DOI: https://doi.org/10.34123/jurnalasks.v16i1.587

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

**Jurnal Aplikasi Statistika & Komputasi Statistik**
**Vol. 16 No. 1, June 2024**
***e- ISSN: 2615-1367, p-ISSN:2086-4132***

# Application of Geographically Weighted Logistic Regression in Modeling The Human Development Index in East Java

**Toha Saifudin[1*], Leni Sartika Panjaitan[2], Sabrina Falasifah[3], Yan Dwi Pracoko[4]**

[1,2,3,4]*Airlangga University, Surabaya, Indonesia*
**Corresponding Author: E-mail address: [1]tohasaifudin@fst.unair.ac.id, [2]leni.sartika.panjaitan-2020@fst.unair.ac.id, [3]sabrina.falasifah-2020@fst.unair.ac.id, [4]yan.dwi.pracoko-2020@fst.unair.ac.id*

**ARTICLE INFO**

***Article history:***
Received 13 June, 2023
Revised 10 May, 2024
Accepted 07 May, 2024
Published 30 June, 2024

***Keywords:***
AIC; HDI; GWLR Method; Fixed Gaussian Kernel; Logistic Regression

**Abstract**

**Introduction/Main Objectives:** pinpoint the factors influencing HDI, taking into consideration location and spatial factors. **Background Problems:** The Human Development Index (HDI) in East Java often fails to reflect actual conditions accurately, as disparities exist among districts and cities, with some falling below government expectations. Novelty: GWLR extends logistics regression by incorporating spatial factors, allowing for the identification of regional differences and influential factors affecting HDI based on actual data. **Research Methods:** To address this issue, the Geographically Weighted Logistic Regression (GWLR) method is employed. The independent variables used are Expected Years of Schooling ($X_1$), Open Unemployment Rate ($X_2$), and Morbidity Rate ( $X_3$) in 2021, while dependent variable is the Human Development Index (Y). **Finding/Results:** The study reveals that GWLR provides a superior model compared to Ordinary Logistic Regression, indicated by a lower Akaike Information Criterion (AIC) of 28.72. Additionally, the GWLR model with Fixed Gaussian Kernel weights outperforms other weighting methods. At 90% confidence level, the significant variables influencing HDI are expected years of schooling ( $X_1$) and the open unemployment rate ($X_2$). Given the relatively low HDI in Indonesia, the East Java Government should focus on improving these key areas to enhance HDI across districts and cities in the region.

## 1. Introduction

The Sustainable Development Goals (SDGs) consist of 17 goals that must be attained [1]. Within the realm of sustainable development, human development is paramount. The SDGs address human development, specifically the third goal (ensuring healthy lives and promoting well-being), the fourth goal (ensuring inclusive and equitable quality education), and the seventh goal (promoting sustained, inclusive, and sustainable economic growth).

The United Nations Development Program (UNDP) defines human development as a continuous process that involves making choices that contribute to a long and healthy life, acquiring knowledge, and living a decent life with access to essential resources. The Human Development Index (HDI) serves as a metric to assess a community's level of development in terms of health, education, income, and

other related factors. HDI achievements are categorized into four groups: very high (with values of 80 and above), high (with values between 70 and below 80), medium (with values between 60 and below 70), and low (with values below 60) [2].

HDI in East Java varies significantly across districts/cities due to distinct location characteristics and differing development priorities. Despite the annual increase in HDI, the growth does not entirely mirror the true state of HDI in the region. Disparities exist among districts/cities, with some falling below government expectations.

Geographically Weighted Logistic Regression (GWLR) is a variant of logistic regression that incorporates spatial or location factors when predicting a categorical dependent variable [3]. By incorporating a weighting function, the model takes into account the geographical location of the observed data. The local nature of GWLR allows it to address the issue of spatial heterogeneity, making it a valuable tool in overcoming this challenge.

Indah Manfaati Nur and M. Al Haris [4] conducted a study on the factors affecting HDI in Central Java using the Geographically Weighted Logistic Regression (GWLR) method. The research revealed that the influencing factors included the number of health facilities, literacy rate, morbidity rate, and open unemployment rate. Additionally, Lili et al. [5] conducted a similar study on the factors influencing HDI in Kalimantan, also utilizing the Geographically Weighted Logistic Regression (GWLR) method. Their findings indicated that the influencing factors consisted of the percentage of open unemployment, the percentage of the population with university degrees, the percentage of the poor, and the number of health workers, including doctors, midwives, nurses, and pharmacists.

Based on this explanation, a study was conducted in East Java to model HDI using the Geographically Weighted Logistic Regression (GWLR) method. This modeling aimed to pinpoint the factors influencing HDI, taking into consideration location and spatial factors. The distinct models for each location enable researchers to gain a more specific insight into the challenges faced in different areas. The goal of this research is to assist the East Java Government in enhancing these influencing factors to elevate HDI levels in every district/city in the province and work towards achieving the third, fourth, and eighth Sustainable Development Goals (SDGs).

## 2. Material and Methods

### *2.1 Human Development Index (HDI)*

The Human Development Index (HDI) serves as a holistic indicator that measures human development by considering three fundamental aspects of well-being: health, education, and living standards. Established by the United Nations Development Program (UNDP), this index utilizes a range of metrics to assess development achievements, including average years of schooling, life expectancy, school enrollment rates, and per capita income levels [6]. To gauge the level of human development in different countries and regions worldwide, the commonly used metric is the Human Development Index (HDI). This index employs a three-dimensional geometric average, with each dimension being standardized on a scale of 0 to 1. A higher HDI value indicates a higher level of human development. Unlike more conservative measures such as income and economic growth, the HDI acknowledges that factors like healthcare, education, access to social services, and others significantly impact human well-being.

Development goes beyond the mere enhancement of per capita income; it also encompasses various other facets of society. Relying solely on Gross Domestic Product (GDP) growth is inadequate when addressing human development. It is imperative to consider additional factors such as societal challenges, shifts in people's perceptions and behaviors, and more [7-8]. The quality of Human Resources (HR) stands out as a critical factor for successful development. Human resources, as a development target, serve as a driving force for progress that impacts the success of development initiatives. With proficient human resources, training programs could be effectively carried out. Competent human resources play a crucial role in a country's development [9-10-11].

