Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

Geographically Weighted Random Forests for Human Development Index of Central Java Prediction

Shaifudin Zuhdi[1], Isna Nurul Fatatik[2], Izlah Nur Fadlila Herawati Prihasno[2], Hasri Akbar Awal Rozaq[3*]

[1]Department of Informatics, Sebelas Maret University, Indonesia
[2]Department of Data Science, Sebelas Maret University, Indonesia
[3]Graduate School of Informatics, Department of Computer Science, Gazi University, Ankara, Türkiye

Email: [1]szuhdi@staff.uns.ac.id

Received : Jul 30, 2025; Revised : Aug 25, 2025; Accepted : Aug 27, 2025; Published : Sep 2, 2025

Abstract

The geographically weighted regression (GWR) model has been widely used in various types of predictions, including human development index predictions. Similarly, the random forests (RF) model has also been widely used in various value predictions. The GWR model always assumes a local linear relationship between dependent and independent variables. The RF model only produces one global model that cannot represent conditions at each location. The GWR model is susceptible to multicollinearity in each independent variable, which can lead to overfitting if multicollinearity in the model is high. To address the vulnerability of the GWR model to multicollinearity, the RF model and the GWR model can be combined. Since the RF model is not vulnerable to multicollinearity in the independent variables, the modification becomes the geographically weighted random forests (GWRF) model to improve the shortcomings of the GWR and RF models. The GWR and GWRF models were constructed using data from districts and cities in Central Java Province, which was selected as the study area due to evident disparities in human development index achievements. These disparities highlight the presence of spatial heterogeneity that conventional models fail to adequately capture. To rigorously evaluate model performance, data from 2023 were employed as training data, while data from 2024 served as testing data. This research introduces a novel integration of spatial econometric and machine learning approaches, providing a more robust framework for addressing complex spatial variations in human development outcomes. The GWRF model is capable of producing a model that does not overfit when there is multicollinearity among independent variables. The GWRF model offers a novel integration of machine learning and spatial modelling, outperforming both GWR and RF by not only delivering high predictive accuracy under complex variable relationships but also capturing nuanced local spatial heterogeneity that conventional approaches fail to address.

Keywords : *GWR, Human Development Index, Random Forests, Spatial Analysis.*

1. INTRODUCTION

The Human Development Index (HDI) is one of the indicators of a country or region's success in implementing human development. Several factors, including population density, the percentage of poor people, and gross regional domestic product, influence the HDI of a region. In addition to these influencing factors, the HDI is also affected by the spatial conditions of each region [1]. The HDI in Sub-Saharan Africa has different factors [2] compared to the factors influencing the HDI in North Sulawesi [3] and the factors influencing the HDI in Eastern Indonesia [4]. Thus, the HDI is also influenced by the geographical location of a region. This indicates that there is a spatial weight that influences the HDI value of a region.

Spatial data is data that contains information about the location or geography of an area based on longitude and latitude values. Modeling methods involving spatial attributes are commonly defined by

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

geostatistics. One method that can be used in spatial analysis is Geographically Weighted Regression (GWR), which is an extension of the ordinary regression model that adds weights by taking spatial influence into account. GWR provides a local approach to regression by emphasizing geographical proximity, which aligns with Tobler's first law of geography. This offers more nuanced insights into spatial data compared to the global models produced by ordinary regression [5]. However, according to Quinones et al., the GWR model has limitations in capturing non-linear effects. Additionally, the GWR model is more prone to overfitting and tends to assume that all coefficients vary spatially without proper variable selection [6], [7].

Machine learning (ML) models have high capabilities for prediction from data mining [8], which is usually flexible and non-linear. In addition, ML has also become a modern data analysis method in recent years. ML has many types of methods, including Support Vector Machine (SVM) [9], Random Forests (RF) [10], and Gradient Boosting (GB) [11]. Several studies have compared the performance of various ML methods, including a study conducted by Appiah-Badu et al., which found that RF outperformed K-Nearest Neighbor (K-NN) in predicting rainfall in Ghana [12]. Nurwatik et al. also noted that the RF model was more effective in modeling landslide risk in Malang, Indonesia, compared to K-NN and Naïve Bayes [13]. Based on the advantages of RF over other ML models, it is necessary to incorporate spatial heterogeneity into the RF model. Thus, a geographically weighted random forests model was developed, in which the RF model coefficients representing global data are divided into local sub-models in accordance with the GWR model approach [6]. The GWRF model has the advantage of not being susceptible to multicollinearity in independent variables [14].

