JURNAL STATISTIKA DAN APLIKASINYA

e-ISSN 2620 – 8369
VOLUME X (ISSUE X) MONTH YEAR: XX-XX
https://doi.org/10.21009/JSA.081XX

SPATIAL REGRESSION AND MACHINE LEARNING FOR JAVA'S OPEN UNEMPLOYMENT ANALYSIS (INDONESIA 2023)

Amira Shohifa[1*], Pinky Siwi Nastiti[2], Nathania Fiorella Harditama[3]

[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia
Depok City 16424, West Java, Indonesia

Corresponding author's e-mail: * *amira.shohifa@ui.ac.id* [1*], *pinky.siwi@ui.ac.id* [2], *nathania.fiorella@ui.ac.id* [3]

Article History:
Available online.

Keywords:
Geospatial Analysis; Predictive Modeling; Random Forest; Spatial Autocorrelation; Unemployment Determinants

ABSTRACT

Background – Open Unemployment Rate (TPT) is a critical socio-economic issue in Indonesia, especially in Java, with significant regional disparities. Traditional models fall short in capturing its complex, non-linear, and spatial aspects. This study addresses this by comparing spatial regression and machine learning to improve TPT prediction and identify key determinants.

Purpose – This research aims to analyze socio-economic factors influencing TPT across 118 districts/cities in Java (2023). It compares spatial regression and machine learning models to predict TPT and identify the most significant predictors.

Methodology – Data from BPS included TPT (dependent) and Labor Force Participation Rate (TPAK), Life Expectancy at Birth (UHH), Expected Years of Schooling (HLS), Adjusted Per Capita Expenditure (PKD), and Poverty Percentage (PPM)(independent variables). Analysis began with OLS and Moran's I for spatial autocorrelation. Spatial regression models (SAR, SEM, SARMA) were developed. For comparison, machine learning algorithms (Random Forest, Decision Tree, SVR, GBM) were applied, incorporating hyperparameter tuning, cross-validation, and spatial features.

Findings – Moran's I confirmed significant spatial dependency. An optimized Random Forest (RF CV) model showed the best TPT prediction performance (R^2=0.926). The spatial lag of TPT, TPAK, and spatial coordinates (longitude) were the most crucial predictors in the best RF model.

Research limitations – This study used cross-sectional data (2023), limiting temporal analysis. The independent variables, while relevant, were not exhaustive. The Random Forest model's interpretation is descriptive of variable importance, not direct causal inference.

Originality/value – This research offers a comprehensive comparison of spatial regression and tuned machine learning models with spatial features for TPT in Java. It highlights the Random Forest model's superior predictive capability and identifies crucial (spatial) predictors, providing insights for precise data-driven policy.

1. Introduction

Java Island is the most populous island in Indonesia and serves as the country's economic center. Despite overall improvements in employment indicators, significant disparities in the Open Unemployment Rate (TPT) still persist across its districts and cities. The TPT, defined as the percentage of the labor force actively seeking work but currently unemployed, remains high in certain regions, such as Banten Province, which recorded an OUR of 6.64% in 2025—one of the highest in Java according to the Central Bureau of Statistics (BPS) [1].

These spatial disparities indicate that unemployment is not randomly distributed and may be influenced by local socioeconomic factors such as the Labor Force Participation Rate (TPAK), Life Expectancy at Birth (UHH), Expected Years of Schooling (HLS), Per Capita Expenditure (PKD), and Poverty Rate (PPM). Previous studies have shown the importance of accounting for spatial dependence when analyzing regional unemployment [2]. Spatial regression models, including the Spatial Autoregressive Model (SAR), Spatial Error Model (SEM), and Spatial Autoregressive Moving Average (SARMA), are increasingly used to capture these effects more accurately [3]. Meanwhile, machine learning methods such as Random Forest and Gradient Boosting have also gained popularity for handling complex, nonlinear relationships in socioeconomic data [4].

However, there is still limited research that integrates both spatial econometric approaches and machine learning models in analyzing regional unemployment. Most existing studies either focus on one approach or limit their analysis to a single province. This study addresses that gap by comparing spatial regression models and machine learning algorithms in modeling the TPT across districts and cities in Java Island for the year 2023. Understanding the combined spatial and nonlinear influences of multiple factors is essential for producing accurate predictive models and informing data-driven employment policy decisions.

2. Methods

Material and Data

The data used in this study consist of information from 118 regencies/municipalities across six provinces in Java Island, Indonesia, for the year 2023. The primary variables analyzed include the Open Unemployment Rate (TPT) (%) as the dependent variable. Independent variables comprise the Labor Force Participation Rate (TPAK) (%), Life Expectancy at Birth (UHH) (years), Expected Years of Schooling (HLS) (years), Adjusted Per Capita Expenditure (PKD) (thousand rupiah/person/year), and Poverty Rate (PPM) (%). Geographic coordinate data (latitude and longitude) for each regency/municipality were also utilized to facilitate spatial analysis. All data were obtained from official publications and the website of Statistics Indonesia (Badan Pusat Statistik/BPS) for the relevant year. A sample of the data structure is presented in Table 1.

Table 1. Sample of Data Used for Regencies/Cities in Java Island, 2023

Regencies /Cities	Long.	Lat.	TPT (%)	TPAK (%)	UHH (years)	HLS (years)	PKD (Rp)	PPM (%)
Cilacap	109.015	-7.725	8.74	66.6	74.25	12.67	11432	10.99
Banyumas	109.3	-7.5167	6.35	64.6	73.98	13.26	12492	12.53
Purbalingga	109.3667	-7.3833	5.61	73.45	73.37	12.02	10964	14.99
.	.	.	.	.	.	.	.	.

.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
Kota Surabaya	112.75	-7.25	6.76	68.73	74.75	14.85	18977	4.65
Kota Batu	112.5167	-7.8667	4.52	78.99	73.29	14.56	13603	3.31

Source: Statistics Indonesia (BPS) West Java Province, 2024.

Research Method

This study employed a quantitative approach involving several analytical stages. The research object is the Open Unemployment Rate (TPT) and its relationship with selected socio-economic variables across districts/cities in Java Island for the year 2023. The analytical process adopted is as follows:

1. **Descriptive Analysis and Spatial Data Preparation:** Initial data exploration involved descriptive statistics and visualizations (boxplots, thematic maps) for each variable to understand their distributions and spatial patterns. Spatial weight matrices, necessary for spatial analysis, were constructed based on queen contiguity.
2. **Baseline Linear Regression Modeling:** An Ordinary Least Squares (OLS) regression model was first estimated to examine the initial relationships between TPT and the independent variables. Classical assumption tests (normality of residuals via Shapiro-Wilk, multicollinearity via VIF, heteroscedasticity via Breusch-Pagan, and residual autocorrelation via Durbin-Watson) were conducted.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon \quad (1)$$

Where Y is the dependent variable, X_1, X_2, ..., X_k are the independent variables, β_0 is the intercept, β_1, β_2, ..., β_k are the regression coefficients, and ε is the error term assumed to be normally distributed.

Linear regression is categorized into two types: simple linear regression, which involves one independent variable, and multiple linear regression, which involves two or more independent variables. The model is typically estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals. For the OLS estimators to be valid and unbiased, several assumptions must be met, including:

- Linearity: The relationship between dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of residuals is constant across all levels of the independent variables.
- Normality: The residuals are normally distributed.

These assumptions are typically tested using various statistical diagnostics, such as:

- Shapiro-Wilk test for normality of residuals,
- Breusch-Pagan test for heteroscedasticity,
- Durbin-Watson test for autocorrelation.

3. **Spatial Autocorrelation Diagnostics:** Moran's I statistic was used to test for spatial autocorrelation in the dependent variable (TPT), each independent variable, and the residuals from the OLS model to determine the presence and nature of spatial dependencies.

