teacher's role and the strengthening of classroom interaction over reliance on high-tech systems that may not be contextually appropriate.

Furthermore, positioning AI and deep learning technologies as the main solutions in education often creates an illusion of efficiency that neglects the complexity of human learning processes. While technology can serve as an effective support tool—for instance, through learning analytics or automated feedback—it is dangerous to assume that AI can replace the pedagogical functions of teachers in fostering students' mathematical understanding. Therefore, mathematics educators need to understand and distinguish between deep learning as a technology and deep learning as a pedagogical approach. This distinction ensures that educational innovation strategies remain grounded in humanistic, contextual, and transformative educational principles.

METHOD

This study is a conceptual inquiry aimed at critically and thoroughly clarifying the two main meanings of the term *deep learning* in the context of mathematics education: as a pedagogical approach (deep learning approach) and as a technology in artificial intelligence (deep learning in AI). This review does not employ quantitative or qualitative empirical methods; instead, it relies on a literature analysis and a conceptual reflection approach that is both systematic and argumentative.

Type of Study

This research falls under the category of non-empirical studies based on theoretical analysis, focusing on concept formation, clear distinctions in meaning, and drawing pedagogical implications from existing frameworks of thought. This method is commonly used in the fields of philosophy of education, theory development, and critical discourse in educational studies [10].

Data Sources and Literature

The primary sources in this study are secondary literature consisting of:

1. International journal articles indexed in Scopus and WoS,
2. Academic books on mathematics pedagogy and educational technology,
3. State-of-the-art reviews in the field of AI in education,
4. Policy report documents and white papers from educational institutions.

The literature was purposively collected using keywords such as: *"deep learning approach"*, *"deep learning in education"*, *"mathematics pedagogy"*, *"AI in mathematics teaching"*, and *"teacher role in digital learning"*, drawn from databases such as Scopus, ERIC, Google Scholar, and SpringerLink. The publication years were focused on the period from 2000 to 2024 to capture both historical and contemporary perspectives.

FINDINGS AND DISCUSSION

This review highlights the urgent need to clarify the term *deep learning*, which currently carries two dominant meanings in the practice of mathematics education: as a pedagogical approach and as a technology in artificial intelligence (AI). These two meanings have very different epistemological roots and practical consequences. This literature review aims to correct the reduction of meaning that often occurs in educational policy narratives and the implementation of digital learning.

Deep Learning as AI Technology in Mathematics Learning

In recent literature, the term *deep learning* is more frequently associated with artificial intelligence, particularly with machine learning techniques that use multi-layered artificial neural networks (deep neural networks). In the field of mathematics education, this technology has been widely applied to develop various adaptive and data-driven learning applications.

Recognition of Handwritten Mathematical Expressions

Convolutional Neural Network (CNN) models have been widely used to recognize handwritten mathematical expressions input by students on digital devices. This capability not only automates corrections but also enables real-time identification of students' conceptual errors [3]. Such applications shift the teacher's role from merely checking answers to analyzing learning processes, aligning closely with the principles of formative assessment.

Process-Based Automatic Evaluation

Models such as Recurrent Neural Networks (RNNs) and Transformers can evaluate the sequence of steps students take to solve math problems. These systems assess not just the final answers but also the students' thought processes, making it possible to track their cognitive paths

during learning [11]. This supports the *assessment for learning* principle and enhances the teacher's ability to provide targeted interventions.

Intelligent Tutoring Systems Based on Natural Language

Language-based models like BERT and GPT have been developed into intelligent tutoring systems (ITS) for mathematics education. These systems can answer student questions, offer explanations, and even detect confusion through patterns in natural language interaction [12]. This approach augments the teacher's role in providing personalized support beyond classroom hours.

Adaptive Learning Systems

Deep learning technology is also used in adaptive learning systems, where content and activities are tailored to each student's profile and needs. Models such as autoencoders or multi-layer perceptrons classify learning styles and predict learning difficulties, which then inform automatic instructional decision-making [13].

Challenges in Implementing Deep Learning as a Technology

Despite its promise, the implementation of *deep learning* in the context of mathematics education is not free from significant structural and ethical challenges. One of the main obstacles is the limited availability of contextual and representative data from actual classroom practices. Such data is crucial for training deep learning models to accurately capture the nuances of real learning experiences.

In addition, the implementation of this technology requires adequate digital readiness, including access to high-powered hardware and stable, high-speed internet connectivity, both of which remain challenges in many regions, particularly in developing countries. Moreover, deep learning models are generally "black boxes", meaning their internal operations are difficult to explain or understand, even for teachers. This opacity creates a barrier to pedagogical reflection and instructional decision-making based on a deep understanding of students.

On the other hand, ethical and privacy concerns also demand serious attention. The use of student data without explicit consent, the potential for algorithmic bias, and concerns over excessive surveillance of students' learning activities create moral dilemmas that must be addressed with caution [14][15]. In the context of developing countries like Indonesia—especially in areas with limited access, such as Papua—these challenges become even more complex. Severe limitations in digital infrastructure, combined with diverse learning cultures, call for contextual and adaptive approaches. Therefore, advanced technology alone is not sufficient; pedagogical approaches that are sensitive to local realities must serve as the foundation for any AI-based educational innovation.

In contrast, deep learning in the pedagogical tradition refers to a learning strategy that encourages students to deeply understand concepts, connect various ideas, think critically, and transfer knowledge to new contexts. This concept is grounded in theoretical frameworks such as *approaches to learning* by [7] and the constructivist learning framework developed by [8].

Deep Learning as a Pedagogical Approach

In contrast, within the pedagogical tradition, *deep learning* refers to a learning approach that places conceptual understanding as the central goal of learning, not merely procedural mastery or rote memorization. This concept is rooted in the *approaches to learning* theory introduced by [7], which distinguishes between surface learning and deep learning. Students who engage in surface learning tend to memorize information without truly understanding its meaning, whereas students who learn deeply connect new information to their existing knowledge structures, understand the relationships among concepts, and can apply that understanding to new situations.

This approach is further enriched by constructivist theories in education, as explained by [8], which emphasize that learning is an active process wherein students construct meaning from their own learning experiences. Within this framework, teachers are no longer positioned as the sole source of knowledge, but rather as facilitators who design learning environments that foster exploration, dialogue, and reflection. Learning is not viewed as a linear process, but as a complex social and cognitive activity that demands full student engagement and awareness of the learning context.

A key feature of the pedagogical deep learning approach is the active involvement of students in constructing meaning from new information. This process occurs not only at the cognitive level but also at the affective and metacognitive levels—students learn to understand, reflect, and regulate their learning strategies. Learning takes place in a reflective, contextual, and integrated manner, recognizing that the concepts being studied are interconnected and linked to students' real-life experiences. In this regard, the teacher plays an essential role as both facilitator and mediator of meaning, not merely delivering content, but also creating a dialogic space for building shared understanding. In a study by [16], it was emphasized that creative and spatial thinking should be encouraged through student reflection activities, aligning with the characteristics of pedagogical deep learning.

Assessment in this approach does not focus solely on outcomes, but rather emphasizes the learning process that students undergo [17]. Formative assessments, classroom observations, student reflections, and project-based assessments are the main tools used to determine how well students truly understand concepts and can apply them in authentic contexts [18].

In Indonesia's pluralistic context—geographically, culturally, and socioeconomically—the pedagogical deep learning approach is highly relevant. Challenges such as educational inequality, linguistic and cultural diversity, and the complex needs of students call for approaches that are not standardized, but rather responsive to context. Teachers must understand students' sociocultural realities and use them as starting points in designing instructional strategies. In areas such as Papua, for example, overly tech-dependent approaches that fail to consider infrastructure readiness and local culture risk creating new educational divides. Therefore, strengthening pedagogical approaches that are deep, humanistic, and contextual is key to achieving mathematics education that is not only cognitively sound but also socially just.

Distinction and Synthesis: Rejecting the Reduction of Meaning

This review emphasizes that reducing the term *deep learning* to merely a technological concept is a form of dangerous conceptual reductionism. When AI technology is assumed to be capable of replacing the teacher's role in creating meaningful learning, we overlook the reality that

human learning is social, complex, and contextual. However, this does not mean that the two meanings cannot be integrated. On the contrary, deep learning technology can be used to support pedagogical deep learning, as long as teachers remain the primary agents in designing, interpreting, and facilitating the learning process. In other words, technology must be embedded within a pedagogical framework, not the other way around.