In 2021, East Java recorded an HDI of 71.4, signifying a substantial level of human development in the region. The HDI value for East Java is calculated based on data obtained from diverse sources, including BPS, the Ministry of Health, the Ministry of Education and Culture, and other relevant institutions. These sources supply data on critical indicators such as life expectancy at birth, years of

schooling, and per capita income. The 2021 East Java HDI report underscores the significant progress made by the province across several domains. For instance, life expectancy at birth in East Java has increased from 69.8 years in 2015 to 70.9 years in 2021. Moreover, the literacy rate for individuals aged 15 and above has shown an upward trend, rising from 96.2% in 2015 to 98.2% in the period of 2015-2021 [12].

Several factors impact the attainment of the HDI growth goal in East Java [13-14].

1. Expected Years of Schooling

   BPS report highlights that the Expected Years of Schooling (HLS) represents the projected length of education (in years) that children of a specific age group are expected to undergo in the future [12]. The significance of expected schooling duration in shaping educational development at each level is duly acknowledged by considering the expected duration of schooling for individual children. This underscores the commitment to providing quality education to all children [15].
2. Open Unemployment Rate

   The open unemployment rate specifically denotes the proportion of individuals who are not currently engaged in any form of employment. Unemployment could be attributed to various factors, and one of the significant causes is a decline in economic growth. Additionally, the decline in industrial development and the substitution of human labor with advanced technology directly contribute to unemployment [16].
3. Pain Rate

   The pain rate, as stipulated in Health Law no. 28, refers to the numerical or proportional representation of individuals afflicted with a particular illness within a specific community over a designated timeframe [17]. Morbidity statistics serve as an accurate reflection of the prevailing circumstances, as they exhibit a strong correlation with factors such as local poverty levels, living conditions, the standard of potable water, and the quality of healthcare services [18].

## *2.2 Data Source*

This research makes use of secondary data obtained from BPS East Java [12]. The focus of this study comprises all districts and cities within East Java in 2021, encompassing a total of 38 districts/cities.

### *2.2.1 Research Variable*

The research encompasses two categories of variables: response variables and predictor variables. The specifics are outlined as follows:

1. Response Variables

   The dependent variable (Y) represents the Human Development Index (HDI) of all districts/cities in East Java in 2021. This variable is measured on a nominal scale. According to the classification by BPS (2014), HDI is considered low if it is below 60, medium if it is $60 \leq IPM < 70$, high if it ranges from $70 \leq IPM < 80$, and very high if it is $IPM \geq 80$. For this research, binary responses are utilized, allowing the classification based on the BPS criteria as follows:

$0 = Low\ if\ the\ HDI\ number < 70$

$1 = High\ if\ the\ HDI\ number \geq 70$

2. Predictor Variables

   The data used as predictors corresponds to the year 2021. Table 1 provides an inventory of several predictor variables that are significant for this specific research.

**Table 1.** Research predictor variables

| Var | Variable Name | Data Type |
|---|---|---|
| X1 | Expected Years of Schooling | Continuous |
| X2 | Open Unemployment Rate | Continuous |
| X3 | Pain Rate | Continuous |

## 2.3 Binary Logistic Regression Analysis

Logistic regression is a statistical technique commonly used to determine the association between a categorical dependent variable and one or more independent variables [19-20]. Unlike linear regression, logistic regression utilizes a binary dummy variable for the dependent variable, eliminating the need for normality, heteroscedasticity, or autocorrelation assumption tests [21]. However, it cannot handle multicollinearity, which occurs when predictor variables counteract each other. Despite this limitation, logistic regression offers the advantage of Y estimating a wide range of values for the dependent variable, potentially extending beyond the conventional 0 to 1 range. In the case of a dependent variable encompassing two categories with values of 0 and 1, binary logistic regression is employed. When utilizing a model, the response variable adheres to a Bernoulli distribution.

$$f(y_i) = \pi_i(x_i)^{y_i}(1-\pi_i(x_i))^{1-y_i} \quad (1)$$

In the context provided, $\pi_i$ signifies the probability assigned to the i event, while $y_i$ refers to the ith random variable that could take on the values of 0 or 1.

The multivariable logistic regression model is created by representing the value of E(Y=1|X) as π(x), leading to the following equation:

$$\pi(x_i) = \frac{expexp\,(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k)}{1+expexp\,(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k)} \quad (2)$$

where k denotes the number of predictor variables

Equation (2) could be represented as equation (3) in the following manner:

$$\pi(x) = \frac{expexp\,(g(x))}{1+expexp\,(g(x))} \quad (3)$$

To facilitate the estimation of regression parameters, the logistic regression model could be converted into the π(x) equation (3). This transformation leads to the logit form of logistic regression, resulting in the following equation:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \quad (4)$$

Estimating unknown parameters in logistic regression involves the utilization of the *Maximum Likelihood Estimation* (MLE) technique. This method focuses on maximizing the likelihood function to estimate the β parameter, necessitating that the data follows a specific distribution. In binary logistic regression, every observation satisfies a *Bernoulli distribution*, allowing for the derivation of a *likelihood* function. The following section presents a systematic description of the *likelihood* function for a binary logistic regression model.

$$L(\beta) = \prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad (5)$$

To simplify calculations, the likelihood function is maximized as a log function so that from equation (6) the *log-likelihood* function is obtained as follows:

$$l(\beta) = \sum_{i=1}^{n}[y_i\ ln\ ln\ \pi(x_i)\ +\ (1-y_i)\ ln\ ln\ (1-\pi(x_i)) \quad (6)$$

Obtaining estimates for the parameter vector β involves maximizing the *likelihood* function. This is done by equating the first derivative of the likelihood function for each parameter to zero. $\left(\frac{\partial l(\beta)}{\partial \beta} = 0\right)$, It follows from equation (7) that:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{N}X_{ij}(y_i-\pi(x_i)) = 0 \quad (7)$$

given j=0,1,2,…,k, it follows that $X_{i0}$=1. Due to the non-linear nature of the derivative of $l(\beta)$ concerning $\beta_j$ for j=0,1,2,…,k, the *Newton-Raphson* method is utilized to calculate the estimate of $\boldsymbol{\beta}$ denoted as $\hat{\beta}$.