Recent advancements in spatial machine learning have highlighted the potential of hybrid models that combine the advantages of multiple methodologies. The integration of spatial factor optimization techniques with GWRF models has shown promising results in addressing spatial heterogeneity challenges [15]. Research conducted by Li et al. demonstrated that GWRF models significantly outperformed traditional RF and GWR approaches in estimating regional forest carbon density, achieving superior predictive accuracy when combined with remote sensing data [16]. Additionally, studies on macro-level crash frequency prediction have shown that GWRF models are not susceptible to multicollinearity issues while maintaining high prediction accuracy when appropriate bandwidth selection is implemented [17]. The computational efficiency of GWRF has been further enhanced through spatially weighted formulations that improve prediction power while addressing spatial dependence commonly found in geographical data [18], [19]. Furthermore, recent developments in neural network architectures, particularly graph neural networks and neural processes, have been proposed as complementary approaches to traditional spatial modeling, offering enhanced capabilities for handling spatial autocorrelation and prediction uncertainty [20].

The application domain of GWRF continues to expand beyond traditional environmental modeling to encompass diverse fields including public health, urban planning, and economic development. Contemporary research has identified key challenges in data-driven geospatial modeling, including issues related to imbalanced data, spatial autocorrelation, and model generalization, which GWRF models are uniquely positioned to address [21]. Recent applications of machine learning techniques, including GWRF, to human development index estimation have demonstrated the capability to provide high-resolution spatial estimates that reveal significant population misclassification due to aggregation bias [22]. Moreover, hybrid approaches combining GWRF with convolutional neural networks have been developed to model spatial heterogeneity more effectively, integrating global autocorrelation analysis through Moran's I statistics [23]. Advanced geographically weighted implementations, including spatiotemporal proximity neural networks and geographically weighted versions of principal component analysis, have emerged to handle complex nonlinear space-time interactions [24]. The development of specialized geospatial random forest variants, such as GeoRF, has

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

introduced novel bivariate splits designed specifically for geographic coordinates, demonstrating superior performance over traditional spatial statistical models [25], [26]. This study employs data from districts and cities in Central Java Province, chosen due to the pronounced disparities in human development index distribution between major urban centers and surrounding rural districts. Such disparities underscore the existence of spatial inequalities that conventional approaches often overlook, thereby providing a compelling rationale for adopting advanced spatial–machine learning methodologies [27].

2. METHOD

This research was conducted in several stages: (1) collecting datasets from the central statistics agency, (2) performing GWR modeling, (3) performing GWRF modeling, and (4) evaluating the GWRF model. The following is an overview of the research process.

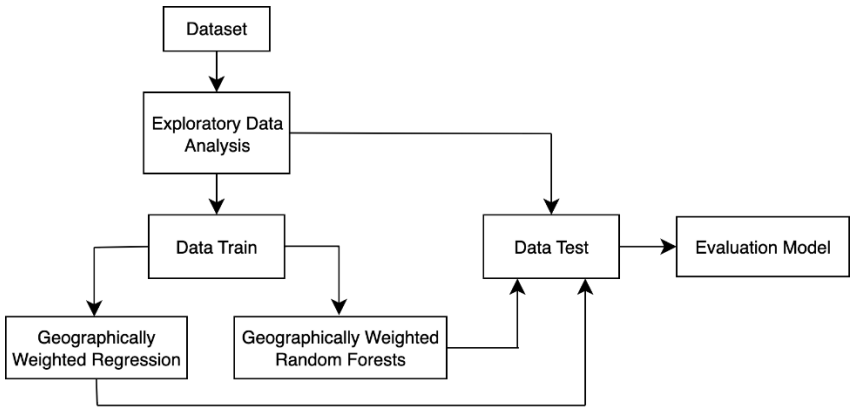


Figure 1. Research Process

2.1. Dataset

The dataset is sourced from the Central Bureau of Statistics, consisting of independent variables [28]: life expectancy (LE), expected years of schooling (EYS), average years of schooling (AYS), and adjusted per capita expenditure. The dependent variable is the Human Development Index (HDI) value. The variables LE, EYS, AYS, and adjusted per capita expenditure were chosen because they are the main dimensions in measuring a country's performance in human development. Since the GWRF model considers spatial effects, latitude and longitude data are used. The dataset used consists of data from the districts and cities in Central Java, totaling 35 observation locations. The dataset used is from the years 2023 and 2024, where the 2023 data is used as training data for the model and the 2024 data is used as testing data for the model.