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2} \quad (2)$$

where x_i represents the value of the variable at location *i*, and w_{ij} is the element of the spatial weight matrix in the *i*-th row and *j*-th column. The value of Moran's I ranges from [-1,1], where I = 0 indicates no spatial autocorrelation, I = -1 indicates perfect negative spatial autocorrelation, and I = 1 indicates perfect positive spatial autocorrelation.

4. **Spatial Regression Modeling:** Based on the diagnostic tests (including Lagrange Multiplier tests for spatial lag and spatial error dependence), relevant spatial regression models were estimated. The models considered were the Spatial Autoregressive Model (SAR), Spatial Error Model (SEM), and Spatial Autoregressive Moving Average (SARMA). Parameter estimation and significance testing were performed for each model. Residuals from these spatial models were also checked for remaining spatial autocorrelation.

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (3)$$

- Spatial Autoregressive Model (SAR)

  The SAR model assumes that the dependent variable in one region is directly influenced by the values of the dependent variable in neighboring regions. A significant and positive ρ suggests that high unemployment in surrounding regions is associated with high unemployment in a given region.

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4)$$

- Spatial Error Model (SEM)

  The SEM model assumes that spatial dependence exists not in the dependent variable directly, but in the error terms. This suggests that unobserved variables influencing unemployment may be spatially structured. A significant λ indicates that neighboring areas share similar unmeasured influences.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (5)$$

- Spatial Autoregressive Moving Average (SARMA)

The SARMA model captures spatial effects in both the outcome variable and the unobserved components. It is the most flexible of the three and is appropriate when both types of spatial dependencies—direct and in the residuals—are present in the data.

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \tag{6}$$

5. **Machine Learning Modeling:** Several machine learning algorithms were employed for comparative purposes in predicting TPT. These included Decision Tree, Random Forest, Support Vector Regression (SVR), and Gradient Boosting Machine (GBM). To optimize their performance, hyperparameter tuning and 5-fold cross-validation were implemented. Spatial features, specifically geographical coordinates (latitude, longitude) and a spatially lagged TPT variable, were incorporated as predictors in these models. The potential for residual spatial autocorrelation in the best-performing ML model was also assessed.
6. **Model Evaluation and Comparison:** All developed models (OLS, SAR, SEM, SARMA, and the optimized ML models) were evaluated and compared based on standard performance metrics: Akaike Information Criterion (AIC) for regression models, Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Visual analysis of residual maps was also used to assess model fit. The best overall model was then selected based on these comprehensive evaluations.

The theoretical underpinnings of Moran's I [5, 6], spatial effect tests such as the Lagrange Multiplier tests [5], specific spatial regression models (SAR, SEM, SARMA) [5], machine learning algorithms including Decision Tree [7], Random Forest [7, 8], Support Vector Regression (SVR) [7, 9], and Gradient Boosting Machine (GBM) [7, 10], as well as model evaluation criteria (AIC, R^2, RMSE, MAE) [7, 11], are detailed extensively in standard econometric and machine learning literature.

3. Results

This section presents the main findings from the analysis of the Open Unemployment Rate (TPT) across 118 regencies/cities in Java Island, Indonesia, for the year 2023.

3.1 Descriptive Overview of Variables

Descriptive analysis revealed considerable variation in TPT and its potential socio-economic determinants across Java. The TPT ranged from 1.52% in Pangandaran Regency to a high of 10.52% in Cimahi City. Similar geographical disparities and distributional characteristics were observed for the Labor Force Participation Rate (TPAK), Life Expectancy at Birth (UHH), Expected Years of Schooling (HLS), Adjusted Per Capita Expenditure (PKD), and Poverty Rate (PPM). Thematic maps indicated spatial clustering for these variables, suggesting regional patterns that warrant spatial analysis. Detailed visualizations including boxplots, bar charts of highest values, and distribution maps for each variable are presented in **Figure 1** to **Figure 6**.

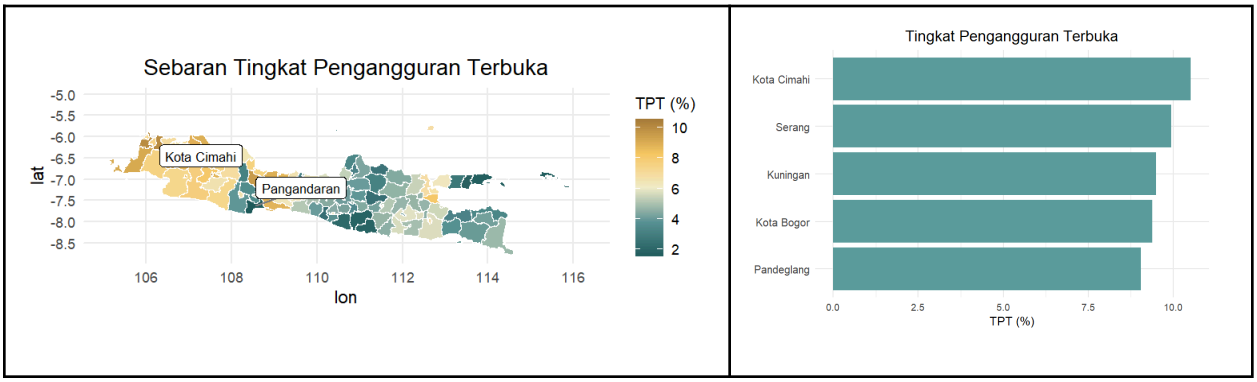

Figure 1. Descriptive Visualization of Open Unemployment Rate (TPT): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that Cimahi City had the highest Open Unemployment Rate (TPT) among all regencies/municipalities in Java Island in 2023, recording a rate of 10.52%. Conversely, the region with the lowest TPT was Pangandaran Regency, with a rate of 1.52%.

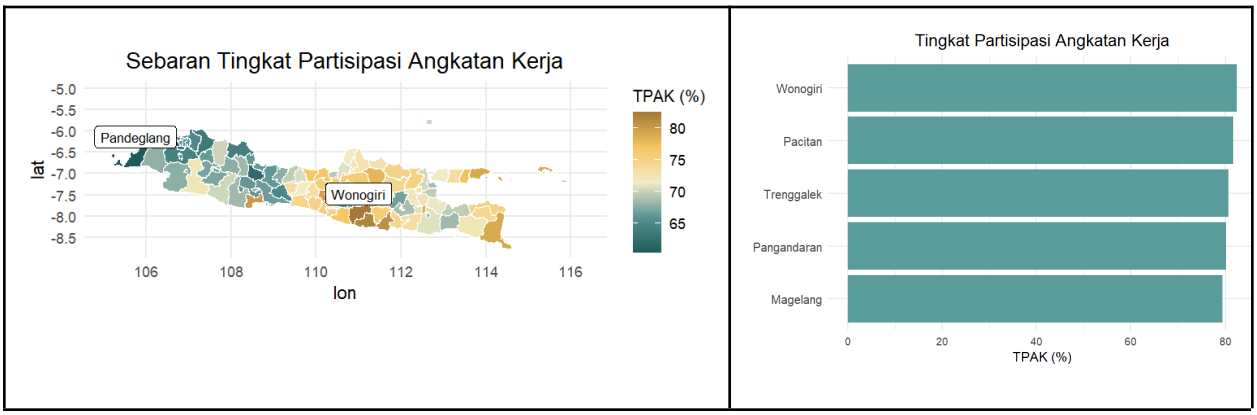

Figure 2. Descriptive Visualization of Labor Force Participation Rate (TPAK): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that Wonogiri Regency had the highest Labor Force Participation Rate (TPAK) among all regencies/municipalities in Java Island in 2023, recording a rate of 82.45%. In contrast, the lowest TPAK was observed in Pandeglang Regency, with a rate of 60.33%.