## 2.4 Testing Spatial Assumptions

Spatial influences in data are determined through various testing methods, including the examination of spatial dependence and spatial heterogeneity. The *Breusch-Pagan* test method is employed to assess spatial heterogeneity, whereas the *Morans'I* test method is utilized to evaluate spatial dependence. In the case of partial testing using the *Breusch-Pagan* test, the hypothesis is as follows:

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

$H_1 = At\ the\ very\ least, there\ is\ one\ \sigma_i^2 \neq \sigma$

The test statistics applied in the *Breusch-Pagan* test are.

$$BP = \left(\frac{1}{2}\right) f^T Z (Z^T Z)^{-1} Z^T f \tag{8}$$

The vector components of $f$ are given by $f_i = \left(\frac{e_i}{\sigma^2} - 1\right)$, where $e_i = y_i - \hat{y}$ and $\sigma^2$ represents the variance of y. On the other hand, $Z$ is an $n \times p$ matrix that includes standardized normal vectors for each observation. The critical region, denoted as $H_0$, is rejected when the BP test statistic value exceeds $BP > \chi^2_{(p)}$, suggesting spatial heterogeneity in the data.

In the context of spatial dependency testing using the *Morans'I* test method, the null hypothesis $H_0$ could be formulated as follows:

$H_0$ : The data does not exhibit any spatial dependency.

$H_1$ : The data exhibits spatial dependency.

The *Morans'I* test employs the following test statistics for conducting the analysis.

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{9}$$

where is the value $E(I) = -\frac{1}{n-1}$ and $var(I) = \frac{(n^2 S_1 - n S_2 + 3W^2)}{W^2(n^2-1)} - [E(I)]^2$ . Spatial dependency in the data is inferred when the test statistic value $|Z| = Z_{\frac{\alpha}{2}}$, leading to the rejection of the critical region $H_0$.

## 2.5 Geographically Weighted Logistic Regression (GWLR) Model

Geographically Weighted Logistic Regression (GWLR) is a regression analysis technique that incorporates spatially varying coefficients into a logistic regression model [22]. Unlike traditional logistic regression models where coefficients are assigned to all observations, GWLR considers location dependence of coefficients to explain spatial heterogeneity in the data [23]. The GWLR method incorporates the geographic location of areas by using a weighting function, where each observation is assigned a weight $(w_{ij})$. Thus, the resulting model from equation (2) could be expressed as follows:

$$\pi(x_i) = \frac{exp(\sum_{k=0}^{p} \beta_k(u_i, v_i) x_{ik})}{1 + exp(\sum_{k=0}^{p} \beta_k(u_i, v_i) x_{ik})} \tag{10}$$

where the regression coefficient $\beta_k(u_i, v_i)$ denotes the impact of the predictor variable $x_{ik}$ at a particular location $(u_i, v_i)$.

In the case of Geographically Weighted Logistic Regression (GWLR), the logit form is given by:

$$\left(\pi(x_i)\right) = ln\left(\frac{\pi(x_i)}{1-(x_i)}\right) = x_i \beta(u_i, v_i) \tag{11}$$

where $0 < \pi(x_i) < 1$

The response variable ($Y_i$) follows a *Bernoulli distribution* with the probability function given by:

$$Pr(Y = y_i) = \pi^{y_i}\left(1 - \pi(x_i)\right)^{1-y_i} \quad (12)$$

where value $y_i = 0,1$

Utilizing the *Maximum Likelihood Estimator* (MLE) technique, the parameters of the GWLR model are estimated. The *log-likelihood* function of the logistic spatial model is derived by maximizing the likelihood function in log form:

$$ln\ ln\ L\left(\beta(u_i, v_i)\right) = ln\ ln\ \left(\prod_{i=1}^{n}\left[\frac{(expexp\,(x_i^T\beta))^{y_i}}{1+expexp\,(x_i^T\beta)}\right]\right) \quad (13)$$

## 2.6 Odds Ratio

The odds ratio value is determined by comparing the probability of success to the probability of failure. The equation used to estimate the odds ratio value is:

$$Odd = \frac{\pi(x)}{1-\pi(x)} = exp(X_i\beta) \quad (14)$$

where $i = 1,2, \dots, n$ and $0 < \pi(x) < 1$

## 2.7 Selection of Best Bandwidth and Model

The radius of the circle, known as *bandwidth*, assumes that the points within this radius continue to exert influence. In GWLR modeling, the role of *bandwidth* is of utmost importance as it directly affects the accuracy of the model by adjusting the variance and bias in the data.

The Akaike Information Criterion (AIC) method is the preferred approach for model selection. When dealing with a large sample size in the context of the GWLR method, the AIC formula is utilized:

$$AIC = 2k - 2ln(\hat{L}) \quad (15)$$

The calculation for AIC in the context of small sample sizes through the GWLR method is outlined in the equation:

$$AICc = AIC + \frac{2k^2-2k}{n-k-1} \quad (16)$$

The model that demonstrates the lowest AIC value is regarded as the most favorable model, where *k* denotes the number of parameters and *n* represents the number of samples.

## 2.8 GWLR Model Fit Test

To assess the level of geographic influence, hypothesis testing for the GWLR model involves conducting similarity tests and simultaneous tests between logistic regression models. A similarity test was carried out between the GWLR model and the logistic regression model to determine the significant level of influence. The hypothesis being tested is as follows:

$H_0: \beta_k(u_i, v_i) = \beta_k\ ; k = 1,2, \dots, p$ (The GWLR and logistic regression models exhibit no substantial disparity)

$H_1: At\ the\ very\ least, there\ is\ one\ \beta_k(u_i, v_i) \neq \beta_k$ (The GWLR and logistic regression models display significant disparities)

The first stage of the Maximum Likelihood Ratio Test (MLRT) method involves deriving the test statistics by determining the set of parameters

within the population $(\Omega) = \{\beta_0(u_i, v_i),\ \beta_1(u_i, v_i), \dots, \beta_k(u_i, v_i)\}$. Then, the likelihood function is constructed as shown below:

$$L(\Omega) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{17}$$

$$L(\hat{\Omega}) = L(\Omega)$$

$$= \prod_{i=1}^{n} \pi(x_i)^{y_{1}j}(1-\pi(x_i))^{y_{0}i} \tag{18}$$

Begin by establishing the parameters set under $H_0(\omega), \omega = \{\beta_0(u_i, v_i)\}$. Proceed to formulate both the likelihood function and the maximum likelihood function.

$$L(\omega) = \prod_{i=1}^{n} \pi(x_i)^{y_{1}i}(1-\pi(x_i))^{y_{0}i} \tag{19}$$

with

$$L(\hat{\omega}) = L(\omega)\ \prod_{i=1}^{n} \left\{ \left(\frac{n_{1i}}{n}\right)^{y_{1}i} \left(\frac{n_{0i}}{n}\right)^{y_{0}i} \right\} \tag{20}$$

the ratio equation of the maximum likelihood function under $H_0$ and the maximum likelihood function under the population could be expressed as follows, where $n_{1i}$ represents the number of i observations included in category 1, $n_{0i}$ represents the number of i observations included in category 0, and n represents the total number of observations.