Table 1. Sample of the Dataset

City	LE	EYS	AYS	Outcome (IDR)	HDI	Latitude	Longitude
Cilacap	74.25	12.67	7.39	11432	71.83	-7.73333	109
Banyumas	73.98	13.26	7.87	12492	73.86	-7.48321	109.14
Purbalingga	73.37	12.02	7.34	10964	70.24	-7.39075	109.364
Banjarnegara	74.47	11.82	6.86	10226	69.14	-7.40271	109.681
Kebumen	73.83	13.37	7.86	9734	71.37	-7.67868	109.657

Table 1 shows a sample of the dataset for 2023 that will be used as training data for the model. The dataset used consists of LE, EYS, AYS, outcome, HDI, latitude, and longitude from 29 districts and 6 cities in Central Java Province, totaling 35 observation locations. Each city has important variables

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

related to human development and geographical characteristics. Health and education variables show significant variation across cities. LE ranges from 69.96 years in Brebes to 77.93 years in Salatiga City. For education, the EYS ranges from 11.8 in Wonosobo to 15.55 in Semarang City, while the AYS varies from 6.4 in Brebes to 11.24 in Salatiga City. The adjusted per capita expenditure ranges from 9587 in Pemalang to 16650 in Salatiga City.

The Human Development Index (HDI) as a composite indicator shows that Salatiga City has the highest value (84.99), while Brebes has the lowest value (67.95). The latitude and longitude variables will be used to calculate the spatial weights of each observation location.

2.2. Multicollinearity Test

Multicollinearity testing is conducted to determine whether there is a correlation between independent variables in a regression model [29]. The presence or absence of multicollinearity in a regression model can be determined from the variance inflation factor (VIF) value. VIF is calculated using a formula.

$$\boldsymbol{VIF = \frac{1}{(1-R^2)}} \quad (1)$$

Interpreting VIF values is very important in detecting multicollinearity issues. The lower the tolerance value, the higher the likelihood of serious multicollinearity in the model. A VIF value of 1 indicates an ideal condition where there is no correlation between independent variables. When the VIF value is less than 5, this indicates that the correlation between independent variables is still within acceptable or moderate limits.

However, when the VIF value reaches 5 or higher, this is a warning signal that there is significant multicollinearity between the predictor variables in the regression model. This high multicollinearity can cause regression coefficients to become unstable and difficult to interpret correctly. Therefore, VIF serves as a very useful diagnostic tool for researchers to identify and measure the severity of multicollinearity issues in multiple regression analysis, so that necessary corrective measures can be taken to improve the quality of the regression model [30].

2.3. Geographically Weighted Regression Model

The GWR model accommodates different relationships that occur at various points in space [31]. It allows researchers to explore the relationship between independent and dependent variables that shift from one location to another. The mathematical equation for the GWR model is as follows.

$$y_i = \beta_{0i}(u_i, v_i) + \sum_{n=1}^{k} \beta_{ni}(u_i, v_i)x_{ni} + \varepsilon_i \quad (2)$$

y_i shows the HDI value at observation location-i, β_0 is the intercept parameter value that is also possessed by each observation location-i with different value. β_n is the parameter for n independent variables at each observation location-i.

2.4. Random Forests Model

The RF model is an ensemble method that combines the ideas of bagging (bootstrap aggregation) and randomly selecting subspaces. RF also originates from a collection of decision trees formed using the bagging method [32]. This method uses several decision trees to train different subsets of data. After each decision tree with its respective subset of data has formed a model, it will be used for prediction. Then, the prediction results from each decision tree are combined as a whole to improve accuracy. In regression models, this aggregation is the average of all prediction trees, which can be denoted as.

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (3)$$

With $\hat{f}(x)$ being the prediction result from tree-t, and T being the number of trees in the forest.

2.5. Human Development Index

The Human Development Index (HDI) is a composite measure that assesses the average level of achievement in a region based on several factors. The basic elements of human development include life expectancy, education, and a decent standard of living [33]. These three basic elements have very broad meanings because they are related to many factors. Life expectancy at birth is used to measure the health dimension. A combination of indicators for expected years of schooling and average years of schooling is used to measure the knowledge dimension. The purchasing power indicator is used to measure the decent standard of living dimension.

2.6. Geographically Weighted Random Forests

The data shows spatial heterogeneity, meaning that the relationship between the independent and dependent variables differs for each observation location. The standard RF model cannot accommodate differences in the relationship between variables at each observation location. Therefore, the GWRF model was developed to form an RF model that can accommodate local models from each observation location; in other words, there is an RF model for each observation location. There is an equation in the RF mode.

$$Y_i = ax_i + e, i = 1, \dots, n \quad (4)$$

Which defines the dependent variable at location-i expanded to obtain the equation.

$$Y_i = a_1 x_{i1} + a_2 x_{i2} + e, i = 1,2, \dots, n \quad (5)$$

The global GWRF model and the local GWRF model equations are as follows.