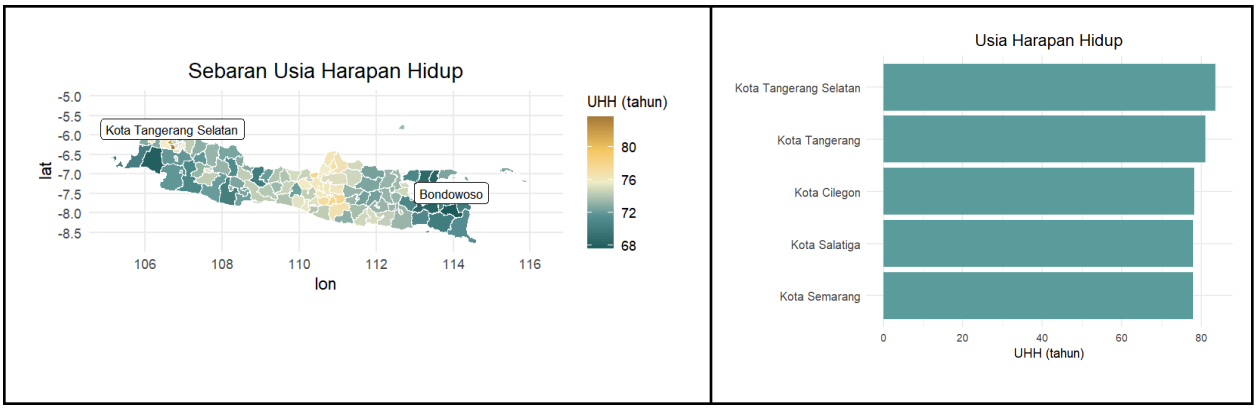

Figure 3. Descriptive Visualization of Life Expectancy at Birth (UHH): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that South Tangerang City had the highest Life Expectancy at Birth among all regencies/municipalities in Java Island in 2023, recording a value of 83.57 years. In contrast, Bondowoso Regency had the lowest life expectancy, with a value of 67.6 years.

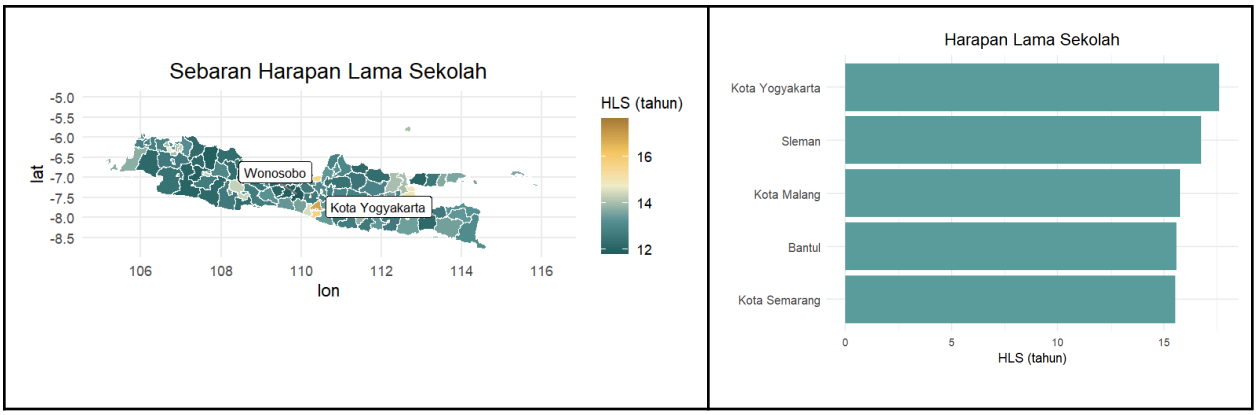

Figure 4. Descriptive Visualization of Expected Years of Schooling (HLS): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that Yogyakarta City had the highest Expected Years of Schooling in Java Island in 2023, with a value of 17.62 years. Meanwhile, the lowest was recorded in Wonosobo Regency, at 11.8 years.

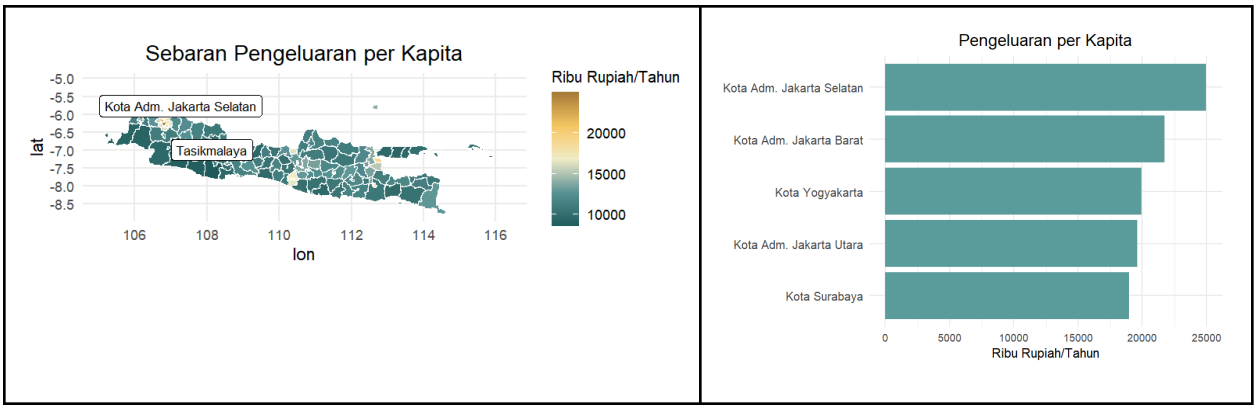

Figure 5. Descriptive Visualization of Adjusted Per Capita Expenditure (PKD): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that South Jakarta Administrative City had the highest Per Capita Expenditure in Java Island in 2023, amounting to Rp24,975 per person per year. On the other hand, Tasikmalaya Regency had the lowest, with Rp8,562 per person per year.

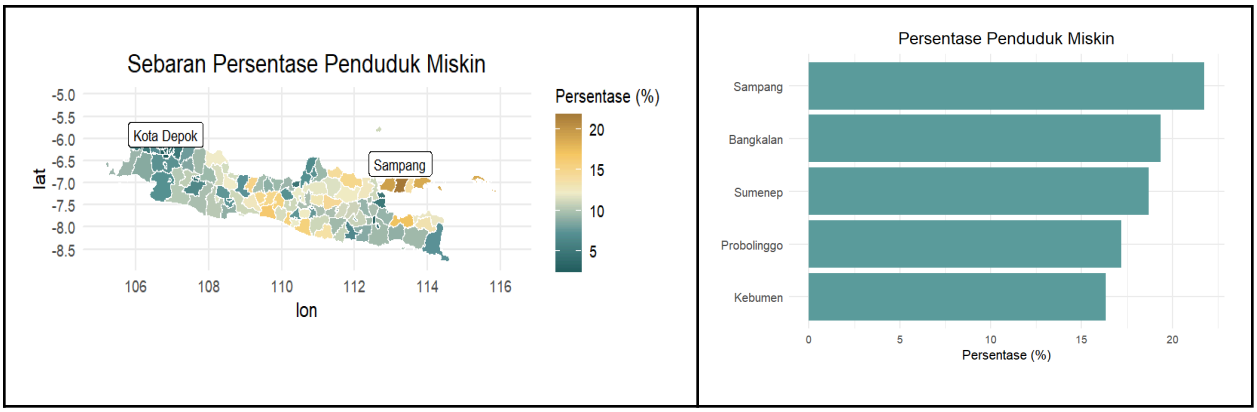

Figure 6. Descriptive Visualization of Poverty Percentage (PPM): Boxplot, Bar Chart of Highest Values, and Distribution Map

Based on the distribution map and bar chart, it was found that Sampang Regency had the highest Poverty Rate in Java Island in 2023, recorded at 21.76%. In contrast, the lowest Poverty Rate was observed in Depok City, at 2.38%.