CONCLUSION

The importance of making a clear distinction between *deep learning* as a pedagogical approach and *deep learning* as an AI-based technology. Both have their contributions to the development of mathematics education, but they cannot be equated semantically or practically. The pedagogical approach to deep learning emphasizes active student engagement in constructing meaning, reflecting critically, and transferring knowledge across diverse contexts, with the teacher serving as the central facilitator. Meanwhile, deep learning technology can enhance learning processes through data analysis, automated feedback, and content personalization—as long as its use is guided and supervised by strong pedagogical principles. The danger of reductionism arises when technology is placed at the center of educational processes and replaces the role of teachers, even though learning is inherently a social, complex, and contextual process. Therefore, AI-based educational innovation must prioritize humanistic, ethical, and locally grounded approaches, so that education becomes not only technologically smart but also pedagogically just.

Based on these conclusions, several relevant recommendations can be made. First, educators and policymakers should broaden their understanding of the pedagogical meaning of deep learning, so as not to be trapped in technological euphoria that neglects the human dimension of learning. Second, teacher training is needed to help educators critically understand the role of technology in learning, enabling them to integrate it reflectively and contextually. Third, the development of learning technologies, especially those based on deep learning, must consider algorithmic transparency, student data protection, and the involvement of teachers as active partners. Fourth, further research is urgently needed to explore models of integration between technology and pedagogy that fit the educational context of Indonesia, including in areas with infrastructure challenges and cultural diversity, such as Papua.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[2] Muhammad Athoillah, Rani Kurnia Putri, Andri Kurniawan, Sri Rahmawati Fitriatien, and Dian Majid, *Akselerasi Teknologi Informasi "Cakap dan Beretika di Era Digital,"* Cetakan Pertama., vol. 1. Bandung: PT. Refika Aditama, 2024.

[3] Y. Zhou, L. Jin, and C. Luo, "Handwritten Mathematical Expression Recognition with Contextual Attention Convolutional Sequence-to-Sequence Model," *IEEE Access*, vol. 7, pp. 158271–158280, 2019, doi: 10.1109/ACCESS.2019.2949483.

[4] R. K. Putri and M. Athoillah, "Enhancing handwritten numeric string recognition through incremental support vector machines," *Journal of AppliedMath*, vol. 2, no. 1, p. 373, 2024.

[5] D. R. Costa and C. W. Chen, "Exploring the relationship between process data and contextual variables among Scandinavian students on PISA 2012 mathematics tasks," *Large Scale Assess Educ*, vol. 11, no. 1, 2023, doi: 10.1186/s40536-023-00155-x.

[6] R. K. Putri and M. Athoillah, "Detection of facial mask using deep learning classification algorithm," *Journal of Data Science and Intelligent Systems*, vol. 2, no. 1, pp. 58–63, 2024.

[7] F. Marton and R. Säljö, "On Qualitative Differences in Learning: I—Outcome and Process," *British Journal of Educational Psychology*, vol. 46, no. 1, pp. 4–11, 1976.

[8] J. Biggs and C. Tang, *Teaching for Quality Learning at University*, 3rd ed. McGraw-Hill Education, 2007.

[9] B. Jaworski, *Theory and Practice in Mathematics Teaching Development: Critical Inquiry as a Mode of Learning in Teaching*. Springer, 2006.

[10] K. F. Punch, *Introduction to Social Research: Quantitative and Qualitative Approaches*, 3rd ed. London: SAGE Publications, 2014.

[11] X. Xiong, D. Yang, C. P. Rosé, and Y. Wang, "A Deep Learning Approach for Automatic Feedback Generation in Mathematics Problem Solving," *Journal of Educational Data Mining*, vol. 12, no. 3, pp. 1–26, 2020.

[12] A. Almatrafi, M. Malkawi, and M. Alharthi, "Intelligent Tutoring Systems and Natural Language Processing in Mathematics Education: A Review," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 3, pp. 157–172, 2021.

[13] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.

[14] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators?," *International Journal of Educational Technology in Higher Education*, vol. 16, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.

[15] W. Holmes, M. Bialik, and C. Fadel, *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston, MA: Center for Curriculum Redesign, 2019.

[16] R. Oughton *et al.*, "Developing 'deep mathematical thinking' in geometry with 3- and 4-year-olds: a collaborative study between early years teachers and university-based mathematicians," *Mathematical Thinking and Learning*, vol. 00, no. 00, pp. 1–20, 2022, doi: 10.1080/10986065.2022.2119497.

[17] F. D. L. Anis Munfarikhatin, Rizqi Anugrah, "Game Etnomatematika Kweritop Dalam Pembelajaran Konsep Bangun Datar Peserta Didik Smp," in *Prosiding MAHASENDIKA III Tahun 2024*, Denpasar: Program Studi Pendidikan Matematika, Universitas Mahasaraswati Denpasar, 2023, pp. 79–89.

[18] and E. M. Siti Nurleha, Anis Munfarikhatin, "Development Of Mathematics Learning Media Concept Board (Pakeling Bandar) For Students With Special Needs," *International Journal of Indonesian Education and Teaching*, vol. 8, no. 2, pp. 262–274, 2024.

Classifying Disadvantaged Districts/Cities in Indonesia: A Support Vector Machine Approach

Wanda Surianto(1), Lia Mauliani(2), Ridha Ferdhiana (3), Nurhasanah(4)

[1,2,3,4] Department of Statistics, Universitas Syiah Kuala, Banda Aceh

Jl. Syech Abdurrauf No. 3 Kopelma Darussalam, Banda Aceh, Aceh
e-mail: wanda.surianto@usk.ac.id(1), lia_liani@mhs.usk.ac.id(2), ridha.ferdhiana@usk.ac.id(3), nurhasanah@usk.ac.id(4)

ABSTRAK

Kesenjangan antara wilayah di Indonesia ditandai dengan adanya istilah daerah tertinggal dan daerah tidak tertinggal. Pemerintah menetapkan status daerah tertinggal setiap lima tahun sekali, surat keputusan peraturan presiden no 63 tahun 2020 menetapkan 62 kabupaten/kota di Indonesia sebagai daerah tertinggal. Penelitian ini bertujuan untuk mengklasifikasikan status daerah tertinggal menggunakan algoritma *Support Vector Machine* (SVM) dengan tiga jenis kernel: linear, polinomial, dan *Radial Basis Function* (RBF). SVM dipilih karena kemampuannya dalam menangani data berdimensi tinggi dan tugas klasifikasi non-linear. Dataset yang digunakan berasal dari BPS dan JDIH tahun 2022, mencakup 20 variabel yang merepresentasikan indikator sosial ekonomi, infrastruktur, dan layanan publik. Distribusi data menunjukkan ketidakseimbangan, dengan hanya 62 dari 514 kabupaten/kota yang dikategorikan sebagai daerah tertinggal. Parameter optimal ditentukan secara eksperimental: kernel linear (C = 0,1), polinomial (C = 1, d = 3), dan RBF (C = 1, γ = 0,1). Berdasarkan hasil evaluasi, kernel linear memberikan performa terbaik pada dataset yang digunakan, dengan akurasi 0,94, presisi 0,91, recall 0,81, dan skor F1 sebesar 0,85. Model mengklasifikasikan 45 kabupaten/kota sebagai tertinggal dan 469 sebagai tidak tertinggal. Sebanyak 29 kabupaten/kota menunjukkan perbedaan klasifikasi dibandingkan dengan klasifikasi resmi pemerintah. Perbedaan ini dapat mencerminkan perubahan kondisi di lapangan atau keterbatasan dalam kriteria kebijakan, yang menunjukkan potensi pendekatan berbasis data untuk mendukung perencanaan pembangunan wilayah yang lebih tepat sasaran dan berkeadilan.
Kata kunci: Daerah tertinggal; Klasifikasi; Kernel; Support Vector Machine (SVM)

ABSTRACT

The terms "underdeveloped" and "non-underdeveloped" regions highlight the gap between regions in Indonesia. The government determines the status of underdeveloped regions every five years. Presidential Decree No. 63 of 2020 determines 62 districts/cities in Indonesia as underdeveloped regions. This study aims to classify disadvantaged district status using the Support Vector Machine (SVM) algorithm with three kernel types: linear, polynomial, and Radial Basis Function (RBF). SVM was selected for its effectiveness in handling high-dimensional data and non-linear classification tasks. The dataset, sourced from BPS and JDIH in 2022, comprises 20 variables covering socioeconomic, infrastructure, and public service indicators. The data distribution is imbalanced, with only 62 out of 514 districts labeled as disadvantaged. Optimal parameters were determined experimentally: linear (C = 0.1), polynomial (C = 1, d = 3), and RBF (C = 1, γ = 0.1). Based on evaluation results, the linear kernel achieved the best performance on the given dataset, with an accuracy of 0.94, precision of 0.91, recall of 0.81, and F1-score of 0.85. The model classified 45 districts as disadvantaged and 469 as non-disadvantaged. A total of 29 districts showed discrepancies compared to the official classification. These differences may indicate either changing ground conditions or limitations in policy criteria, highlighting the potential of data-driven approaches to support more targeted and equitable regional development planning.
Keywords: Disadvantaged regions; Classification; Kernel; Support Vector Machine (SVM)

INTRODUCTION

Disadvantaged regions refer to regencies whose areas and communities are underdeveloped compared to other regions on a national scale [1]. The designation of disadvantaged status is determined by the government through a Presidential Regulation (Perpres) issued every five years. The most recent regulation, Perpres No. 63 of 2020, identifies 62 districts/cities as disadvantaged, with the next update scheduled for 2025 [2]. These disparities may be driven by various factors, including limited economic activity, low human resource quality, inadequate infrastructure, high incidence of disasters and conflicts, and restricted access to transportation, telecommunications, and information services [3]. These multidimensional issues necessitate urgent and coordinated responses from both local and central governments to ensure social and economic equity and to prevent regions from falling behind.