$$Y_i = a_1(u_i, v_i)x_i + e, i = 1, \dots, n \quad (6)$$

Where $a_1(u_i, v_i)$ denotes the calibrated RF model prediction at observation location-i. Here (u_i, v_i) represent the coordinates or values of latitude and longitude. Each observation location will be built into a submodel by only considering the observations in its vicinity. The area used in the submodel is the neighborhood or kernel size, and the maximum distance at which a location influences other locations around it is called the bandwidth. To determine the bandwidth, there are two types of kernels: adaptive and fixed. One kernel that can be used is the fixed bisquare kernel, which has the following mathematical equation.

$$w_{ij} = \left(1 - \left(\frac{d_{ij}}{b}\right)^2\right)^2 \quad (7)$$

For $d_{ij} < b$ and $w_{ij} = 0$ for other conditions. Where w_{ij} s the weight for each observation location, d_{ij} is the distance between location-i and location-j, and b is the bandwidth value.

2.7. Mean Absolute Error

Mean Absolute Error (MAE) is one of the most important evaluation metrics for assessing the performance of predictive models, particularly in the context of research comparing Geographically Weighted Regression (GWR) and Geographically Weighted Random Forests (GWRF) models for predicting the Human Development Index. MAE is an ideal metric for measuring prediction accuracy

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

because it provides a clear picture of the average absolute error between predicted values and actual HDI values at each district and city location in Central Java Province. In the context of spatial models like GWR and GWRF, which produce different predictions for each geographic location, MAE helps evaluate how well the model captures local variations and addresses issues such as multicollinearity and overfitting. Here is the formula for MAE [34].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

The MAE formula works by calculating the difference between each actual value (yi) and the predicted value (ŷi), then taking the absolute value to eliminate negative signs. The use of absolute values is important because it ensures that prediction errors above and below the actual value are treated equally and do not cancel each other out. After all absolute differences are calculated, the values are summed (Σ) and divided by the number of samples (n) to obtain the average error. The final MAE result shows the average magnitude of prediction errors in the same units as the target variable, so if the target is in rupiah, then the MAE will also be in rupiah, making it very easy to interpret.

3. RESULT

3.1. Exploratory Data Analysis

The dataset used has 35 observation locations, namely districts and cities in Central Java province. Each observation location has 4 independent variables and 1 dependent variable, namely HDI. The following is the distribution of HDI from districts and cities in Central Java in 2023, which is used as training data.

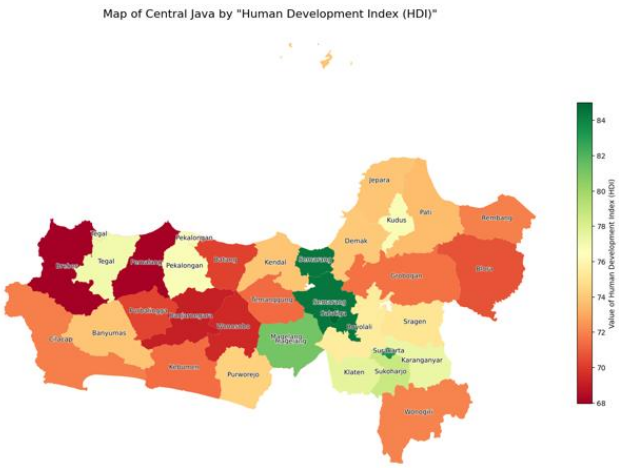


Figure 2. Distribution of Human Development Index in Regencies and Cities in Central Java

Figure 2 shows the distribution of HDI values for 29 districts and 5 cities in Central Java in 2023. Regions that tend to have a high HDI are major cities with easy access to education. Central Java Province is mostly dominated by areas with an HDI below 80, namely regions classified as regencies. This indicates a development disparity between cities and districts.

Figure 3 shows the distribution of independent variables used as training models for GWR and GWRF. The life expectancy variable in regencies and cities in Central Java predominantly has high values. The expected length of schooling variable in regencies and cities in Central Java predominantly has moderate values. Similarly, the average length of schooling and adjusted per capita expenditure variables predominantly have moderate values.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