3.2 Baseline Regression and Spatial Dependence

The Ordinary Least Squares (OLS) regression model was initially estimated to examine the relationships between TPT and the independent variables. The parameter estimates are presented in **Table 2**. The OLS model indicated that TPAK and PPM had a statistically significant negative relationship with TPT. Classical assumption tests were conducted: residuals were found to be normally distributed (Shapiro-Wilk test, **Table 3**), no significant multicollinearity was detected (VIF values, **Table 4**), residuals were homoscedastic (Breusch-Pagan test, **Table 5**), and no significant classical autocorrelation was found in the residuals (Durbin-Watson test, **Table 6**).

Table 2. Parameter Estimation Results of the Ordinary Least Squares (OLS) Model

Parameter	Estimation	Standard Error	t_{value}	$p - value$
β_0 (Intercept)	35.2	4.921	7.154	0.000000000096
β_1 (TPAK)	-0.2491	0.03035	-8.205	0.000000000000453
β_2 (UHH)	-0.1091	0.06206	-1.758	0.08148
β_3 (HLS)	-0.1682	0.1624	-1.036	0.30264
β_4 (PKD)	-0.00003007	0.00006562	-0.458	0.64762
β_5 (PPM)	-0.1525	0.04570	-3.338	0.00115

Based on Table 2, the multiple linear regression model obtained is as follows:

$$TPT = 35.2 - 0.2491\,TPAK - 0.1091\,UHH - 0.1682\,HLS - 0.00003007\,PKD - 0.1525\,PPM$$

The Ordinary Least Squares (OLS) regression results indicate that two independent variables—Labor Force Participation Rate (TPAK) and Poverty Rate (PPM)—are statistically significant at the 5% significance level, as evidenced by their p-values of less than 0.05. Both variables have negative coefficients, suggesting that higher labor force participation and lower poverty levels are associated with lower Open Unemployment Rates (TPT).

On the other hand, Life Expectancy at Birth (UHH) has a p-value of 0.08148, indicating marginal significance at the 10% level, while Expected Years of Schooling (HLS) and Per Capita Expenditure are not statistically significant, with p-values of 0.30264 and 0.64762 respectively. Although statistically insignificant, their negative coefficients suggest a potential inverse relationship with TPT.

Table 3.Shapiro-Wilk Normality Test Results for OLS Model

Shapiro Wilk Test	$p - value$
0.98912	0.4781

The p-value of 0.4781 > 0.05 indicates that the residuals are normally distributed, satisfying the normality assumption.

Table 4. Multicollinearity Test Results (VIF Values) for OLS Model

Variabel Independen	X_1	X_2	X_3	X_4	X_5
Nilai VIF	1.2339	1.2999	1.5757	2.0607	1.6958

All VIF values are below 10, indicating no multicollinearity problem among the independent variables.

Table 5. Breusch-Pagan Homoscedasticity Test Results for OLS Model

Breusch-Pagan	$p - value$
7.8523	0.1646

The p-value of 0.1646 > 0.05 indicates that the variance of residuals is constant, thus the homoscedasticity assumption is met.

Table 6. Durbin-Watson Autocorrelation Test Results for OLS Model

DW-Statistics	$p - value$
2.0278	0.4998

With a DW value of 2.0278 and p-value of 0.4998 > 0.05, there is no indication of autocorrelation in the residuals.

However, Moran's I tests revealed significant positive spatial autocorrelation in the dependent variable (TPT, **Table 7**), most independent variables (TPAK, **Table 8**; UHH, **Table 9**; HLS, **Table 10**; PKD, **Table 11**; PPM, **Table 12**), and critically, in the OLS residuals (Moran's I = 0.2542, p-value < 0.05, **Table 13**). This confirmed the necessity for spatial modeling approaches.

Table 7. Spatial Autocorrelation Test Results (Moran's I) for Open Unemployment Rate (TPT)

Moran's I Statistics	$p - value$
0.6582	< 2.2e-16

Moran's I = 0.6582 with p-value < 2.2e-16 indicates a strong and significant positive spatial autocorrelation in TPT.

Table 8. Spatial Autocorrelation Test Results (Moran's I) for Labor Force Participation Rate (TPAK)

Moran's I Statistics	$p - value$
0.6033	< 2.2e-16

Moran's I = 0.6033 with p-value < 2.2e-16 shows a significant positive spatial pattern in TPAK across regions.

Table 9. Spatial Autocorrelation Test Results (Moran's I) for Life Expectancy at Birth (UHH)

Moran's I Statistics	$p - value$
0.5402	< 2.2e-16

Moran's I = 0.5402 and p-value < 2.2e-16 suggest significant spatial clustering in life expectancy values.

Table 10. Spatial Autocorrelation Test Results (Moran's I) for Expected Years of Schooling (HLS)

Moran's I Statistics	$p - value$
0.3676	9.016e-09

Moran's I = 0.3676 and p-value = 9.016e-09 indicate moderate but significant positive spatial autocorrelation in HLS.

Table 11. Spatial Autocorrelation Test Results (Moran's I) for Adjusted Per Capita Expenditure (PKD)

Moran's I Statistics	$p - value$
0.5179	1.61e-15

Moran's I = 0.5179 with p-value = 1.61e-15 confirms significant spatial dependence in per capita expenditure.

Table 12. Spatial Autocorrelation Test Results (Moran's I) for Poverty Percentage (PPM)

Moran's I Statistics	$p - value$
0.5053	1.365e-14

Moran's I = 0.5053 with p-value = 1.365e-14 indicates strong positive spatial autocorrelation in poverty rates.

Table 13. Spatial Autocorrelation Test Results (Moran's I) for OLS Model Residuals

Moran's I Statistics	$p - value$
0.2542	5.118e-05

Moran's I = 0.2542 and p-value = 5.118e-05 show significant residual spatial autocorrelation, suggesting that the OLS model does not fully account for spatial effects.

3.3 Spatial Regression Model Analysis

Lagrange Multiplier (LM) tests indicated significant spatial lag and spatial error dependencies, suggesting SAR, SEM, and SARMA models were all potentially appropriate (**Table 14**).

- The Spatial Autoregressive Model (SAR) estimation (koefisien di **Table 15**) identified TPAK as a significant negative predictor, with a significant spatial lag coefficient (ρ=0.5034). The AIC for SAR was 399.6184 (**Table 16**), and its residuals showed no remaining spatial autocorrelation (**Table 17**).
- The Spatial Error Model (SEM) estimation (koefisien di **Table 18**) also found TPAK significant, with a significant spatial error coefficient (λ=0.5469). Its AIC was 409.6056 (**Table 19**), and residuals were spatially random (**Table 20**).
- The Spatial Autoregressive Moving Average (SARMA) model estimation (koefisien di **Table 21**) showed a significant spatial lag coefficient (ρ=0.57595) while the spatial error coefficient was not significant. TPAK was identified as a significant predictor. The AIC was 401.0987 (**Table 22**), with spatially random residuals (**Table 23**).

Table 14. Lagrange Multiplier (LM) Test Results for Spatial Regression Model Specification

Test	$p - value$
Lagrange Multiplier (error)	0.0002238
Lagrange Multiplier (lag)	0.0000007325

Robust LM (error)	0.4178
Robust LM (lag)	0.0006721
Lagrange Multiplier (SARMA)	0.000003399

Table 15. Parameter Estimation Results for the Spatial Autoregressive (SAR) Model

Parameter	Estimation	$p - value$
ρ	0.5034	0.0000001474
β_0 (Intercept)	18.727	0.0001135
β_1 (TPAK)	-0.15545	0.00000008432
β_2 (UHH)	-0.06673	0.2063074
β_3 (HLS)	-0.045696	0.7413537
β_4 (Pengeluaran)	-0.000020351	0.7110041
β_5 (PPM)	-541.5	0.1774808

The spatial autoregressive coefficient $\rho = 0.5034$ is significant ($p < 0.001$), indicating strong spatial dependence in the dependent variable (TPT). Among the predictors, only TPAK is statistically significant ($p < 0.05$) with a negative coefficient, suggesting that higher labor force participation is associated with lower TPT. Other variables, including UHH, HLS, expenditure, and PPM, are not significant.