One of the government's efforts to address this issue is the issuance of Presidential Regulation No. 63 of 2020 concerning Disadvantaged, Frontier, and Outermost (3T) regions. According to this regulation, 62 districts/cities are classified as disadvantaged regions [2]. Through the Ministry of Villages, Development of Disadvantaged Regions, and Transmigration, several intervention programs have been implemented, improving the status of some districts. However, with ongoing changes in regional administrative structures such as expansion, creation, and mergers, continuous reassessment of regional development status is necessary.

A region is categorized as disadvantaged or not based on a composite index that includes three key components: the Social Resilience Index, the Economic Resilience Index, and the Ecological/Environmental Resilience Index [4]. Additionally, a region's ability to manage resources particularly in terms of potential, information/value, innovation, and entrepreneurship supports its progress toward becoming non-disadvantaged [5].

Various statistical and machine learning methods can be applied to classification analysis, including Support Vector Machine (SVM) [6]. SVM is a technique that separates two classes of data by finding the optimal hyperplane that maximizes the margin between them [7][8]. With kernel functions such as linear, polynomial, and radial basis function (RBF), SVM can handle both linear and non-linear classification problems, particularly in binary classification tasks [9].

A study in Maluku Province applied SVM with a linear kernel (C = 1) and eight variables, achieving 76.31% accuracy in classifying districts/cities [10]. Another study using the Naïve Bayes algorithm on 208 districts/cities yielded 90.5% accuracy with an 80:20 train-test split [11]. However, these methods have limitations Naïve Bayes assumes feature independence, and decision trees may overfit with high-dimensional data. Therefore, this study investigates whether SVM can effectively classify the disadvantaged status of all districts/cities in Indonesia using publicly available indicators. This research is titled *"Classification of Disadvantaged District/City Status in Indonesia Using Support Vector Machine (SVM)"*.

METHOD

Data Source

This study uses secondary data sourced from the 2022 publication of the Central Statistics Agency (BPS) and Presidential Regulation No. 63 of 2020 issued by the Government of the Republic of Indonesia. The regulation is available through the Regulatory Database of the Legal Documentation and Information Network (JDIH BPK). The dataset comprises 514 observations, covering all districts and cities across 34 provinces in Indonesia.

Research Variable

This study utilizes a total of 20 variables, which include one dependent variable (denoted as Y) and nineteen independent variables (denoted as X_1 to X_{19}). The independent variables are categorized into three composite indices: the Social Resilience Index, comprising variables X_1 to X_7; the Economic Resilience Index, comprising variables X_8 to X_{15}; and the Ecological/Environmental Resilience Index, comprising variables X_{16} to X_{19}.

Table 1. Research Variables

Variable	Variable Name	Unit
Y	Disadvantaged Region Status	Classification label (0 = Disadvantaged; 1 = Non-disadvantaged)
X_1	Expected Years of Schooling	Years
X_2	Life Expectancy	Years
X_3	Mean Years of Schooling	Years
X_4	Number of Midwives	Persons
X_5	Number of Community Health Centers	Units
X_6	Number of Family Planning Clinics	Units
X_7	Percentage of Households with Electricity	Percent
X_8	Per Capita Expenditure	Rupiah
X_9	Open Unemployment Rate	-
X_{10}	Percentage of Population Living in Poverty	Percent
X_{11}	Gross Regional Domestic Product	Rupiah
X_{12}	Per Capita Food Expenditure	Rupiah
X_{13}	Poverty Depth Index	-
X_{14}	Poverty Severity Index	-
X_{15}	Number of Markets	Units
X_{16}	Number of Earthquake Events	Events
X_{17}	Number of Landslide Events	Events
X_{18}	Number of Flood Events	Events
X_{19}	Paved Roads	Km

Analysis Methods

Classification is applied when the target variable in a study is categorical. Its primary objective is to predict labels or assign data to specific categories [12]. When the aim is to uncover the predictive structure of a problem, classification focuses on identifying the variables or interactions among variables that contribute to a particular outcome [13].

In machine learning, algorithms often adopt a binary classification framework, which involves only two target classes. For multi-class problems, one class is treated as positive and the others as negative, with the process repeated for each class [14]. Support Vector Machine (SVM) aims to construct a hyperplane that separates data points with positive and negative labels. Rather than using any separating line, SVM seeks the hyperplane that maximizes the margin the shortest distance to the nearest data points, known as support vectors which defines the optimal boundary between classes

[15]. By using kernel functions, SVM is capable of projecting data into a higher-dimensional space to facilitate the separation of classes that cannot be linearly separated in the original space [16].

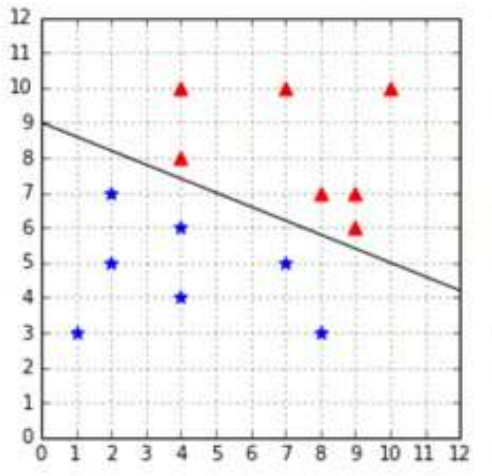


Figure 1. Data partitioning with a hyperplane

Figure 1 illustrates a hyperplane that separates two classes, represented as the positive class (+1) and the negative class (−1). The hyperplane in the above illustration can be expressed by Equation (1), while Equations (2) and (3) indicate that the illustrated method divides the dataset into positive and negative classes..

$$\boldsymbol{wx} + b = 0 \quad (1)$$

$$\boldsymbol{wx} + b > 0 \quad (2)$$

$$\boldsymbol{wx} + b < 0 \quad (3)$$

The weight vector (W) represents the line perpendicular to the origin and the hyperplane. The bias (b) refers to the position of the hyperplane relative to the origin. Equations (4) and (5) below are used to calculate the values of **b** and **W**.

$$b = -1/2 \ (\boldsymbol{w}.\boldsymbol{x}^{+} + \boldsymbol{w}.\boldsymbol{x}^{-}) \quad (4)$$

$$\boldsymbol{w} = \sum_{i=1}^{n} a_i y_i \boldsymbol{x_i} \quad (5)$$

H_1 is the supporting hyperplane for the +1 class, represented by the function $wx_1 + b = +1$, while H_2 is the supporting hyperplane for the −1 class. To determine the optimal hyperplane between the two classes, Equations (7) and (8) is used.

$$Margin = |dH_1 - dH_2| = 2\frac{2}{\|W\|} \quad (6)$$

$$Minimize\ j_1[\boldsymbol{w}] = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad (7)$$

$$yi(xi.\boldsymbol{w} + b) - 1 \geq 0, i = 1, \ldots, n \quad (8)$$

The value of the support vectors influences the level of accuracy in determining the most suitable method the higher the support vector value, the greater the resulting accuracy. Therefore, parameter selection becomes a critical aspect of the problem. A trial and error technique is employed to obtain the optimal accuracy value [17].