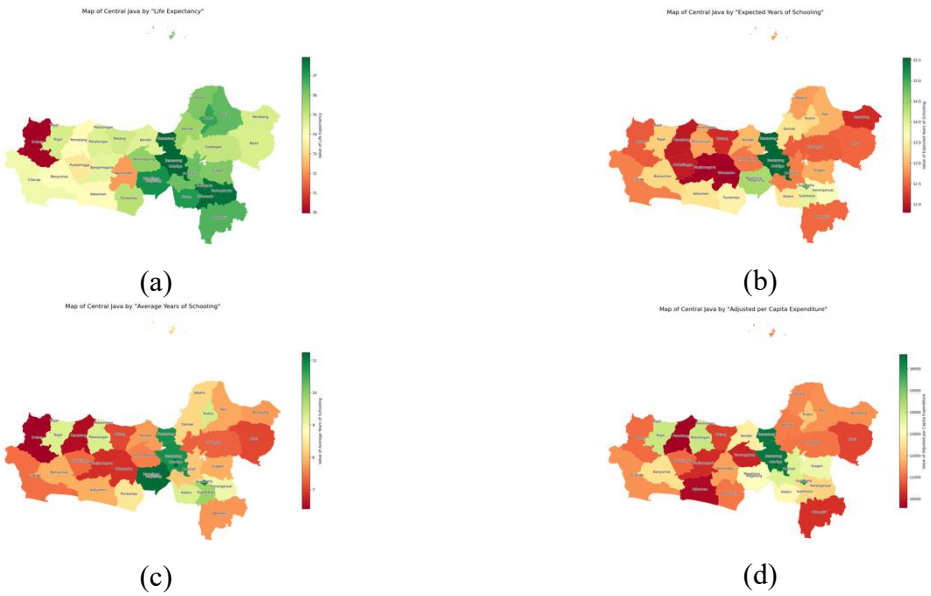


Figure 3. Distribution of Independent Variables of Life Expectancy (a), Expected Years of Schooling (b), Average Years of Schooling (c), and Adjusted per Capita Expenditure (d)

3.2. Correlation Test

The following is a heatmap showing the correlation values between independent variables.

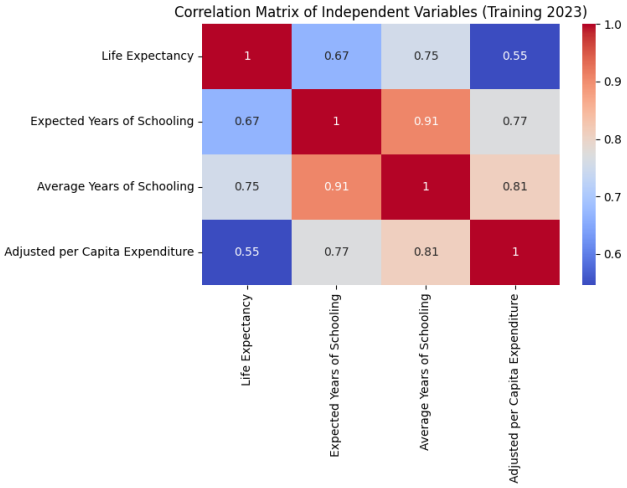


Figure 4. Correlation Results between Independent Variables

Figure 4 shows that there is a positive correlation between the independent variables. The positive correlation tends to be moderate to strong. The correlation between the variables of expected length of schooling and average length of schooling is very strong and has the potential for multicollinearity. Therefore, it is necessary to perform a multicollinearity test using VIF.

Table 2. Multicollinearity Test

Variables	VIF
Life Expectancy	501,9849
Expected Years of Schooling	1052,8093
Average Years of Schooling	208,2365
Adjusted per Capita Expenditure	140,1924

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

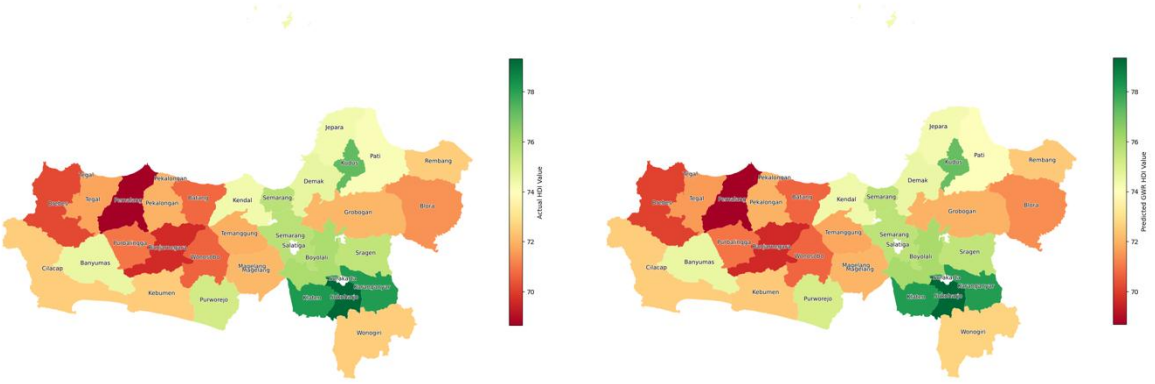
Table 2 shows that the VIF values of each independent variable are very high, indicating that there is very strong multicollinearity between the independent variables.