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{i=1}^{n} \left\{ \left(\frac{n_{1i}}{n}\right)^{y_{1}i} \left(\frac{n_{0i}}{n}\right)^{y_{0}i} \right\}}{\prod_{i=1}^{n} \hat{\pi}(x_i)^{y_{1}i}(1-\hat{\pi}(x_i))^{y_{0}i}} \tag{21}$$

Equation (21) provides insight into what could be derived from the GWLR model deviation:

$$D(\hat{\beta}^*) = -2\ ln\ ln\ \Lambda\ = 2\left[ln\ ln\ L(\hat{\Omega})\ - ln\ ln\ L(\hat{\omega})\ \right] \tag{22}$$

For instance, if $D(\hat{\beta})$ represents the deviance of the logistic regression model with $db_1$ degrees of freedom, and $D(\hat{\beta}^*)$ represents the deviance of the GWLR model with $db_2$ degrees of freedom, then the test statistics used to examine the correlation between these two models are:

$$F_{hit} = \frac{D(\hat{\beta})/db_1}{D(\hat{\beta}^*)/db_2} \tag{23}$$

The F distribution with $db_1$ and $db_2$ degrees of freedom is attained by the test statistic in equation (23). To assess the GWLR model parameters collectively, the testing criteria involve rejecting $H_0$ when the $F_{hit} > F_{\alpha,db_1,db_2}$. Evaluate the GWLR model parameters concurrently by considering the following hypotheses:

$H_0: \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \cdots = \beta_p(u_i, v_i) = 0\ ; i = 1,2, \dots, n$

$H_1: At\ the\ very\ least, there\ is\ one\ \beta_k(u_i, v_i) \neq 0\ ; k = 1,2, \dots, p$

According to the MLRT model, the test statistics used for simultaneous testing are as follows:

$$G^2 = 2\left[ln\ ln\ L(\hat{\Omega})\ - ln\ ln\ L(\hat{\omega})\ \right] \tag{24}$$

The test statistic $G^2$ in equation (24) approximates the chi-square distribution with degrees of freedom $v = n - k - 1$. The criteria for the test involve rejecting $H_0$ if the value of $G^2 > \chi^2_{\alpha,v}$.

## *2.9 Partial Test of the GWLR Model*

GWLR model parameter testing determines which parameters have the greatest impact on the model. The partial test hypothesis for the $\beta_k$ parameter is as follows:

$H_0: \beta_k(u_i, v_i) = 0$ ; *i = 1, 2, ..., n; k=1, 2, ..., p*
$H_1: \beta_k(u_i, v_i) \neq 0$

The *Wald* test could be utilized to obtain test statistics for this test, as demonstrated below:

$$Z_{hit} = \frac{\widehat{\beta}_k(u_i,v_i)}{Se\left(\widehat{\beta}_k(u_i,v_i)\right)} \; ; k = 1,2, \dots p \tag{25}$$

The test statistic presented in equation (25) provides an approximation of the standard normal distribution. The purpose of the test is to reject the null hypothesis, $H_0$ if the absolute value $|Z_{hit}| > Z_{\frac{\alpha}{2}}$.

## *2.10 GWLR Model Classification Accuracy*

The classification accuracy of the GWLR model could be assessed by computing the Apparent Error Rate (APPER) value, which represents the likelihood of error in object classification. The APPER value could be calculated using the following formula:

$$APPER = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \times 100 \tag{26}$$

where,
$n_{11}$ = Number of observed errors that fall into the error category based on the prediction results
$n_{12}$ = Number of observed errors that fall into the success category based on the prediction results
$n_{21}$ = Number of observed successful events that fall into the error category based on the prediction results
$n_{22}$ = Number of observed success events that fall into the success category based on the predicted results

## *2.11 Research Procedure*

The GWLR method allows for the analysis procedure to be conducted in the following manner:

1. Based on a thematic map generated with ArcGIS software, the map illustrates the factors that impact the Human Development Index of districts/cities in East Java.
2. The process of modeling and estimating Human Development Index data utilizing the Geographically Weighted Logistic Regression (GWLR) approach could be carried out using GWR4 software through the subsequent steps:
    a. Validate the fundamental assumptions of spatial regression, specifically examining spatial heterogeneity through the Breusch-Pagan technique.
    b. Verify the core assumptions of spatial regression, particularly focusing on spatial dependence using Moran's I approach.
    c. Establish the optimal bandwidth (h) by utilizing the CV method.
    d. Compute the weighting matrix using kernel functions such as Fixed Gaussian, Fixed Bisquare, and Adaptive Gaussian. Subsequently, identify the most suitable kernel function by comparing the AIC values.
    e. Evaluate the appropriateness of the GWLR model using logistic regression.
    f. Conduct partial testing on the GWLR parameters and identify the variables that impact the HDI in each district/city.
    g. Develop a GWLR model specific to each district/city within East Java Province.
    h. Estimate the Odd Ratio value of predictor variables that exhibit a significant impact.
    i. Contrast the Logistic Regression model with the GWLR Model.
3. Examine and explain the different factors that play a crucial role in determining the Human Development Index in East Java Province and assess the precision of the categorization.
4. Formulate conclusions

## 3. Result and discussion

The obtained data is utilized to classify dependent and independent variables based on districts/cities in East Java. Thematic maps are employed for this purpose. The subsequent classification represents the research variables categorized according to districts/cities in East Java.