Analysis Steps

The data analysis in this study was performed using Microsoft Excel and Python, supported by the Google Colab programming platform. The analysis was conducted through the following stages:

1. Data cleaning;
   The dataset used in this study contains fewer than 100 missing values per variable. To address this issue, the missing values were imputed using the mean value of each respective variable.
2. Descriptive statistical analysis;
3. Data normalization using Z-Score Standardization to standardize the variables;
   Data normalization transforms variable values to a standard scale to prevent dominance caused by scale differences. One commonly used method is Z-score Standardization, which transforms the data so that it has a mean of 0 and a standard deviation of 1 [23]. Z-normalization for a value *x* in a dataset can be calculated using the equation (9).
   $$x_i' = \frac{x_i - \mu}{\sigma} \quad (9)$$
4. Dataset split;
   To obtain a more accurate model with better generalization capability, the training and evaluation processes were carried out using K-Fold Cross Validation. In this method, the dataset is divided into K equal parts (folds), where each fold is used as the testing set in turn, while the remaining folds are used for training. This approach ensures that all data are used for both training and testing, leading to more stable evaluation results that are not dependent on a single train-test split. In this study, data splitting was performed using K = 5 and K = 10 as variations.
5. Inferential analysis using the Support Vector Machine (SVM) method, which involved the application of three kernel functions:
   Linear function: The linear kernel can be expressed in the following equation:
   $$\phi(x_i, x_j) = x_i^T x_j \quad (10)$$
   $K\ (x, x')$ is a kernel function where xxx and $x.x'$ are feature vectors of two data points to be processed. This linear kernel function is the simplest type of kernel, representing the dot product of the two vectors [18][20]. The parameter **C** values used are 0.1, 1, and 10. The selection of these parameters is based on a previous study on the classification of disadvantaged regions [22].
   Polynomial Kernel Function: The polynomial kernel has two parameters: c, which represents the constant term, and d, which denotes the degree of the kernel. The polynomial kernel equation can be written as follows:
   $$\phi(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \quad (11)$$
   $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the feature vectors of two data points to be processed, while d represents the predefined degree of the polynomial. As the degree increases, the kernel function maps the input data into higher-dimensional space, enabling the SVM to perform non-linear classification on more complex datasets. The values of parameter **C** used are 0.1, 1, and 10, while the values of parameter **d** used are 2, 3, and 4. The selection of these parameters follows a previous study on the classification of disadvantaged regions [22].
   Radial basis function (RBF): RBF kernel projects feature vectors into an infinite-dimensional space using the parameter γ (gamma). In the SVM test, treatment using RBF produces a higher

level of accuracy. This is because usually linear and polynomial kernels take less time and provide lower accuracy than rbf or Gaussian kernels [19]. However, unlike the polynomial kernel, the RBF kernel is less dependent on the degree of the function, which makes it more effective in addressing overfitting issues commonly encountered in polynomial kernels [21]. The following equation represents the Radial Basis Function (RBF) kernel.

$$\phi(x_i, x_j) = exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \quad (12)$$

This equation indicates that the closer $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the larger the resulting value. The RBF kernel has the property that higher weights are assigned to data points that are closer together, and these weights decrease exponentially as the distance between the points increases. The values of parameters **C** and **γ** used in the RBF kernel are 0.1, 1, and 10. The selection of these parameters considers the combinations of **C** and **γ** values used in previous studies [22].

6. Result interpretation and conclusion

   In classification tasks, model performance is commonly evaluated using several metrics, including the confusion matrix, accuracy, precision, and recall.

RESULT AND DISCUSSION

Descriptive Analysis

Descriptive analysis provides a general overview of the data. Indonesia consists of 514 districts/cities across 34 provinces. According to Presidential Regulation No. 63 of 2020 on 3T regions, 62 districts/cities (12%) are classified as disadvantaged, while the remaining 452 (88%) are non-disadvantaged. This shows that the majority of regions fall into the non-disadvantaged category. The distribution highlights regional disparities and can guide policymakers in formulating development strategies. The descriptive statistics are presented in Table 2.

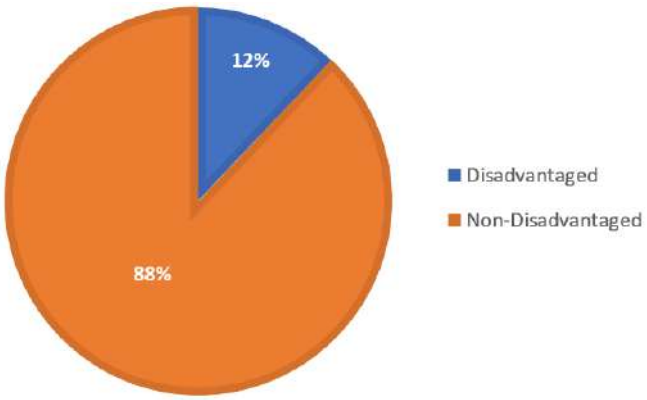


Figure 2. The disadvantaged region status of districts/cities in Indonesia.

The province with the highest number of disadvantaged districts/cities is Papua, with a total of 22, followed by East Nusa Tenggara, which has 13 disadvantaged districts/cities. Furthermore, there are three provinces Papua, Maluku, and East Nusa Tenggara where the number of disadvantaged regions exceeds the number of non-disadvantaged ones. Java and Kalimantan Island have no disadvantaged districts/cities, while Sumatra Island has only a small number. Overall, it can be observed that disadvantaged districts and cities are predominantly concentrated in the eastern region of Indonesia.

Table2. Descriptive Statistics of Study

Variable	Minimum	Maximum	Mean
X_1	4.07	17.81	13.09
X_2	55.70	77.82	69.93
X_3	1.58	13.03	8.55
X_4	0.00	10281.00	641.33
X_5	3.00	101.00	22.26
X_6	7.00	1060.00	76.00
X_7	27.70	100.00	97.54
X_8	4190.00	24221.00	10643.32
X_9	0.12	75.64	5.96
X_{10}	2.28	42.03	11.68
X_{11}	255.92	72961651.00	591218.80
X_{12}	43.84	5740045.00	601974.40
X_{13}	0.17	13.90	1.97
X_{14}	0.02	7.59	0.53
X_{15}	0.00	137.00	19.35
X_{16}	0.00	11.00	0.12
X_{17}	0.00	103.00	1.68
X_{18}	0.00	30.00	3.16
X_{19}	0.00	105589.00	477.30

According to BPS, the national average for expected years of schooling in Indonesia is 12.85 years, which closely aligns with the dataset average of 13.09 years. However, there is a wide disparity across regions, with values ranging from as low as 4.07 years to as high as 17.81 years. This indicates that while some districts still experience very limited access to education, others have expectations reaching higher education levels. Such gaps reflect serious inequalities in educational opportunities.

Similar disparities are observed in other indicators within the Social Resilience Index. The number of midwives ranges from 0 to 10,281, and community health centers from 1 to 1,010 units highlighting unequal access to basic health services. In contrast, life expectancy shows relatively low variability, ranging from 55.70 to 77.82 years, with an average of 69.93 years. This suggests that despite differences in health infrastructure, longevity remains fairly consistent. Overall, these descriptive statistics emphasize the need for targeted classification and intervention to strengthen social resilience in underdeveloped regions.

Selection of the Best Ratio and Parameters

Following the evaluation of various parameters across two different ratios for each kernel function, the subsequent step involves selecting the parameters that demonstrate the best performance based on the classification test results. The optimal parameters are determined using classification performance indicators, namely accuracy, precision, recall, and f1 score.

Table 3. Best Parameters for Linear Kernel Function

K-folds	Parameters	Classification Performance			
		accuracy	**precision**	**recall**	**f1 score**
10-folds	c = 0.1	0,94231	0,97	0,85	0,89
	c = 1	0,92308	0,88	0,88	0,88
	c = 10	0,94231	0,90	0,93	0,91
5-folds	c = 0.1	0,96117	0,98	0,88	0,92
	c = 1	0,95146	0,92	0,89	0,91
	c = 10	0,93204	0,88	0,86	0,87

Based on Table 3, it can be seen that the best parameter for classifying districts/cities in Indonesia using the linear kernel is C = 10 with 10-folds, while for the 5-folds, the best C parameter for classifying the status of disadvantaged districts/cities is 0.1.