3.3. GWR Model Results

The 2023 dataset was used to train the GWR model. The GWR model obtained from the training data was then used to predict the 2024 dataset and as testing data. Figure 5 shows that the 2024 IPM prediction results using the GWR model have the same distribution as the actual IPM values. The GWR model has a model evaluation in Table 3, which shows that the GWR model has a very large R2 value. Referring to the results of the multicollinearity test, which shows multicollinearity in each independent variable, the GWR model obtained has the potential to be an overfitting model and may result in a significant number of false positives in the local model.

Table 3. Results of the GWR Model Evaluation

Evaluation	Value
MAE	0,0596
RMSE	0,0796
R^2	0,9997



(a) (b)

Figure 5. Distribution of HDI Values Actual in 2024 (a) and Predicted in 2024 (b)

Figure 6 shows that the GWR model produced IPM prediction values that were very similar to the original values, a condition commonly referred to as overfitting. This occurred because there was very high multicollinearity between the independent variables.

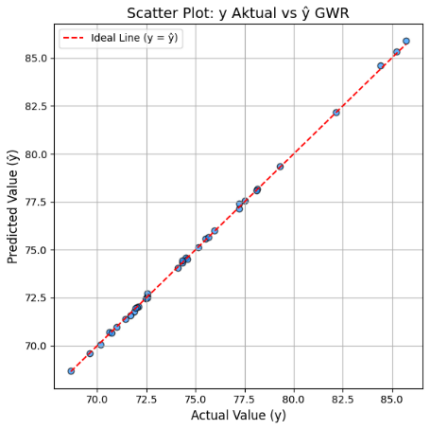
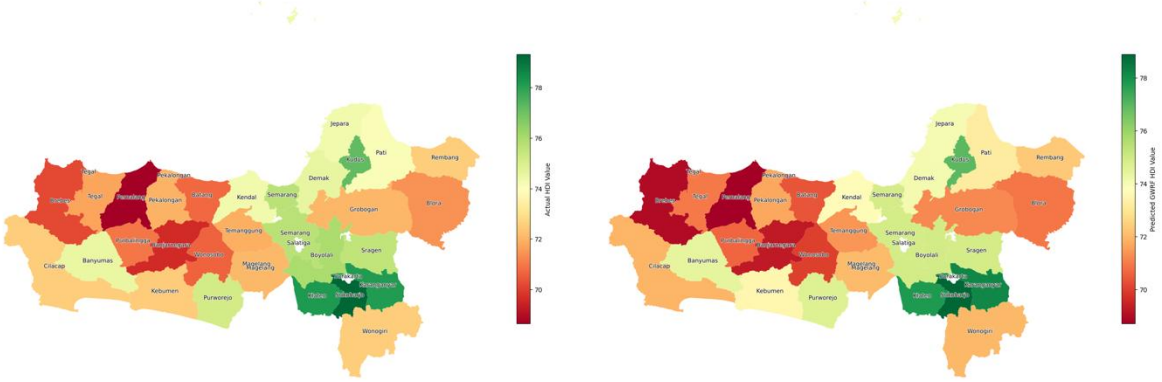


Figure 6. Comparison of Actual HDI Data for 2024 and GWR Model Predictions

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

3.4. Model GWRF

The GWRF model was constructed using a 2023 dataset consisting of four independent variables and one dependent variable. The model was constructed using a fixed bisquare kernel, which was also used when constructing the GWR model. Figure 7 shows the distribution of HDI values for districts and cities in Central Java for prediction results using the GWRF model, which has the same distribution value as the actual data. Figure 8 shows that the 2024 HDI prediction results using the GWRF model do not overfit as occurs in the GWR model. This indicates that the GWRF model is capable of overcoming multicollinearity conditions that occur between independent variables. The GWRF model produces parameters that generalize well to data not used for training. In regions with relatively low HDI values, there is a slightly larger deviation between the actual and predicted values, indicating the influence of other variables not accounted for in the model.



(a) (b)

Figure 7. Distribution of Actual HDI Values (a) and Predicted Values (b) for 2024