Table 16. Akaike Information Criterion (AIC) Value for the SAR Model

Model	AIC
SAR	399.6184

The SAR model has an AIC value of 399.6184. This value can be used to compare with other models (e.g., SEM, SARMA), where a lower AIC indicates a better model fit.

Table 17. Spatial Autocorrelation Test Results (Moran's I) for SAR Model Residuals

Moran's I Statistics	$p - value$
-0.0289	0.618

Moran's I = -0.0289 with a p-value of 0.618 shows no significant spatial autocorrelation in the residuals. This confirms that the SAR model effectively accounts for spatial dependence in the data.

Table 18. Parameter Estimation Results for the Spatial Error (SEM) Model

Parameter	Estimation	$p - value$
λ	0.5469	0.000026748
β_0 (Intercept)	24.794	0.00003158
β_1 (TPAK)	-0.18054	0.00000007456
β_2 (UHH)	-0.067286	0.32455
β_3 (HLS)	-0.045655	0.78941
β_4 (Pengeluaran)	-0.000011934	-0.045655
β_5 (PPM)	-0.086129	0.06402

The spatial error coefficient $\lambda = 0.5469$ is significant ($p < 0.001$), indicating the presence of spatial autocorrelation in the error term. Among the explanatory variables, only TPAK is statistically significant ($p < 0.05$) with a negative coefficient, suggesting that higher labor force participation is associated with lower unemployment. Other variables, UHH, HLS, expenditure, and PPM are not statistically significant.

Table 19. Akaike Information Criterion (AIC) Value for the SEM Model

Model	AIC
SEM	409.6056

The SEM model has an AIC value of 409.6056, which can be compared with other models (e.g., SAR, SARMA) to evaluate model performance, where lower AIC indicates a better fit.

Table 20. Spatial Autocorrelation Test Results (Moran's I) for SEM Model Residuals

Moran's I Statistics	$p - value$
-0.0173	0.5514

Moran's I = -0.0173 with a p-value of 0.5514 shows no significant spatial autocorrelation in the residuals. This suggests that the SEM model effectively addresses spatial dependence in the error component.

Table 21. Parameter Estimation Results for the Spatial Autoregressive Moving Average (SARMA) Model

Parameter	Estimation	$p - value$
ρ	0.57595	0.0000035978
λ	-0.15061	0.48062
β_0 (Intercept)	-0.14239	0.003664

β_1 (TPAK)	-0.059041	0.00002105
β_2 (UHH)	-0.064847	0.243083
β_3 (HLS)	0.064847	0.625977
β_4 (Pengeluaran)	-0.0000191	0.709313
β_5 (PPM)	-0.043674	0.287517

The spatial autoregressive coefficient $\rho = 0.57595$ is statistically significant ($p < 0.001$), indicating strong spatial dependence in the dependent variable. However, the spatial error coefficient $\lambda = -0.15061$ is not significant ($p = 0.48062$), suggesting weak spatial dependence in the residuals. Among the explanatory variables, only TPAK is significant ($p < 0.05$), with a negative coefficient, showing that higher labor force participation is associated with lower unemployment. Other variables (UHH, HLS, expenditure, and PPM) are not significant.

Table 22. Akaike Information Criterion (AIC) Value for the SARMA Model

Model	AIC
SARMA	401.0987

The AIC value for the SARMA model is 401.0987. This is lower than the AIC of SAR and SEM, indicating a better overall model fit among the spatial regression models evaluated.

Table 23. Spatial Autocorrelation Test Results (Moran's I) for SARMA Model Residuals

Moran's I Statistics	$p - value$
-0.0006137949	0.4528

Moran's I = -0.0006 with a p-value of 0.4528 indicates no significant spatial autocorrelation remaining in the residuals, confirming that SARMA successfully accounts for both spatial lag and spatial error structures.

Visual inspection of residual maps (**Figure 7** for OLS, **Figure 8** for SAR, **Figure 9** for SEM, and **Figure 10** for SARMA) suggested that while SAR and SEM improved upon OLS, the SARMA model provided the most randomly distributed residuals with lower extreme values among the spatial regression models.

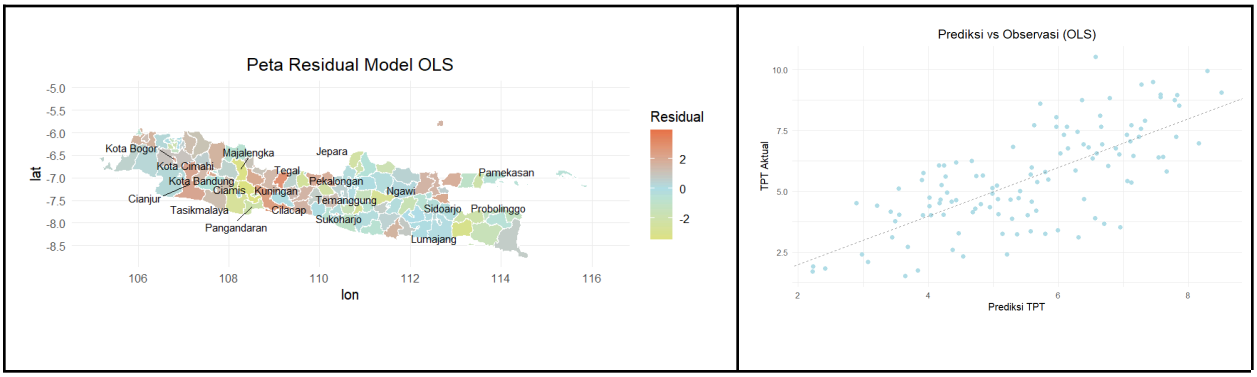

Figure 7. Residual Distribution Map and Predicted vs. Observed Plot for the OLS Model

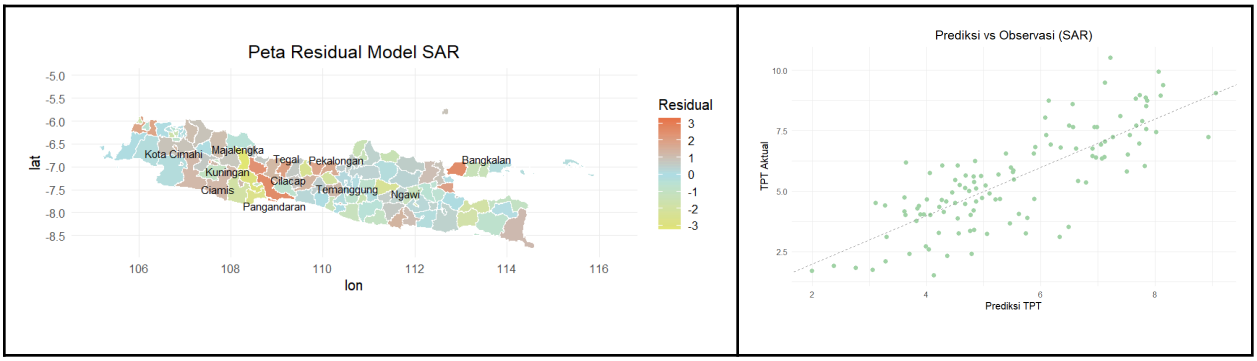

Figure 8. Residual Distribution Map and Predicted vs. Observed Plot for the SAR Model