Table 4. Best Parameters for Polynomial Kernel Function

K-folds	Parameters	Classification Performance			
		accuracy	**precision**	**recall**	**f1 score**
10-folds	d = 2; c = 0.1	0,86538	0,84	0,69	0,73
	d = 2; c = 1	0,88462	0,86	0,74	0,78
	d = 2; c = 10	0,88462	0,83	0,78	0,80
	d = 3; c = 0.1	0,88462	0,94	0,70	0,75
	d = 3; c = 1	0,92308	0,96	0,80	0,85
	d = 3; c = 10	0,88462	0,86	0,74	0,78
	d = 4; c = 0.1	0,86538	0,84	0,69	0,73
	d = 4; c = 1	0,88462	0,86	0,74	0,78
	d = 4; c = 10	0,86538	0,80	0,73	0,75
5-folds	d = 2; c = 0.1	0,89320	0,88	0,68	0,73
	d = 2; c = 1	0,91262	0,95	0,72	0,78
	d = 2; c = 10	0,92233	0,88	0,80	0,83
	d = 3; c = 0.1	0,90291	0,95	0,69	0,75
	d = 3; c = 1	0,92233	0,96	0,75	0,81
	d = 3; c = 10	0,92233	0,91	0,78	0,82
	d = 4; c = 0.1	0,89320	0,88	0,68	0,73
	d = 4; c = 1	0,90291	0,89	0,71	0,76
	d = 4; c = 10	0,92233	0,88	0,80	0,83

Based on Table 4, the best parameter combination for the polynomial kernel function in the simulation test is d = 3 and C = 1. By comparing the classification model performance indicators, the same optimal parameter combination is obtained for both 10-folds and 5-folds training and testing data ratios.

Table 5. Best Parameters for RBF Kernel Function

K-folds	Parameters	Classification Performance			
		accuracy	precision	recall	f1 score
10-folds	γ = 0.1; c = 0.1	0,80769	0,40	0,50	0,45
	γ = 0.1; c = 1	0,84615	0,80	0,64	0,67
	γ = 0.1; c = 10	0,86538	0,84	0,69	0,73
	γ = 1; c = 0.1	0,80769	0,40	0,50	0,45
	γ = 1; c = 1	0,80769	0,48	0,50	0,45
	γ = 1; c = 10	0,80769	0,40	0,50	0,45
	γ = 10; c = 0.1	0,80769	0,40	0,50	0,45
	γ = 10; c = 1	0,80769	0,40	0,50	0,45
	γ = 10; c = 10	0,80769	0,40	0,50	0,45
5-folds	γ = 0.1; c = 0.1	0,84466	0,42	0,50	0,46
	γ = 0.1; c = 1	0,90291	0,95	0,69	0,75
	γ = 0.1; c = 10	0,88350	0,82	0,68	0,72
	γ = 1; c = 0.1	0,84466	0,42	0,50	0,46
	γ = 1; c = 1	0,84466	0,42	0,50	0,46
	γ = 1; c = 10	0,84466	0,42	0,50	0,46
	γ = 10; c = 0.1	0,84466	0,42	0,50	0,46
	γ = 10; c = 1	0,84466	0,42	0,50	0,46
	γ = 10; c = 10	0,84466	0,42	0,50	0,46

Based on Table 5, the optimal parameters for the RBF kernel were $\gamma = 0.1$ and $C = 10$ for both 10-fold and 5-fold dataset splits. However, the best overall performance was achieved using the linear kernel with $C = 0.1$ on the 5-fold dataset, resulting in an accuracy of 0.96117, precision of 0.98, recall of 0.88, and F1 score of 0.92.

Table 6. Best Model Confusion Matrix (Linear Kernel, c = 0.1)

Actual	Prediction		Total
	disadvantaged	non-disadvantaged	
Disadvantaged	39	23	62
Non-disadvantaged	6	446	452
Total	45	469	514

After evaluating all dataset split configurations, the linear kernel with $C = 0.1$ was identified as the most effective for classifying disadvantaged districts/cities in Indonesia using the SVM algorithm. The best-performing model was the linear kernel, as determined by the SVM-based analysis conducted in this study [24]. Based on Table 6, the model correctly classified 39 districts/cities as disadvantaged and 446 as non-disadvantaged, while 23 disadvantaged areas were misclassified as non-disadvantaged and 6 non-disadvantaged areas were misclassified as disadvantaged. Using this optimal setting, the model achieved an accuracy of 0.94, precision of 0.91, recall of 0.81, and an F1 score of 0.85.

CONCLUSION

This study identified the optimal kernel functions and parameter configurations for classifying the disadvantaged status of districts and cities in Indonesia using the Support Vector Machine (SVM) algorithm. The best-performing settings were c = 0.1 for the linear kernel, demonstrated the highest performance, achieving an accuracy of 94%, precision of 91%, recall of 81%, and an F1 score of 85%. When compared to the official 2020 classification outlined in Presidential Regulation No. 63, the SVM model predicted a reduction in the number of disadvantaged regions from 62 to 45 in 2022, with 29 districts/cities showing a change in status 23 improved to non-disadvantaged, while 6 previously non-disadvantaged areas were reclassified as disadvantaged. These findings underscore the potential of SVM as a valuable tool to support policy decisions and monitor regional development progress. However, the study is not without limitations, as it relies solely on cross-sectional data from a single year (2022), without incorporating temporal dynamics. To address this, future research should consider using longitudinal data, techniques to improve model interpretability, and applying ensemble methods to enhance robustness. Furthermore, involving policy stakeholders through feedback mechanisms could increase the model's practical relevance and effectiveness in guiding targeted development interventions and early identification of at-risk regions.

REFERENCE

[1] Pemerintah Indonesia. Peraturan Presiden Republik Indonesia No 131 Tahun 2015 Tentang Penetapan Daerah Tertinggal Tahun 2015-2019. Sekretariat Negara. Jakarta. 2015.

[2] Pemerintah Indonesia. Peraturan Presiden Republik Indonesia Nomor 63 Tahun 2020 tentang Penetapan Daerah Tertinggal Tahun 2020–2024. Jakarta: Sekretariat Negara Republik Indonesia. 2020.

[3] E. A. Sari, Meilani T, I. A Shariati, S. Sofyan, R. A. Baihaqi, R. Nooraeni. Klasifikasi Kabupaten Tertinggal Di Kawasan Timur Indonesia Dengan Support Vector Machine. JIKO (Jurnal Informatika dan Komputer). Vol. 3, No. 3, pp 188-195. 2020.

[4] Direktorat Jenderal Pembangunan Desa dan Perdesaan. Indeks Desa Membangun. Jakarta. 2020.

[5] Hamidi, H., Harioso, & Huda. Indeks Desa Membangun. Kementrian Desa, Pembangunan Daerah Tertinggal dan Transmigrasi, Jakarta. 2015.

[6] R. Primartha, Belajar Machine Learning Teori dan Praktik. Bandung : Informatika Bandung, 2018

[7] Mariyam, P. Ana, Ratnawati, D. Eka and Wahyu, Widodo Agis. Klasifikasi Penyakit Gigi dan Mulut Menggunakan Metode Support Vector Machine. Pengembangan Teknologi Informsi dan Ilmu Komputer. Vol. 2, pp. 802-810. 2018.

[8] W. C. Hsu, C. C. Chang and C. J. Lin. A Practical Guide to Support Vector Machine., Taipei: Departement of Computer Science National Taiwan University. 2014.

[9] Domingos, P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. Basic Book, New York. 2015.

[10] Palisoa, N. F., Sinay, L. J., & Matdoan, M. Y. Penerapan Support Vector Machine (SVM) untuk Klasifikasi Kabupaten Tertinggal di Provinsi Maluku. Jurnal Matematika, Statistika dan Terapannya, vol. 02, pp. 79–86. 2023.

[11] Lidya, W., Yozza, H., & Yanuar, F. Klasifikasi Daerah Tertinggal Di Indonesia Menggunakan Metode Naive Bayes Classifier Yanuar. Jurnal Matematika UNAND, IX (1), pp 23–29. 2020.