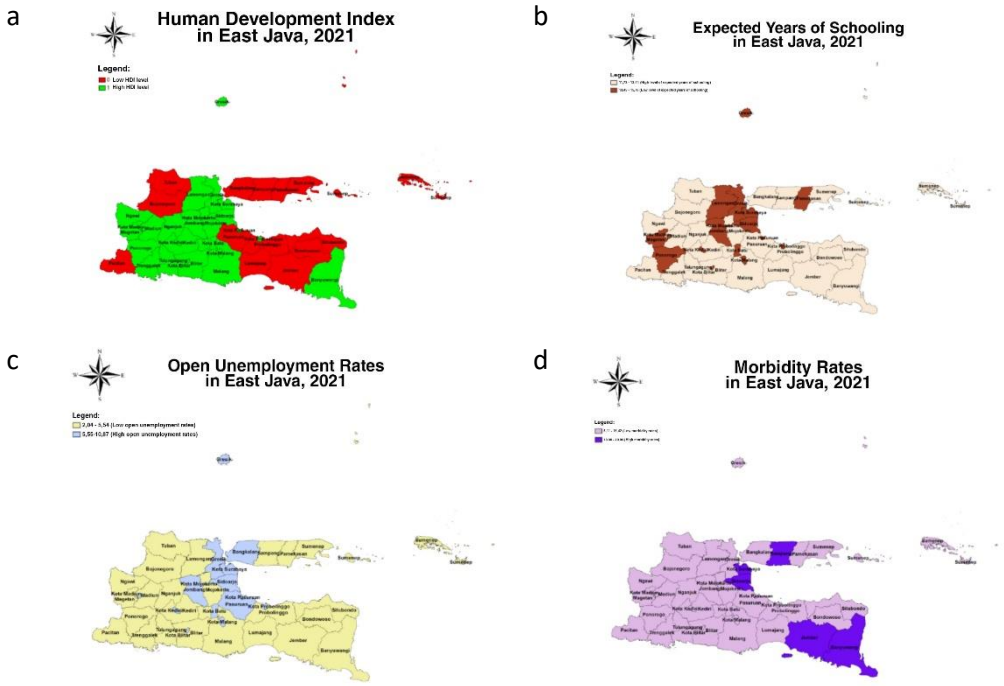


**Figure. 1**. (a) Thematic Map of Human Development Index in East Java; (b) Thematic Map of Expected Years of Schooling in East Java; (c) Thematic Map of Open Unemployment Rates in East Jawa; (d) Thematic Map of Morbidity Rates in East Java.

Before estimating model parameters using spatial regression analysis, the data is subjected to a spatial assumption test. This test is instrumental in assessing whether the data conforms to the fundamental assumptions of spatial regression. A significant level of 10% was applied in this study. In spatial analysis, there could be substantial variations between neighboring locations. By employing a higher significance level, such as 10%, the GWLR model could better capture these discrepancies and offer more precise estimates for each location. Furthermore, given the social context of this research, a 10% significance level is deemed appropriate. The GWLR method analysis entails meeting several assumptions, including the spatial dependency test using Moran's I test and the spatial heterogeneity test using the Breusch-Pagan test.

**Table 3.** Spatial Dependency Test Output

| | |
|---|---|
| Observed | 0.1809297 |
| Expected | -0.02702703 |
| SD | 0.0516518 |
| P-Value | 0.0000056702 |

According to the calculation of Moran's I test, a probability value of 0.000056702 is obtained, which is less than the significant level $\alpha = 10\%$. Therefore, it could be concluded that the null hypothesis $H_0$ is rejected, indicating the presence of spatial dependency in the data. These findings provide evidence that the data has successfully passed the GWLR spatial assumption test, confirming the existence of spatial dependency in the dataset.

**Table 4.** Spatial Heterogeneity Test Output

| | |
|---|---|
| *BP* | 9.6154 |
| *DF* | 3 |
| *P-Value* | 0.02213 |

The Breusch-Pagan test conducted using R software yields a probability value of 0.02213, which is below the significant level of $(\alpha = 10\%)$. Therefore, rejecting the null hypothesis suggests the existence of spatial heterogeneity in the data. This outcome aligns with the GWLR spatial assumption test, indicating spatial heterogeneity in the dataset.

**Table 5.** AIC Value of Each Kernel Weight

| Kernel Fuction Weighting | Bandwith | AIC |
|---|---|---|
| Adaptive Bi-Square | - | - |
| Adaptive Gaussian | - | - |
| Fixed Bi-Square | 1.819 | 28.777 |
| Fixed Gaussian | 0.659 | 28.732 |

Based on the information provided in the table, it is evident that the Kernel Fixed Gaussian weighting function demonstrates the lowest AIC value among the various weighting functions, specifically 28.723, Therefore, the Kernel Fixed Gaussian weighting function is employed to estimate the optimal model in this study. Subsequently, a model suitability test was conducted to assess whether GWLR modeling is more suitable compared to logistic regression modeling.

**Table 6.** GWLR Model Fit Test

| Model | Dev | db | Dev/db | F |
|---|---|---|---|---|
| Logistic Regression | 28.686 | 34 | 0.844 | 2.398 |
| GWLR | 10.103 | 28.690 | 0.352 | |

The analysis reveals that the calculated $F = 2.398$. By considering a significance level of α=10% (0.1), the value is determined to be $F_{(0.1;34;28.690)} = 1.603$. The critical region in this study dictates rejecting $H_0$ when $F > 1.603$, resulting in the rejection of $H_0$. Consequently, a notable difference exists between the logistic regression model and GWLR, indicating the superiority of the GWLR model for modeling purposes.

Following the completion of the analysis, an HDI model was derived for every district/city within East Java Province utilizing the GWLR method. For instance, a specific area in Blitar Regency was selected, resulting in a model equation as presented below:

$$g(x) = -67.957 + 4.007X_1 - 0.059X_2 - 0.169X_3 \quad (27)$$

In Blitar Regency, the partial test results revealed a significant effect of $X_1$ (Expected Years of Schooling). The Odd Ratio calculation was then used to determine the influence of the predictor variables in the model. Notably, only the predictor variables that showed significance in Blitar Regency, such as Expected Years of Schooling, were included in this calculation.

$$Odd\ Ratio = exp\ exp\ (\beta_1) = exp\ exp\ (4.007) = 10.892 \quad (28)$$

In the context of the Odd Ratio calculation, it has been established that a 1 unit increase in Expected Years of Schooling, assuming the other X values remain constant, leads to 10.892 times increase in the Odd value. Moving forward, let's assess the efficiency of the logistic regression model and the GWLR model in the HDI case in East Java.

**Table 7.** Model Goodness Test

| Model | AIC |
|---|---|
| Logistic Regression | 36.686398 |
| GWLR | 28.723450 |

The AIC value in the GWLR model is smaller than that in the logistic regression model, indicating that the GWLR model is more suitable for analyzing HDI data in East Java.