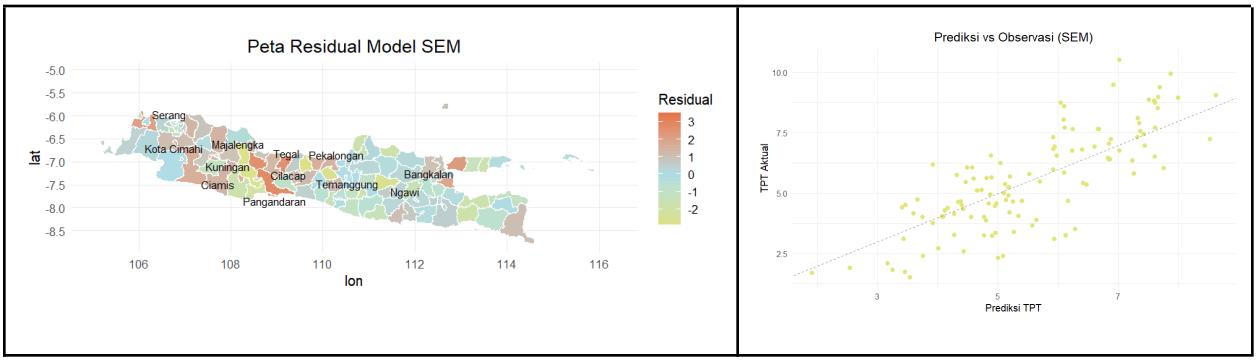

Figure 9. Residual Distribution Map and Predicted vs. Observed Plot for the SEM Model

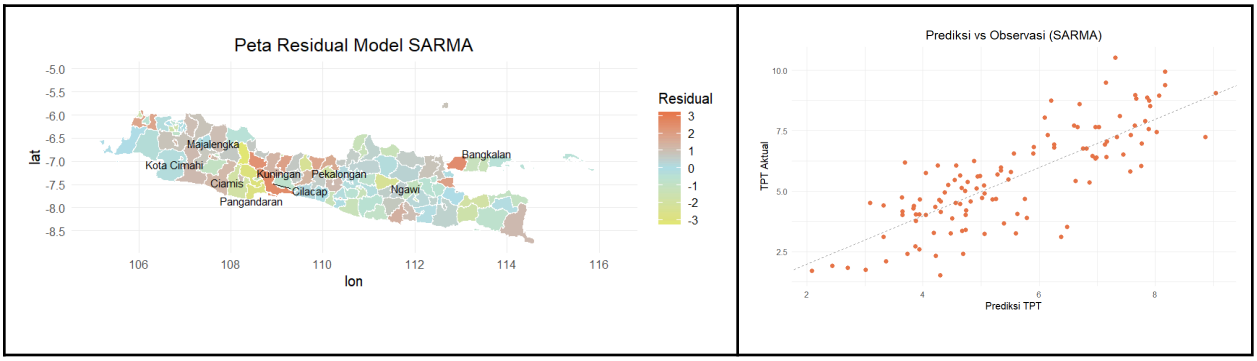

Figure 10. Residual Distribution Map and Predicted vs. Observed Plot for the SARMA Model

3.4 Machine Learning Model Performance

Several machine learning algorithms with incorporated spatial features (coordinates and spatial lag of TPT) were implemented and optimized using hyperparameter tuning and 5-fold cross-validation.

- **Baseline Model Performance:** Initial modeling without extensive tuning showed GBM with the best baseline performance (RMSE = 0.04489, R2=0.9995), while Random Forest had the lowest baseline performance (**Table 24**, **Figure 11**).

Table 24. Comparative Evaluation Metrics of Baseline Machine Learning Models

Model	RMSE	MAE	R^2
Random Forest	1.29479696	1.01581714	0.5919062
SVR	0.20257761	0.17663250	0.9900106
Decision Tree	0.38885176	0.29752512	0.9631935
GBM	0.04489176	0.02138769	0.9995094

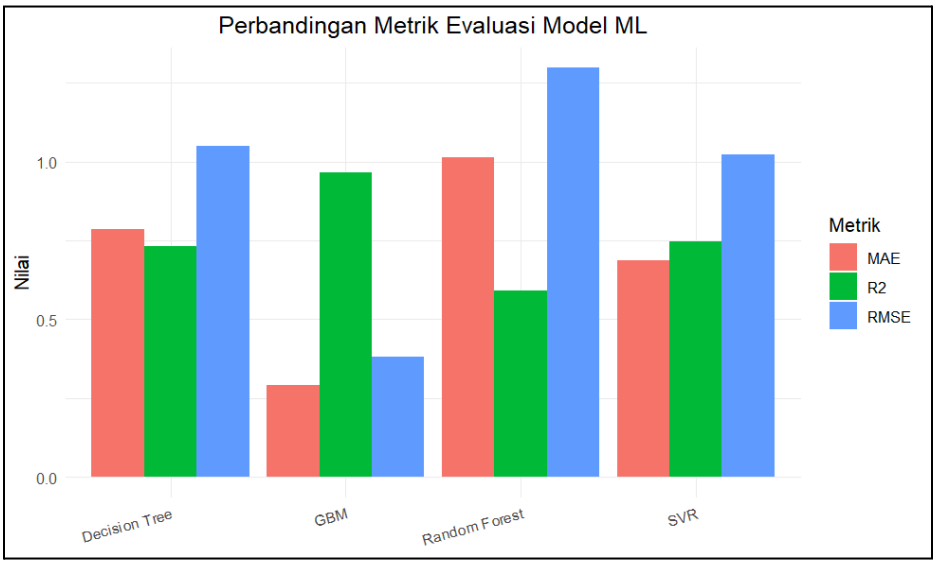

Figure 11. Bar Chart Visualization of Comparative Evaluation Metrics for Baseline Machine Learning Models

- **Performance after Tuning and Cross-Validation:** After optimization, the Random Forest (RF CV) model emerged as superior, achieving the lowest RMSE (0.5507), lowest MAE (0.4232), and highest R2 (0.9262) among all tested ML models (**Table 25**, **Figure 12**). SVR performance decreased, while GBM, though good (R2=0.8621), was outperformed by RF CV.

Table 25. Comparative Evaluation Metrics of Machine Learning Models after Hyperparameter Tuning and Cross-Validation (CV)

Model	RMSE	MAE	R^2
Random Forest	0.5507086	0.4232292	0.9261755
SVR	1.1225803	0.7975597	0.6932451
Decision Tree	1.0943497	0.8244831	0.7984796
GBM	0.7527869	0.5673499	0.8620566

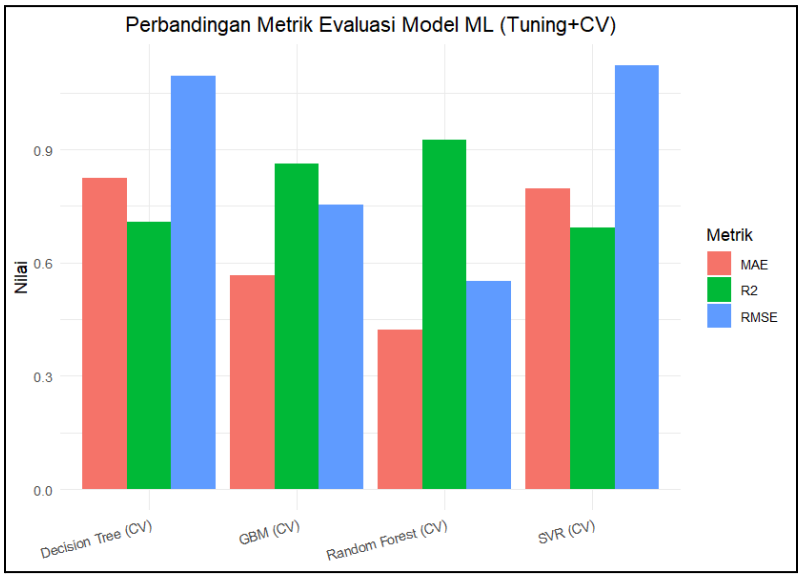

Figure 12. Bar Chart Visualization of Comparative Evaluation Metrics for Machine Learning Models after Hyperparameter Tuning and Cross-Validation (CV)