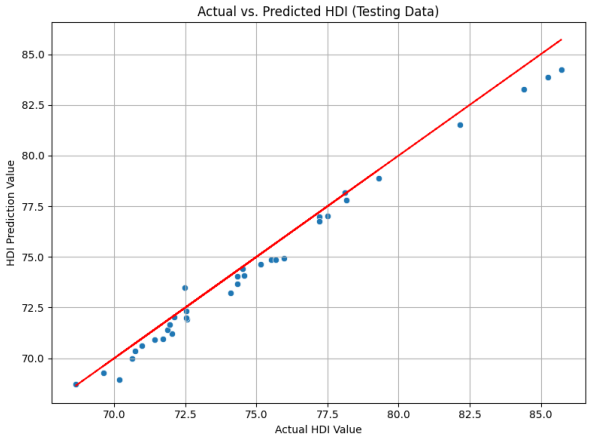


Figure 8. Comparison of Actual Values and GWRF Prediction Results

4. DISCUSSIONS

Data on the influence of independent variables such as life expectancy, expected years of schooling, average years of schooling, and per capita expenditure, adjusted for the dependent variable of HDI for districts and cities in Central Java, show high multicollinearity in each independent variable. The GWR model is able to capture the spatial heterogeneity of the relationship between independent and dependent variables by location. This local model is obtained by changing the coefficients of each independent variable at different locations. Although the GWR model can capture spatial heterogeneity, it is susceptible to multicollinearity issues. Fotheringham and Oshan (2016) show that multicollinearity

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

is not a problem in the GWR model when the sample size is relatively large [35]. Since this study uses a relatively small sample, the solution is to add a machine learning method to overcome the multicollinearity problem.

The selected machine learning method is RF, as RF can address the weaknesses of multicollinearity and non-linearity in global models [36], [37]. The RF method is combined with the GWR model, which is subsequently referred to as the GWRF model. The GWRF model addresses the shortcomings of the GWR model by integrating spatial and data-driven elements. To provide a more comprehensive perspective on the performance of the model used, a comparison of R^2 values is conducted with similar studies that apply GWR and GWRF methodologies in the Indonesian context, particularly to validate the predictive power of the model in this study.

Table 4. Comparative Study

Author	Used Method	Results (R^2)
Kurniati (2022) [38]	GWR (Geographically Weighted Regression)	0.31
Dewi et al. (2024) [39]	RFR (Random Forest Regression)	0.82
Proposed Method	GWRF (Geographically Weighted Random Forest)	0.95
Proposed Method	GWR (Geographically Weighted Regression)	0.99

The comparison results show that the GWR model in this study, with an R^2 of 0.99, performs far better than the study of crime in East Java, which only achieved an R^2 of 0.31. This significant difference can be explained by the different characteristics of the data, where IPM data has a more structured and stable spatial pattern compared to crime data, which tends to be more volatile and influenced by complex social factors. However, the very high R^2 value (0.99) in the GWR model indicates overfitting caused by multicollinearity among the independent variables. This condition occurs when the predictor variables have high correlations with each other, causing the model to become overly sensitive to the training data and lose its generalization ability.

Meanwhile, the GWRF model with an R^2 of 0.95 showed very competitive performance, even surpassing the conventional Random Forest Regression (RFR), which reached 0.82 in a study of stunting in East Java. This confirms the superiority of the hybrid approach that integrates machine learning with geographic weighting. Although there is a 4% decrease in R^2 from GWR to GWRF, this trade-off is advantageous because GWRF can overcome multicollinearity issues through the feature selection mechanism inherent in the Random Forest algorithm. The bootstrap aggregating and random feature sampling techniques in GWRF effectively reduce the impact of multicollinearity by selecting the optimal subset of variables in each decision tree, resulting in a more robust model with better generalization capabilities. Thus, GWRF is a more optimal choice than conventional GWR because it not only provides high prediction accuracy but also avoids the overfitting issues that are prone to occur in GWR when there is multicollinearity among independent variables. The stability and reliability of the GWRF model make it more suitable for practical applications in spatial analysis. Based on the results of this study, the GWRF model is suitable for addressing poverty in regencies and cities in Central Java, as well as in broader areas such as all provinces in Indonesia. Through spatial modeling, the mapping of areas with low to high indicator values will be clearer, and the distribution of the data will be easier to understand.

5. CONCLUSION

In this study, each independent variable has a high correlation with the other independent variables. After conducting a multicollinearity test, it was found that each independent variable had a high VIF value, which means that each independent variable had a multicollinearity problem. When these data were modeled using the GWR model, there was a risk of overfitting. The GWR modeling

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

results yielded a very high R2 value of 0.9997, and a visualization of the comparison between the GWR model predictions and the actual data showed that the model was overfitting. The multicollinearity problem was overcome by adding an RF model to the GWR model, resulting in an RF model that could capture spatial heterogeneity. After the data was modeled using GWRF, a lower R^2 value was obtained compared to the GWR model, namely 0.9769. Based on the scatter plot comparing the GWRF model's prediction results with actual data, it was shown that the model did not overfit. The GWRF model can address the multicollinearity issue experienced by the independent variables in the GWR model with a small sample size. The GWRF model is capable of providing good prediction results. By using 2023 data to build the model and then inputting the independent variables for 2024 to generate prediction values for 2024, the model can produce accurate HDI predictions with an R^2 value of 0.95. Based on this study, when performing spatial modeling with the GWR model on data containing multicollinearity, the issue can be addressed by combining the GWR model with machine learning, such as RF. In this way, the weaknesses of the GWR model can be compensated for by the RF model, and vice versa. This contributes to the diversity of knowledge related to spatial models commonly used in research concerning the Sustainable Development Goals (SDGs).