Upon further analysis of the estimated results of the HDI distribution, significant factors influencing the distribution are identified. This allows for a comparison of HDI distribution before and after estimation, shedding light on the factors influencing districts/cities in East Java. The presentation of HDI Index results in graphs and thematic maps provide a visual representation of the findings.

Estimation results for each district/city in East Java were compared by inputting predictor variables into the equation model. For instance, Bojonegoro Regency was selected, showing an anticipated years of schooling of 12.68, an open unemployment rate of 4.82, and a morbidity rate of 11.02, then it is obtained.

$$\pi(X_j) = \frac{exp(\sum_{k=0}^{p} \beta_k(u_i,v_i)x_{ik})}{1+exp(\sum_{k=0}^{p} \beta_k(u_i,v_i)x_{ik})}$$
$$\pi(X_j) = \frac{exp(-67.96+7.66(12.68)- -0.36(4.82)+0.002(11.02)}{1+exp(-0.36(4.82)+0.002\ (11.02)-0.36(4.82)+0.002(11.02)}$$
$$\pi(X_j) = \frac{expexp\ (27.47)}{1+exp(27.47)}$$
$$\pi(X_j) = 1 \quad (29)$$

Manual calculations are employed in every district/city to allow for a comparison between the estimated results and actual observations, as illustrated in the graph below.

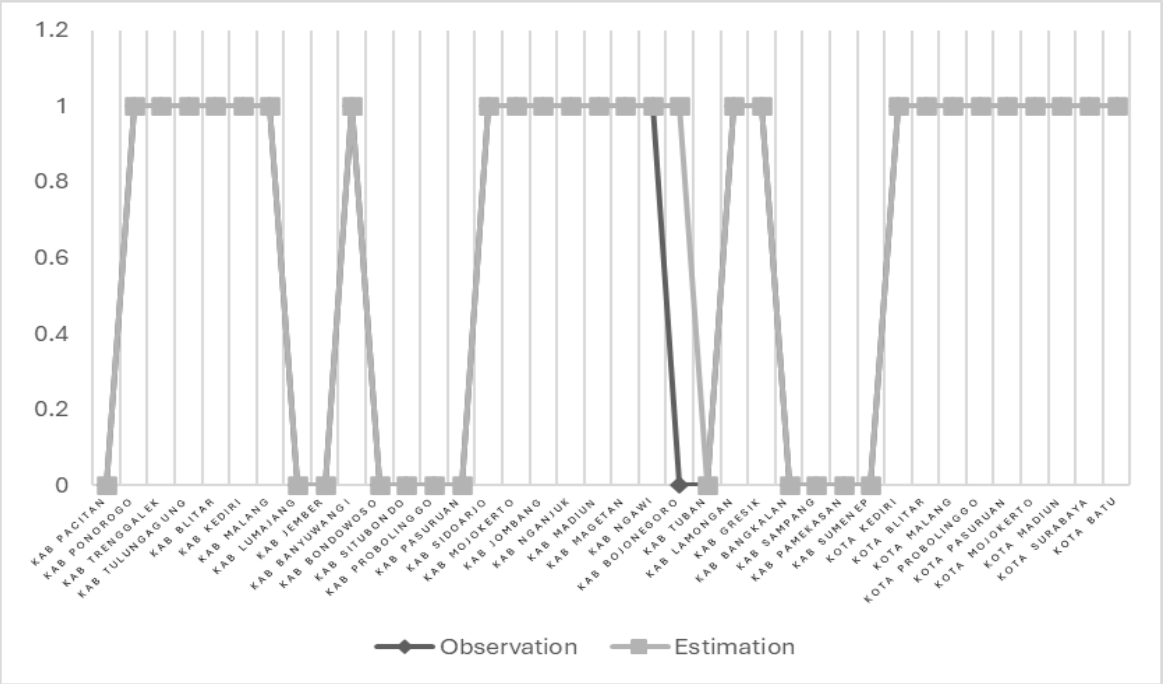


**Figure. 5.** Graph of Comparison of Classification Results with Preliminary Data on the Human Development Index in East Java

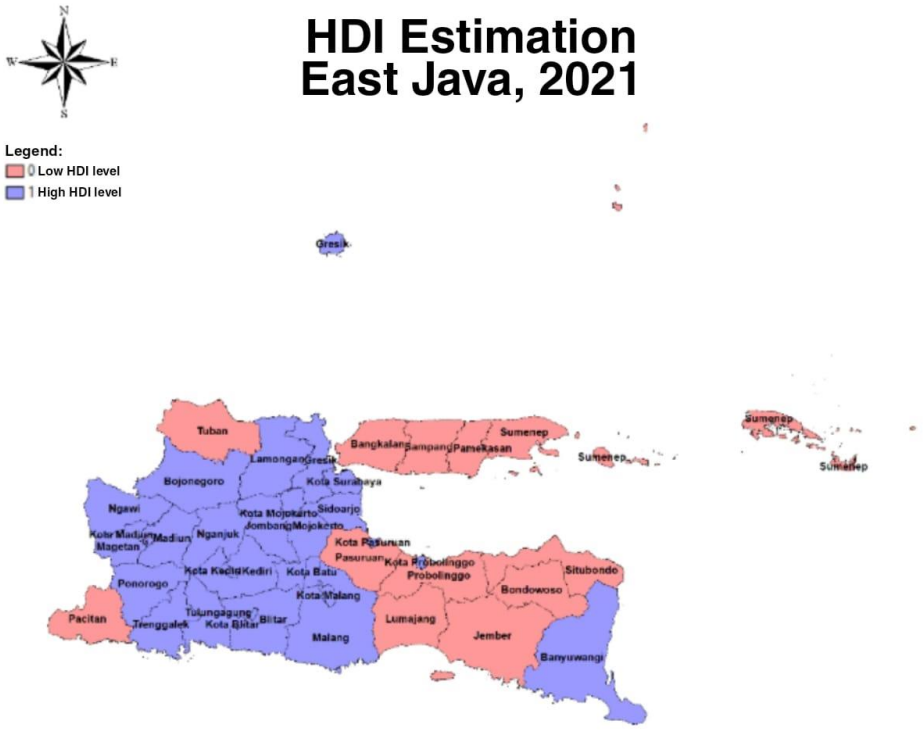


**Figure. 6.** Thematic Map of Human Development Index Estimation Results in East Java

After examining the projected outcomes of the HDI distribution and the factors that impact it, the HDI results in East Java could be categorized using the GWLR method.