- **SpatialRF Diagnostics:** Analysis using the spatialRF package on a Random Forest model incorporating spatial features showed normally distributed residuals (Shapiro-Wilks W = 0.985, p-value = 0.2323) and no significant spatial autocorrelation in residuals (Moran's I = -0.015, p-value = 0.748) (**Figure 13 & Figure 14**). This indicated that the inclusion of spatial features in the Random Forest model effectively captured the spatial structure.

```
Model type
  - Fitted with:                    ranger()
  - Response variable:              TPT

Random forest parameters
  - Type:                           Regression
  - Number of trees:                500
  - Sample size:                    117
  - Number of predictors:           8
  - Mtry:                           2
  - Minimum node size:              5


Model performance
  - R squared (oob):                 0.6036958
  - R squared (cor(obs, pred)^2):    0.9383942
  - Pseudo R squared (cor(obs, pred)):0.9687075
  - RMSE (oob):                      1.281445
  - RMSE:                            0.5708
  - Normalized RMSE:                 0.1981944

Model residuals
  - Stats:
                ┌──────┬─────────┬────────┬──────┬─────────┬──────┐
                │ Min. │ 1st Q.  │ Median │ Mean │ 3rd Q.  │ Max. │
                ├──────┼─────────┼────────┼──────┼─────────┼──────┤
                │-1.72 │  -0.29  │  0.04  │ 0.00 │  0.31   │ 1.50 │
                └──────┴─────────┴────────┴──────┴─────────┴──────┘
  - Normality:
      - Shapiro-Wilks W: 0.985
      - p-value        : 0.2323
      - Interpretation : Residuals are normal

  - Spatial autocorrelation:
               ┌──────────┬──────────┬─────────┬──────────────────┐
               │ Distance │ Moran's I│ P value │ Interpretation   │
               ├──────────┼──────────┼─────────┼──────────────────┤
               │      0.0 │   -0.015 │   0.748 │ No spatial       │
               │          │          │         │ correlation      │
               └──────────┴──────────┴─────────┴──────────────────┘
```

Figure 13. Performance Summary of Random Forest Model from spatialRF::rf() Function (referring to the Model performance output block)

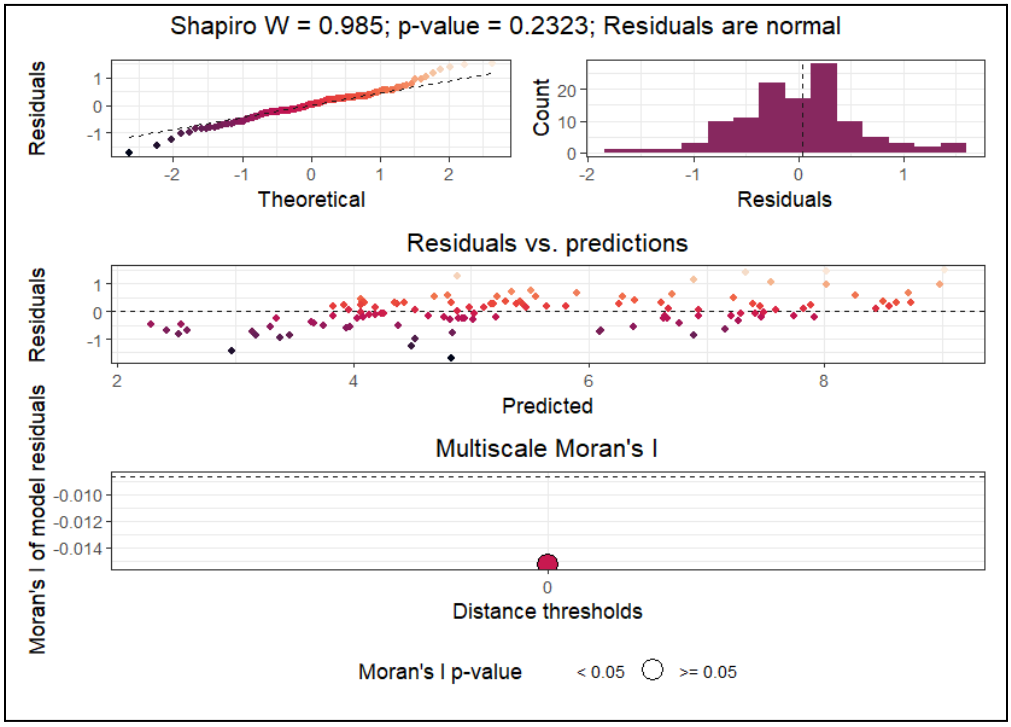

Figure 14. Diagnostic Visualization of Random Forest Model using rf() function from spatialRF Package (QQ Plot, Histogram, Residuals vs. Predictions, Multiscale Moran's I)

- **Residual Analysis of Best ML Model (RF CV):** The Moran's I test on the residuals of the final tuned RF CV model also confirmed no significant spatial autocorrelation (p-value = 0.9267, **Table 26**). The residual map for RF CV (**Figure 15**) further supported this, showing a fairly random distribution.

Table 26. Spatial Autocorrelation Test Results (Moran's I) for Random Forest (CV) Model Residuals

Moran's I Statistics	$p - value$
0.2542	0.9267

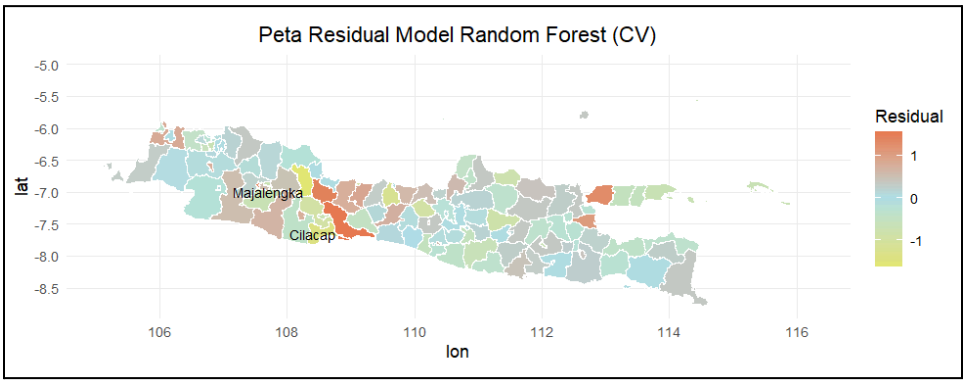

Figure 15. Residual Distribution Map for the Random Forest Model (After Tuning + CV)

3.5 Overall Model Comparison and Key Predictors

A comprehensive comparison of all models is presented in **Table 27**. The Random Forest (RF CV) model demonstrated superior predictive performance (R^2= 0.9262) over OLS (R^2= 0.5210) and all spatial regression models (e.g., SARMA R^2= 0.6579).

Table 27. Comprehensive Comparison of Evaluation Metrics for All Models (OLS, Spatial Regression, and Machine Learning)

Model	RMSE	MAE	R^2	AIC
OLS	1.4028477	1.1149021	0.5209534	425.2416
SAR	1.2056499	0.9453483	0.6461664	399.6184
SEM	1.2494951	0.9805988	0.6199630	409.6056
SARMA	1.0943497	0.9306115	0.6579023	401.0987
Random Forest	0.5507086	0.4232292	0.9261755	NA
SVR	1.1225803	0.7975597	0.6932451	NA
Decision Tree	1.0943497	0.8244831	0.7984796	NA
GBM	0.7527869	0.5673499	0.8620566	NA
OLS	1.4028477	1.1149021	0.5209534	425.2416
SAR	1.2056499	0.9453483	0.6461664	399.6184

Variable importance analysis from the best model (RF CV) revealed that the spatial lag of TPT (lag_TPT) was the most crucial predictor, followed by TPAK, longitude (lon), and per capita expenditure (Pengeluaran) (**Figure 16**). This highlights the significance of spatial interdependencies and specific socio-economic factors in explaining TPT variations.