[12] Wizner, W. Python programming for beginners: 3 books in 1. Springer, London. 2020.
[13] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. Classification and Regression Trees. Chapman & Hall, London. 2020.
[14] T. Jo, Machine learning foundations: Supervised, unsupervised, and advanced learning. Korea: Springer International Publishing. 2021.
[15] C. C. Aggarwal, Data Classification Algorithms and Applications. New York: CRC Press, 2015.
[16] Yalsavar, M., Karimaghaee, P., Sheikh-Akbari, A., Khooban, M.-H., Dehmeshki, J., Al- Majeed, S. Kernel parameter optimization for support vector machine based on sliding mode control. IEEE Access 10, 17003–17017. 2022.
[17] J. A. K. Suykens, M. Signoretto, and A. Argyriou. Regularization, Optimization, Kernels, and Support Vector Machines. 2014.
[18] I Nyoman Setiawan, Robert Kurniawan, Budi Yuniarto, Rezzy Eko Caraka, Bens Pardamean, Parameter Optimization of Support Vector Regression Using Harris Hawks Optimization, Procedia Computer Science, vol. 179, pp. 17-24. 2021.
[19] Wiryawanto T. M. P., Hawani Z., and Ramadhani M. A. Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME", *J Statistika*, vol. 16, no. 1, pp. 384–394, Jul. 2023
[20] M. Athoillah, E. Purnaningrum, and R. K. Putri, "Modified Multi-Kernel Support Vector Machine for Mask Detection," CommIT (Communication and Information Technology) Journal, vol. 16, no. 2, pp. 159–166, 2022.
[21] Cambell, C., & Ying, Y. Learning with Support Vector Machines : Synthesis Lecturers on Artificial Intelligence and Machine Learning. Morgan & Claypool. 2011.
[22] Al-Azies, H., & Anuraga, G. Klasifikasi Daerah Tertinggal di Indonesia Menggunakan Algoritma SVM dan k-NN. Jurnal Ilmu Dasar, 22(1), 31–38. 2021.
[23] James, G., Witten, D., Hastie, T., & Tibshirani, R. An Introduction to Statistical Learning. Springer, London. 2017.
[24] Roshafara F. Forecasting Average Rice Prices at Milling Level According to Quality Using Support Vector Regression ", J Statistika, vol. 17, no. 1, pp. 664–671, Jul. 2024.

Logistic Regression for Sentiment Analysis of Insecurity Phenomena on Platform X

Emeylia Safitri [(1)], Wara Alfa Syukrilla [(2)], Ika Nur Laily Fitriana [(3)]

[1,3]Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka

[2]Psychology Study Program, The Faculty of Psychology, UIN Syarif Hidayatullah Jakarta

Jl. Cabe Raya, Pondok Cabe, Pamulang, Tangerang Selatan 15437

e-mail: emeylia.safitri@ecampus.ut.ac.id[(1)], wara.alfa@uinjkt.ac.id[(2)], ika.nur@ecampus.ut.ac.id[(3)]

ABSTRAK

Fenomena insecurity sebagai gejala psikologis semakin sering menjadi bahan diskusi di media sosial. Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap fenomena insecurity yang diekspresikan melalui unggahan di Platform X. Penggunaan analisis sentimen dalam fenomena insecurity menjadi penting karena fenomena tersebut bersifat subjektif dan seringkali tidak terdeteksi dalam kehidupan nyata. Dalam konteks ini, analisis sentimen menjadi alat yang efektif untuk menggali sentimen pengguna secara sistematis dan objektif. Data dikumpulkan dari tweet berbahasa Indonesia pada Januari 2025 dengan kata kunci terkait seperti "insecure", "minder", dan "overthinking". Setelah melalui proses text preprocessing, data diklasifikasikan menjadi tiga kategori sentimen: positif, netral, dan negatif. Model regresi logistik digunakan sebagai metode klasifikasi dengan validasi silang 10-lipat untuk mengevaluasi kinerja model. Hasil penelitian menunjukkan dominasi sentimen negatif sebesar 73.34%, sedangkan sentimen positif dan netral masing-masing sebesar 20.38% dan 6.28%. Rata-rata akurasi model mencapai 83.13% dengan performa terbaik dalam mendeteksi sentimen negatif. Visualisasi wordcloud memperlihatkan dominasi kata-kata dengan nuansa negatif, seperti "takut", "rendah", dan "sendiri". Temuan ini menyoroti pentingnya pemahaman lebih lanjut terhadap dinamika psikologis masyarakat digital. Penelitian ini juga membuka ruang bagi intervensi berbasis data dalam mendukung literasi kesehatan mental di ranah daring.

Kata kunci: Insecurity, Regresi Logistik, Analisis Sentimen, Platform X, Text Mining

ABSTRACT

The phenomenon of insecurity as a psychological symptom is increasingly becoming a topic of discussion on social media. This study aims to analyze public sentiment toward the phenomenon of insecurity as expressed through posts on Platform X. The use of sentiment analysis in the context of insecurity is crucial because the phenomenon is subjective and often undetectable in real life. In this context, sentiment analysis is an effective tool for systematically and objectively exploring user sentiment. Data was collected from Indonesian-language tweets in January 2025 using related keywords such as "insecure","minder", and "overthinking." After undergoing text preprocessing, the data was classified into three sentiment categories: positive, neutral, and negative. Logistic regression was employed as the classification method, with 10-fold cross-validation used to evaluate model performance. The study's results show a dominance of negative sentiment at 73.34%, with positive and neutral sentiments accounting for 20.38% and 6.28%, respectively. The model's average accuracy reached 83.13%, with the best performance in detecting negative sentiment. Wordcloud visualizations revealed a dominance of negatively nuanced words such as "takut","rendah" and "sendiri." These findings underscore the importance of deeper understanding of the psychological dynamics of the digital public. This study also paves the way for data-driven interventions to support mental health literacy in online spaces.

Keywords: *Insecurity; Logistic Regression; Sentiment Analysis; Platform X; Text Mining*

INTRODUCTION

The development of information and communication technology has brought major transformations in the way individuals interact and express their opinions. Social media has become the primary channel for voicing perspectives on various social, political, and psychological issues. Among the available platforms, Platform X (previously known as Twitter) stands out as one of the most dynamic media with high levels of public participation [1]. In January 2024, Indonesia had approximately 27.5 million X users, making it the fourth largest user base in the world after the United States, Japan, and India. Its widespread use in Indonesia makes the platform a representative reflection of digital society expression.

Platform X has become a representative space for digital expression in Indonesia, whether in political [2], educational [3], or socio-economic [4] issues. One psychological phenomenon frequently appearing in online discourse is insecurity — an emotional condition involving feelings of insecurity, low self-esteem, or lack of confidence. This condition can cause significant stress and have a negative impact on an individual's mental well-being and performance in the workplace [5]. This phenomenon not only reflects individual dynamics but also forms part of the collective narrative shaped in online social interactions. Intensive use of social media among adolescents, especially females, has been identified as a major trigger of feelings of insecurity, resulting from social comparisons to standards of perfection displayed on social media. In one study, 83.5% of young women aged 18–21 experienced insecurity due to social media, particularly Instagram, and they demonstrated strong dependency on such platforms in their daily lives [1].

In the fields of statistics and data science, social media serves as an unstructured data source that poses analytical challenges. Various approaches such as text mining, sentiment analysis, and statistical modeling have been developed to extract meaningful insights from this social data. In this context, text mining becomes the key to extracting useful knowledge from this unstructured text data [6]. Logistic regression model based on Twitter data could achieve up to 89.83% accuracy in classifying public sentiment regarding face-to-face learning [3]. Furthermore, logistic regression and found that location and business activities significantly affect MSME income in Surabaya [4]. The logistic regression method was chosen because it is effective in classifying sentiments into discrete categories such as positive, negative, and neutral, and easy to interpret. This method aligns well with data represented numerically, such as TF-IDF, and is suitable for relatively simple text datasets. Besides being computationally efficient, logistic regression also allows for analysis of the influence of each word feature on the classification results. This supports the research objective of understanding sentiment patterns toward insecurity phenomena in a measurable and in-depth manner.

Logistic regression has also been applied to socio-economic studies [7], found that individuals who were not household heads (OR = 8.14), unmarried (OR = 4.48), and had lower education (OR = 4.47) were at higher risk of unemployment in West Java. Additionally, use of Platform X significantly increased online (R^2 = 12.7%) and offline (R^2 = 6.7%) political participation, with a strong correlation between the two (r = 0.698) [2]. In previous research, sentiment analysis was used to examine public sentiment towards data security issues (close to psychological insecurity towards information) using machine learning with data sources from platform X [8]. Similar research regarding sentiment analysis towards insecurity phenomena has not been widely found.

However, specific studies exploring the insecurity phenomenon on Indonesian social media using text-based sentiment analysis approaches remain limited. Text mining is not only a tool for understanding what people are talking about, but also a key to designing more effective strategies to respond to the dynamics that occur in cyberspace [8]. This study aims to apply sentiment analysis to posts about insecurity on Platform X using logistic regression. Additionally, this research evaluates classification model performance via cross-validation and confusion matrix analysis, while illustrating Indonesian public sentiment patterns related to insecurity. It is expected that the results will contribute scientifically to digital literacy development, increased mental health awareness, and data-based social policy formulation in Indonesia.