**Table 8.** Classification of Human Development Index Results in East Java GWLR Model

| Observation | Estimation | |
|---|---|---|
| | Low (0) | High (1) |
| Low (0) | 12 | 1 |
| High (1) | 0 | 25 |

The estimation results successfully classified 12 regions with low HDI in the appropriate category, but there was one misclassification from low to high

HDI, specifically Bojonegoro Regency. Conversely, the estimation accurately categorized 25 regions with high HDI in the high HDI category, without any incorrect classifications in the low HDI category. As a next step, a thematic map will be developed based on the influential factors affecting HDI in East Java.

**Table 9.** Variables that have a significant influence on HDI

| Significant Variable | Regency/City |
|---|---|
| $X_1$ | Blitar Regency, Kediri Regency, Malang Regency, Probolinggo Regency, Pasuruan Regency, Sidoarjo Regency, Mojokerto Regency, Jombang Regency, Lamongan Regency, Gresik Regency, Bangkalan Regency, Kediri City, Blitar City, Malang City, Pasuruan City, Mojokerto City, Surabaya City, Batu City. |
| $X_2$ | Lumajang Regency |
| $X_3$ | Nothing |

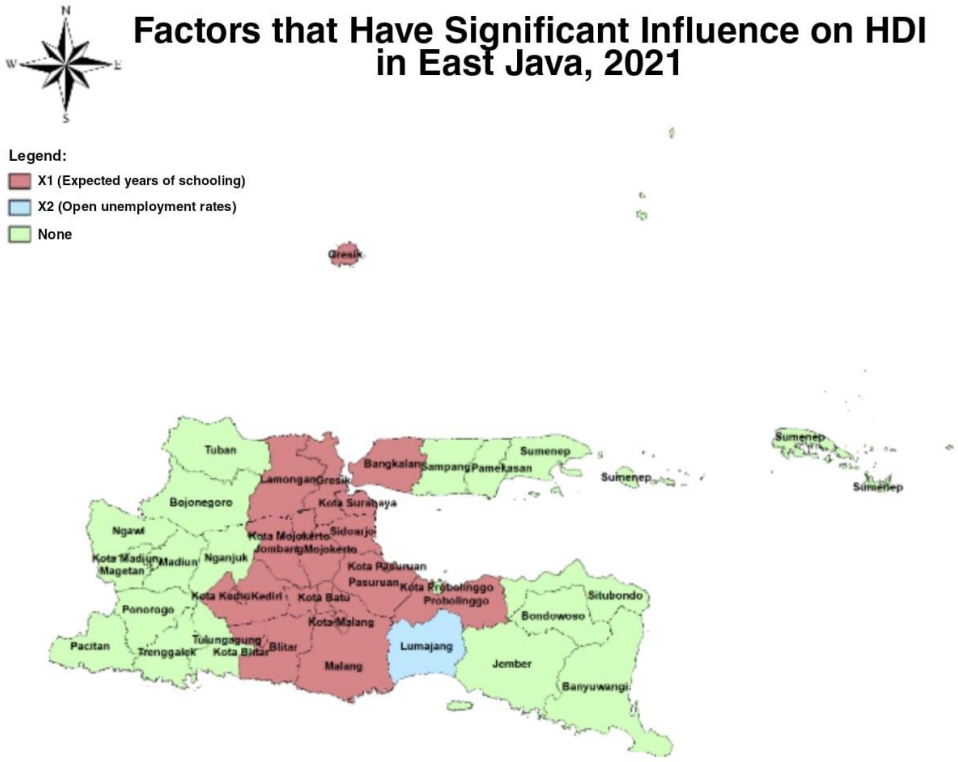


**Figure 7.** Thematic Map of Factors that Significantly Influence the Estimation of HDI in East Java

The provided diagram illustrates the distribution of predictor variables that impact the Human Development Index (HDI) in districts and cities located in East Java. The two influential predictor variables are $X_1$, which represents the expected years of schooling, and $X_2$, which represents the open unemployment rate. Out of all the districts and cities in East Java, 18 of them, accounting for 47.36% of the total, have their HDI value solely influenced by variable $X_1$, indicating the expected years of schooling. Conversely, there is only one district or city, making up 2.63% of the total, where the HDI value is solely influenced by variable $X_2$, representing the open unemployment rate. The figure also demonstrates a tendency for neighboring locations to exhibit similar conditions, with the same predictor variables significantly influencing each other.

The obtained analysis results indicate that the significant variables differ among districts and cities, and in some cases, certain variables may not hold significance at all. This discrepancy could be attributed to the existence of spatial heterogeneity, which reinforces the argument favoring the use of spatial models for effectively describing and analyzing complex phenomena across various regions. Therefore, it is crucial to identify predictor variables that significantly influence the Human Development Index (HDI) in each district and city within East Java Province, to enhance the overall HDI.

## 4. Conclusion

The conclusions that have been derived after conducting analysis and discussion are outlined below:

The year 2021 will see 13 districts/cities in East Java categorized as low HDI, accounting for 34% of the total. Sampang Regency holds the lowest HDI score at 62,8, with Surabaya City leading at 82,31. The analysis suggests that the Fixed Gaussian Kernel function weighting is the optimal model for HDI.

The HDI varies across districts and cities, and the variables that have a significant impact on it differ as well. However, two variables that at least have some influence on the HDI are the expected years of schooling ($X_1$) and the open unemployment rate ($X_2$). According to the GWLR model, the results indicate that districts or cities with higher expected years of schooling tend to have a higher HDI. Conversely, districts or cities with higher levels of open unemployment tend to have a lower HDI.

Based on the discussion, the two districts with the lowest HDI are Sampang Regency and Bangkalan Regency. In Sampang Regency, the HDI is not influenced by the three predictors that were studied. On the other hand, the HDI for Bangkalan Regency is influenced by the Expected Years of Schooling variable ($X_1$), as indicated by the equation $f(x) = -67.957 + 3{,}866X_1 + 0.964X_2 - 0.065X_1$. The higher the expected years of schooling, the higher the HDI value.