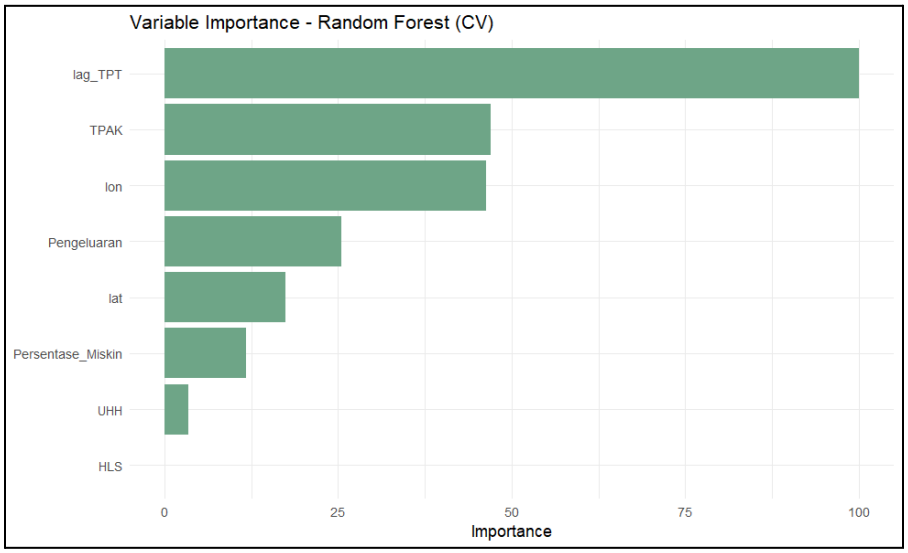

Figure 16. Bar Chart Visualization of Variable Importance from the Random Forest Model (After Tuning + CV)

4. DISCUSSIONS

This study compared spatial regression and machine learning models to analyze the Open Unemployment Rate (TPT) in Java Island, Indonesia, for 2023, identifying key determinants and the most effective modeling approach. The findings offer several points for discussion regarding methodological choices and policy implications.

The initial confirmation of significant positive spatial autocorrelation in TPT and its related variables underscored the necessity of moving beyond non-spatial OLS regression and adopting spatially-aware models. While traditional spatial regression models (SAR, SEM, SARMA) successfully accounted for this spatial dependency in their residuals, this research primarily highlights the superior predictive performance of machine learning. Specifically, an optimized Random Forest (RF CV) model, incorporating spatial features (coordinates and TPT spatial lag), achieved the highest accuracy (R2=0.926), significantly outperforming both OLS and the suite of spatial regression models (e.g., SARMA R2=0.658). This suggests that the complex, non-linear interactions and potential spatial heterogeneity inherent in TPT determinants across Java are more effectively captured by flexible, non-parametric machine learning approaches.

The importance of the TPT spatial lag (lag_TPT) as the top predictor in the best RF model strongly reaffirms the presence of spatial spillover effects in regional unemployment. This aligns with studies like Fauzi et al. (2023) which also emphasized spatial effects, though using different spatial models. The finding that Labor Force Participation Rate (TPAK) and spatial coordinates (longitude, potentially proxying regional economic disparities) are also key predictors offers further avenues for understanding unemployment dynamics. The superior performance of a well-tuned Random Forest with spatial features over traditional spatial econometric models in this context contributes to a growing body of literature, similar to trends observed by Sari & Wulandari (2023) in broader unemployment risk analysis using machine learning, by providing a direct comparison for TPT across a large, diverse region

4.1 Research Implications

The findings carry several implications:

- **Theoretical and Methodological**: This study demonstrates that while established spatial econometric models are vital for diagnosing and incorporating specific spatial dependence structures (lag, error), machine learning models, when thoughtfully augmented with spatial features, can offer enhanced predictive power for complex socio-economic phenomena like TPT. The robust performance of the RF CV model, despite its non-parametric nature, in handling spatial data effectively suggests its value as an alternative or complementary tool in spatial analysis.

- **Policy and Practical**:
    1. The dominance of the TPT spatial lag (lag_TPT) as a predictor strongly suggests that unemployment reduction strategies should adopt a regional perspective, considering inter-district/city spillovers rather than focusing solely on isolated administrative areas.
    2. The identification of TPAK and longitude (as a potential proxy for broader regional economic conditions) as important factors points towards the need for policies that enhance labor market engagement tailored to regional economic structures and address spatial economic disparities across Java.
    3. Given the superior predictive accuracy of the optimized Random Forest model, government agencies could leverage such machine learning tools for more precise TPT

forecasting and identifying high-risk areas, enabling proactive and targeted interventions, aligning with the JSA journal's focus on research impacting economic and social development.

4.2 Limitations and Future Research

This study utilized cross-sectional data for 2023, limiting insights into temporal dynamics. Future research could benefit from employing spatial panel data models. Additionally, incorporating a broader range of variables (e.g., investment, sectoral structure) and exploring other advanced spatial regression (like GWR) or more sophisticated spatial machine learning algorithms could provide deeper understanding. While this study identifies strong predictors, further research focusing on causal inference would be valuable for direct policy formulation.

5. Conclusion

This study successfully modeled the relationship between the Open Unemployment Rate (TPT) and various socioeconomic factors in 118 districts/cities across Java Island in 2023, effectively addressing spatial dependencies. Initial analysis revealed significant positive spatial autocorrelation in TPT and most independent variables, as well as in the residuals of the Ordinary Least Squares (OLS) model, confirming the necessity of incorporating spatial effects into the analysis. While spatial regression models (SAR, SEM, and SARMA) were employed and managed to account for the spatial autocorrelation present in the OLS residuals, their predictive performance was ultimately surpassed by machine learning approaches. The core objective of comparing these methodologies was met, with the optimized Random Forest model (RF CV) demonstrating superior predictive capability for TPT, achieving the highest R^2 of 0.926 and the lowest RMSE (0.5507) and MAE (0.4232) among all models tested, including the spatial regression models. This underscores the effectiveness of machine learning, especially when augmented with spatial features, in capturing complex, non-linear patterns within the data that traditional regression models might miss. Finally, the research identified the most influential variables impacting TPT, fulfilling another key objective. The spatial lag of TPT (lag_TPT) emerged as the most crucial predictor, followed by the Labor Force Participation Rate (TPAK) and spatial coordinates (longitude), indicating the strong influence of neighboring regions and specific socioeconomic dynamics on unemployment rates. This finding provides valuable insights for data-driven policy recommendations, suggesting that unemployment reduction strategies should adopt a regional, interconnected perspective rather than focusing on isolated administrative areas.

6. Acknowledgments

The authors would like to express their sincere gratitude to Dr. Dra. Yekti Widyaningsih, M.Si., for her guidance, insights, and continuous support throughout the research process. Special thanks are also extended to Badan Pusat Statistik (BPS) Jawa Barat for providing the valuable data that made this study possible.

7. References

[1] Badan Pusat Statistik. (2025). *Keadaan Ketenagakerjaan Indonesia Februari 2025*. Jakarta: BPS.

[2] Fauzi, F., Wenur, G. H., & Wasono, R. (2023). Spatial Durbin Model of Unemployment Rate in Central Java. *Parameter: Journal of Statistics*, 3(1), 7–18. https://doi.org/10.22487/27765660.2023.v3.i1.16423

[3] LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. CRC Press.

[4] Sari, R. P., & Wulandari, D. (2023). Machine Learning Implementation for Analyzing Unemployment Risk. *Seminar Nasional Official Statistics*. https://prosiding.stis.ac.id/index.php/semnasoffstat/article/view/2222

[5] Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer.

[6] Cliff, A. D., & Ord, J. K. (1981). *Spatial Processes: Models & Applications*. Pion.

[7] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[8]Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[9] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[11] Gujarati, D. N., & Porter, D. C. (2009). Basic econometrics (5th ed.). McGraw-Hill Education.