METHOD

This study adopts a quantitative content analysis approach based on text mining techniques to explore and analyze public sentiment regarding the phenomenon of insecurity — feelings of uncertainty — within various social, political, and economic contexts. Text mining is a data science method used to extract valuable information from collections of unstructured text data [10]. Text mining is not only a tool for understanding what people are talking about, but also a key to designing more effective strategies to respond to the dynamics that occur in social media [9] This technique enables the identification of patterns, trends, and hidden opinions within large volumes of text data, making it highly relevant for studying public perception dynamics on social media. This study used logistic regression to model the relationship between a categorical dependent variable (multinominal) and one or more independent variables, which can be continuous or categorical. Unlike linear regression, which predicts numerical values, logistic regression predicts the probability of an event or a specific category using the following formula:

$$(Y = 1 \mid X) = \frac{1}{(1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)})} \quad (1)$$

Where, P(Y=1|X) is probability of the event occurring (Y = 1) given the predictor variables X, β_0 is intercept (log-odds when all predictors are 0), β_i is regression coefficients representing the influence of each predictor, and X_i is predictor variables.

The stages of data analysis in this study are as follows:

1. Data Collection
   The data collection process began with the retrieval of tweets related to the topic of insecurity using keywords such as “insecure”, “minder” “insecurity” and “overthinking.” The collection focused on Indonesian-language tweets during January 2025 and filtered for geolocation within Indonesia to ensure cultural relevance. A total of 5,226 tweets were collected for analysis.
2. Text Processing
   Raw data was cleaned from irrelevant elements such as symbols, usernames, URLs, punctuation, and numbers. This included converting text to lowercase (case folding), tokenization into individual words, and stopword removal to retain only meaningful terms. These steps aimed to simplify the text and improve the quality of feature representation.
3. Sentiment Labeling
   After cleaning, the tweets were manually labeled based on their semantic content. Each tweet was classified into one of three sentiment categories—positive, negative, or neutral—

considering the explicit and implicit context. Previous research also carried out manual labeling by researchers to obtain accurate results [11].

4. Wordcloud Visualization
   To obtain an initial view of frequently occurring words, a wordcloud visualization was generated. This allowed identification of dominant keywords related to the insecurity phenomenon and supported the interpretation of sentiment.
5. Text Transformation into Numerical Representation
   Normalized text was then transformed into numerical form using the Term Frequency–Inverse Document Frequency (TF-IDF) method [12]. This representation enables statistical algorithms to recognize the importance of each word across the corpus. The dataset was split into training and testing sets for model training and performance evaluation.
6. Modeling and Performance Evaluation
   Logistic regression was used as the classification algorithm to map tweets into sentiment categories based on the generated feature representations. Model evaluation used k-fold cross-validation (k=10) to avoid overfitting and obtain stable performance estimates. One of the most widely used methods to assess the performance of a classification model is the confusion matrix or classification table [9]. Evaluation metrics included accuracy, precision, recall, and F1-score, calculated from the confusion matrix of classification results.

RESULT AND DISCUSSION

The dataset analyzed consists of 5,226 tweets collected from Platform X during January 2025. Tweets were classified into three sentiment categories—positive, neutral, and negative—based on semantic analysis and manual labeling. The classification results showed that 1,126 tweets (20.38%) were positive, 347 tweets (6.28%) were neutral, and 4,053 tweets (73.34%) were negative. This distribution is visualized in Figure 1.

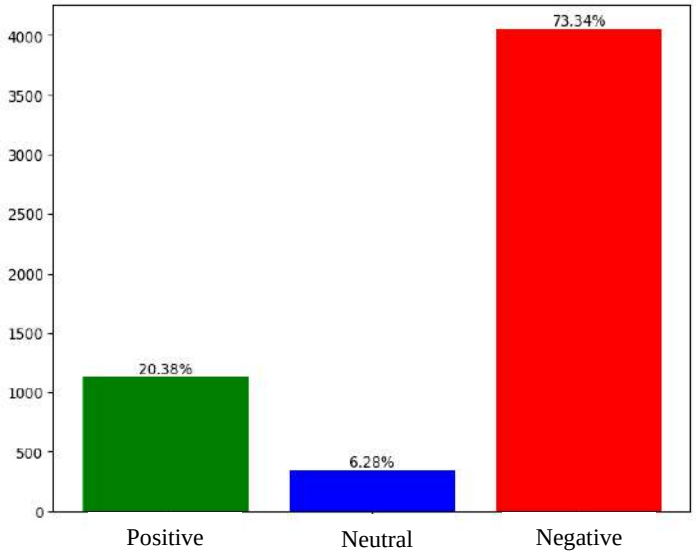


Figure 1. Sentiment visualization

The wordcloud in Figure 2 shows that words such as “merasa” and “diri” appear most prominently, indicating high frequency in the tweet corpus. This dominance suggests that feelings of insecurity are closely linked to individual perceptions of the self. The prominence of the word “percaya” also reflects a need for self-validation and efforts to rebuild confidence. The terms “overthinking” and “minder” highlight excessive rumination and low self-esteem as central themes in discussions of insecurity. Words such as “tidak” “takut” “sendiri” and “rendah” further reinforce the negative emotional spectrum coloring this discourse. On the other hand, words such as “benar” “percaya” “semangat” and “mampu” indicate the presence of affirming narratives and optimistic encouragement in overcoming insecurity. This reflects a dialectic between vulnerability and resilience within public discourse on social media. Overall, the wordcloud visualization provides strong insights into the psychological structure of users expressing insecurity, revealing a complex narrative of uncertainty, self-discovery, and intense emotional dynamics within digital society.

Figure 2. Word Cloud of Dominant Words in Tweets Based on Sentiment Category

Before classification modeling is performed, text data needs to be cleaned and prepared so that it can be represented in a numeric form that matches the input format of the machine learning algorithm. This preprocessing process aims to remove irrelevant elements and standardize the text to be optimal for further analysis. The initial step includes removing usernames (the “@” character), URLs (links with the prefix “http” or “www”), and hashtags (the “#” character and the word after it). These elements generally have no informative value in the context of sentiment analysis, and have the potential to be noise in modeling. In addition, emoticons, non-alphabetic symbols, punctuation, and numbers are also removed to maintain text consistency.

After preprocessing, the text is converted into numeric form using CountVectorizer, which produces a document-term matrix (DTM). This DTM is the basis for classification modeling using logistic regression, where each tweet is represented as a numeric vector based on the frequency of word occurrence. To evaluate model performance objectively and avoid overfitting, the k-fold cross-validation technique is used with a value of k = 10. This method divides the data into 10 random subsets, and in each iteration, one subset is used as test data while the remaining nine are used as training data, the evaluation results of which can be seen in Figure 3. Based on Figure 3, the accuracy scores of each fold are in a relatively narrow and consistent range, without extreme fluctuations. The average cross-validation accuracy was recorded at 83.13%, as indicated by the

red horizontal line on the graph. This value reflects that the logistic regression model has a good and stable classification performance in predicting sentiment based on text features.

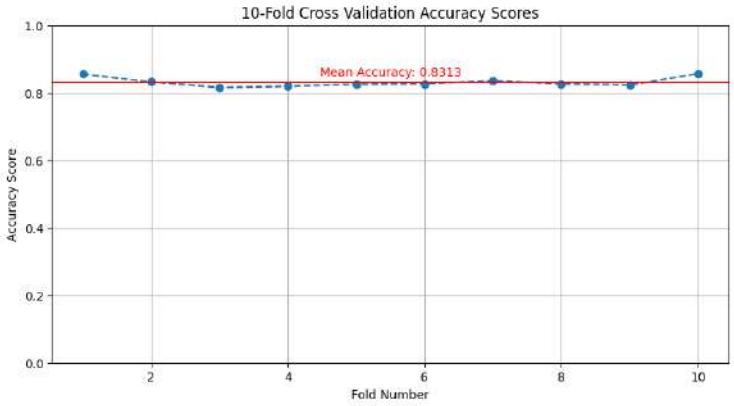


Figure 3. Accuracy Score on 10-Fold Cross Validation Using Logistic Regression Model

Based on Figure 4, the results of the model performance evaluation using the confusion matrix show that the logistic regression model has very good classification performance in detecting negative sentiment. From a total of 816 actual data with negative labels, the model successfully classified 800 data correctly, so that the accuracy for the negative class reached 98%. This shows that the model has a high ability to detect negative opinions on social media. However, the model's performance for the neutral and positive classes was much lower. Only 39% of neutral data (28 out of 71) and 45% of positive data (99 out of 219) were successfully classified correctly. Most of the misclassifications occurred in neutral and positive data that were misclassified as negative, as many as 38 and 115 cases, respectively.