The outcomes and deliberations suggest that special attention should be directed towards regencies/cities with low HDI values, particularly in Sampang Regency and Bangkalan Regency. In Bangkalan Regency, the government should focus on outreach and support to enhance education, particularly the duration of schooling. It is crucial for the people of East Java to actively engage in and endorse the government's initiatives to achieve the SDGs 2030, with the overarching aim of advancing society in Indonesia, especially in East Java. Furthermore, readers should investigate other variables apart from the ones analyzed in this research to facilitate the prompt improvement of districts/cities with low HDI figures, such as Sampang Regency, where the factors contributing to the low HDI remain unknown.

## Ethics approval

Not Required

## Acknowledgments

We gratefully acknowledge the support and contributions of all those involved in this research. Special thanks to Badan Pusat Statistik (BPS) for providing the essential data that made this study possible Our thanks also go to the reviewers and proofreaders for their meticulous efforts and invaluable feedback, which greatly enhanced the quality of this work.

## Competing interests

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

## Funding

This study received no external funding.

## Underlying data

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

# References

[1] UNDP, “Sustainable Development Goals”, undp.org. https://www.undp.org/sustainable-development-goals (Accessed May. 5, 2023).

[2] BPS, *New Method Human Development Index*. Jakarta: Central Statistics Agency, 2014.

[3] D. L. Fika, K. Dadan, and N. B. Naomi, “Estimasi Parameter Model Geographically Weighted Logistic Regression”, *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 9, no. 1, pp. 159-164, 2020.

[4] I. M. Nur and M. Al Haris, “Geographically Weighted Logistic Regression (GWLR) with Adaptive Gaussian Weighting Function in Human Development Index (HDI) in The Province of Central Java”, in *Journal of Physics: Conference Series*, 2021, vol. 1776, p. 012048.

[5] W. Lili, Y. Desi, and N. H. Memi, "Pemodelan Faktor-Faktor yang Berpengaruh Tehadap Indeks Pembangunan Manusia (IPM) di Kalimantan dengan Geographically Weighted Logistic Regression (GWLR)", *Jurnal Eksponensial*, vol. 9, no. 1, pp. 67-74, 2018.
[6] A. Siti, 'Peran Badan Usaha Milik Desa (Bumdes) Terhadap Kesejahteraan Masyarakat di Desa Wanasaba Lauk Kecamatan Wanasaba Kabupten Lombok Timur', Universitas_Muhammadiyah_Mataram, 2023.
[7] I. A. Juliannisa, M. B. N. Ariani, and T. Siswantini, 'Efforts to Increase HDI from an Educational Side in the Johar Baru Community, Central Jakarta', *IKRA-ITH ABDIMAS, DKI Jakarta*, 2023.
[8] H. Lubis, 'Analysis of the Effect of Minimum Wage, GRDP, HDI, and Population on Poverty Levels in Districts/Cities of North Sumatra Province 2015-2019', 2023.
[9] A. Tyas and Ikhsani, 'Natural Resources & Human Resources for Indonesia's Economic Development'. 2015.
[10] A. J. Mahya and W. Widowati, 'Analysis of the Influence of Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditures on the Human Development Index', *Prismatics: Journal of Mathematics Education and Research*, vol. 3, no. 2, 2021.
[11] L. Widyastuti, D. Yuniarti, and M. N. Hayati, 'Pemodelan Faktor-Faktor yang Berpengaruh Terhadap Indeks Pembangunan Manusia (IPM) di Kalimantan dengan Geographically Weighted Logistic Regression (GWLR)', *Jurnal Eksponensial*, vol. 9, no. 1, pp. 67–74, 2018.
[12] BPS East Java, *East Java in numbers 2021*. Surabaya: BPS East Java, 2021.
[13] M. N. Faritz and A. Soejoto, 'Pengaruh pertumbuhan ekonomi dan rata-rata lama sekolah terhadap kemiskinan di Provinsi Jawa Tengah', *Jurnal Pendidikan Ekonomi (JUPE)*, vol. 8, no. 1, pp. 15–21, 2020.
[14] Á. S. Batista, *Logistic Regression: An Introduction to Statistical Model with an Example of Revolving Credit*. Lisbon: Createspace Independent Publishing Platform, 2014.
[15] E. N. Manurung and F. Hutabarat, 'The Influence of Expected Years of Schooling, Average Years of Schooling, and Expenditures per Capita on the Human Development Index', *Scientific Journal of Management Accounting*, vol. 4, no. 2, pp. 121–129, 2021.
[16] E. Permadi and E. Chrystanto, 'Analysis of the Influence of Population, Gross Regional Domestic Product (GRDP), and Regency/City Minimum Wage on the Open Unemployment Rate in Regency/City in East Java Province 2012-2018', *OECONOMICUS Journal of Economics*, vol. 5, no. 2, pp. 86–95, Jun. 2021.
[17] D. S. Amru and E. D. Sihaloho, 'The influence of per capita expenditure and health expenditure on morbidity rates in districts/cities throughout Java', *Asian Business and Economics Scientific Journal*, vol. 14, no. 1, pp. 14–25, 2020.
[18] S. Kardjati, *Aspects of Health and Nutrition for Children Under Five*, First. Jakarta: Indonesian Obor Foundation, 1985.
[19] F. Febrianti and H. Helma, 'Binary Logistic Regression Analysis on Factors that Influence the Willingness of the Nagari Paninauan Community to be Vaccinated with COVID-19', *UNP Journal of Mathematics*, vol. 8, no. 1, pp. 36–44, 2023.
[20] I. Arofah and S. Rohimah, 'Analisis Jalur Untuk Pengaruh Angka Harapan Hidup, Harapan Lama Sekolah, Rata-Rata Lama Sekolah Terhadap Indeks Pembangunan Manusia Melalui Pengeluaran Riil Per Kapita Di Provinsi Nusa Tenggara Timur', *Jurnal Saintika Unpam: Jurnal Sains Dan Matematika Unpam*, vol. 2, no. 1, p. 76, 2019.
[21] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.
[22] P. H. M. Alburquerque, F. A. S. Medina, and A. R. Silva, 'Geographically Weighted Logistic Regression Applied to Scoring Model', in *XL ANDAP Congress*, 2016.
[23] F. D. Lestari, D. Kusnandar, and N. N. Debataraja, 'Estimasi Parameter Model Geographically Weighted Logistic Regression', *Mathematics Scientific Bulletin. Stat. and Applications (Bimaster)*, vol. 9, no. 1, pp. 159–164, 2020.