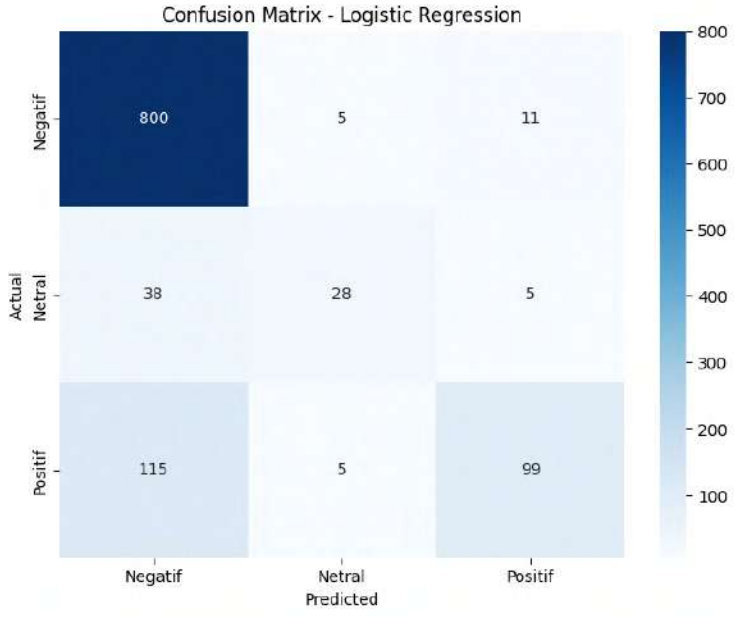


Figure 4. Confusion Matrix

Table 1 Classification Report

Class	Precision	Recall	F1-Score	Support
Negative	0.84	0.98	0.9	816
Neutral	0.74	0.39	0.51	71
Positive	0.86	0.45	0.59	219
Accuracy			**0.84**	**1106**

Confussion matrix in Figure 4 and classsification report in Tabel 1 are the results of modeling 80% of the data, then 20% becomes the test data. Meanwhile, Based on Table 1, it can be seen that the testing of the logistic regression model on the test data produces a confusion matrix that describes the distribution of predictions for three sentiment classes, namely Negative, Neutral, and Positive. The evaluation results show that the model has the best performance in detecting the Negative class. Of the total 816 actual data in the Negative class, the model successfully classified 800 data correctly, resulting in a recall value of 98.04%. This shows that the model has a very high sensitivity to Negative sentiment. In contrast, the model's performance on the Neutral class is moderate. Out of 71 actual Neutral data, only 28 data were predicted correctly, while 38 other data were classified as Negative, and 5 other data were classified as Positive. For the Positive class, the model shows a fairly good level of accuracy, although it is not yet optimal.

CONCLUSION

This study successfully revealed public perceptions of the insecurity phenomenon on Platform X using logistic regression-based sentiment analysis. The main finding shows a strong dominance of negative sentiment (73.34%), significantly higher than positive (20.38%) and neutral (6.28%) sentiments. This indicates that insecurity is generally perceived as a negative psychological experience within Indonesia's digital society. The wordcloud visualization reinforced this by highlighting dominant words such as “afraid,” “inferior,” and “overthinking,” which reflect users’ emotional vulnerability.

From a modeling perspective, logistic regression demonstrated good classification performance, with an average accuracy of 83.13%. However, confusion matrix evaluation revealed limitations in detecting minority classes. Low recall values for neutral (0.39) and positive (0.45) classes indicate the presence of majority class bias, a common challenge in imbalanced datasets. Misclassification of 115 positive and 38 neutral tweets as negative also reduced the model’s sensitivity, especially in early detection applications that require high precision. Academically, this study contributes to the literature on text mining applications in mental health issues in Indonesia and highlights technical challenges in sentiment classification using social media data.

Future studies are recommended to address class imbalance using techniques such as oversampling, under sampling, or class weighting to improve performance in minority classes [14]. More advanced text representations like word embeddings (e.g., Word2Vec or Fast Text), or deep learning approaches based on transformer models (such as BERT), are also suggested to enhance contextual understanding. Research can also be expanded by increasing the dataset size, extending the data collection period, or focusing on specific regions with higher insecurity prevalence. Practically, these findings may benefit mental health practitioners, policymakers, and social organizations in designing more targeted and data-driven intervention strategies.

ACKNOWLEDGMENTS

The author would like to thank the Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka and Psychology Study Program, The Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, as well as all parties involved in supporting this research and publication.

REFERENCE

[1] A. Valentina, G. L. Putri, V. Valiani, and O. H. Putri, "Komunikasi Visual untuk Edukasi Insecurity pada Remaja Perempuan yang Diakibatkan oleh Penggunaan Media Sosial,", *bahasarupa*, vol. 5, no. 2, pp. 237–245, 2022.

[2] D. D. Anwar, K. A. Wibowo and H. S. M. Rohayati, "Pengaruh Penggunaan Media Sosial X (Twitter) Terhadap Partisipasi Politik pada Pemilu Presiden Indonesia Tahun 2024," *Jurnal Riset Komunikasi*, vol. 15, no. 2, pp. 314–328, 2024.

[3] S. A. R. Manaf, A. Alamudi, and A. Fitrianto, "Sentiment Analysis of Twitter Users' Opinion Towards Face-to-Face Learning: Analisis Sentimen Tanggapan Masyarakat Pengguna Twitter terhadap Pembelajaran Tatap Muka,", *Indonesian Journal of Statistics and Its Applications*, vol. 7, no. 1, pp. 15–31, 2023.

[4] A. N. Alifah, A. I. Edina, and M. Almuhayar, "Ordinal Logistic Regression Model of Micro, Small, and Medium-Sized Enterprises Income: A Case Study of Micro, Small and Medium-Sized Enterprises in Surabaya,", *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 2, pp. 143–154, 2024.

[5] C. Lee, G. H. Huang, S. J. Ashford, Jayanti, S. P. Palupi, and K. A. Notodiputro, "Job Insecurity and the Changing Workplace: Recent Developments and the Future Trends in Job Insecurity Research,", *Annual Review of Organizational Psychology and Organizational Behaviour*, vol. 5, no 1, pp. 335–359, 2018.

[6] R Nikhil, N. Tikoo, S. Kurle, H. S. Pisupati, and G. R. Prasa, "Sentiment Analysis of Twitter Users' Opinion Towards Face-to-Face Learning,", *Journal of Emerging Technologies and Innovative Research (JETIR)*, vol. 2, no. 4, pp. 1292-1296, 2015.

[7] D. Jayanti, S. P. Palupi, and K. A. Notodiputro, "A Conditional Logistic Regression Model for Analyzing Unemployment Rates in West Java: Model Regresi Logistik Bersyarat untuk Analisis Tingkat Pengangguran di Provinsi Jawa Barat,", *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 195–204, 2021.

[8] A. Z. Taufan and W. Wibowo, "Analisis Sentimen Terkait Persepsi Keamanan Data Informasi Dan Privasi di Indonesia Menggunakan Pendekatan Machine Learning,", *JINTEKS*, vol. 6, no. 3, pp. 728-736, 2024.

[9] Y. Zhang, Z. Xu, and J. Lin, "Text Mining for Social Media Analysis: Trends and Techniques,", *Journal of Data Science*, vol. 20, no. 1, pp. 65–80, 2022.

[10] S. M. Mohammad. "Sentiment analysis: Automatically detecting valence, emotions, and other affectual states from text,". *In H. L. Meiselman (Ed.), Emotion measurement (2nd ed.) Elsevier*, pp. 323–379), 2021.

[11] A. F. Ningrum and I. F. Amri, "Sentiment Analysis Of Public Opinion On Handling Stunting In Indonesia Using Random Forest,", *J Statistika*, vol. 17, no. 1, pp. 645–654, 2024.

[12] D. K. A. Astutik, A. Indrasetianingsih, F. Fitriani, "Penerapan Text Mining pada Analisis

Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode Density Based Spatial Clustering of Applications With Noise (DBSCAN) dan K-Means,", *J Statistika*, vol. 15, no. 1, pp. 184-194, 2022.

[13] R. Singh, and A. Gupta, "Precision and Recall in Machine Learning Algorithms,", *Journal of Machine Learning Applications*, vol. 11, no. 3, pp. 213–221, 2020.

[14] N. Thamrin, A. F. Baktiar, F. A. Addawiyah, M. Husna, and T. Irwati, "Determinants of Antenatal Care Visits in Indonesia with Synthetic Minority Over-Sampling Techniques for Imbalance Data: Determinan Kunjungan Antenatal Care di Indonesia dengan Teknik Synthetic Minority Over-Sampling untuk Imbalanced Data,", *Indonesian Journal of Statistics and Its Applications*, vol. 7, no. 2, pp. 86–104, 2023.