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest between the authors or with the research object in this paper.

ACKNOWLEDGEMENT

The author would like to express his deepest gratitude to Sebelas Maret University for providing research funding through the Research Grant Group scheme as stated in contract No. 371/UN27.22/PT.01.03/2025. The author also wishes to express gratitude for the support, facilities, academic environment, encouragement, and invaluable resources from Faculty of Information Technology and Data Science provided throughout the conduct of this research.

REFERENCES

[1] X. Lian, Z. Fu, and J. Chen, "Analysis of spatial differences in global regional human development index under planetary pressure and decomposition study of driving factors," *J. Environ. Manage.*, vol. 348, p. 119292, Dec. 2023, doi: 10.1016/j.jenvman.2023.119292.

[2] F. Y. Meilita and M. I. Hasmarini, "Analysis of Factors Affecting the Human Development Index (HDI) in 43 Sub-Saharan African Countries 2018-2022," *J. Ekon. Balanc.*, vol. 20, no. 2, pp. 143–152, Dec. 2024, doi: 10.26618/jeb.v20i2.15474.

[3] J. J. E. Ganda and L. Yola, "Spatial Empirical Analysis on Urban Dwellers' Human Development Index in North Sulawesi, Indonesia," in *Advances in Civil Engineering Materials*, 2023, pp. 465–471, doi: 10.1007/978-981-19-8024-4_40.

[4] F. D. A. Putri, S. Suhendro, and P. Nauli, "Analysis of factors affecting the level of the human development index," *Asian J. Econ. Bus. Manag.*, vol. 1, no. 3, pp. 218–228, Nov. 2022, doi: 10.53402/ajebm.v1i3.229.

[5] M. N. Lessani and Z. Li, "SGWR: similarity and geographically weighted regression," *Int. J. Geogr. Inf. Sci.*, vol. 38, no. 7, pp. 1232–1255, Jul. 2024, doi: 10.1080/13658816.2024.2342319.

[6] S. Georganos *et al.*, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto Int.*, vol. 36, no. 2, pp. 121–136, Jan. 2021, doi: 10.1080/10106049.2019.1595177.

[7] S. Georganos and S. Kalogirou, "A Forest of Forests: A Spatially Weighted and Computationally Efficient Formulation of Geographical Random Forests," *ISPRS Int. J. Geo-Information*, vol. 11, no. 9, p. 471, Aug. 2022, doi: 10.3390/ijgi11090471.

[8] H. Wiemer, L. Drowatzky, and S. Ihlenfeldt, "Applied Aciences Data Mining Methodology for Engineering Applications (DMME)— A Holistic Extension," *Appl. Sci.*, 2019.

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

[9] G. E. I. Selim, E. E. D. Hemdan, A. M. Shehata, and N. A. El-Fishawy, "Anomaly events classification and detection system in critical industrial internet of things infrastructure using machine learning algorithms," *Multimed. Tools Appl.*, vol. 80, no. 8, pp. 12619–12640, 2021, doi: 10.1007/s11042-020-10354-1.

[10] H. Jafarzadeh, M. Mahdianpari, E. Gill, F. Mohammadimanesh, and S. Homayouni, "Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation," *Remote Sens.*, vol. 13, no. 21, p. 4405, Nov. 2021, doi: 10.3390/rs13214405.

[11] E. Y. K. Ng and J. T. Lim, "Machine Learning on Fault Diagnosis in Wind Turbines," *Fluids*, vol. 7, no. 12, 2022, doi: 10.3390/fluids7120371.

[12] N. K. A. Appiah-Badu, Y. M. Missah, L. K. Amekudzi, N. Ussiph, T. Frimpong, and E. Ahene, "Rainfall Prediction Using Machine Learning Algorithms for the Various Ecological Zones of Ghana," *IEEE Access*, vol. 10, pp. 5069–5082, 2022, doi: 10.1109/ACCESS.2021.3139312.

[13] N. Nurwatik, M. H. Ummah, A. B. Cahyono, M. R. Darminto, and J.-H. Hong, "A Comparison Study of Landslide Susceptibility Spatial Modeling Using Machine Learning," *ISPRS Int. J. Geo-Information*, vol. 11, no. 12, p. 602, Dec. 2022, doi: 10.3390/ijgi11120602.

[14] D. Wu, Y. Zhang, and Q. Xiang, "Geographically weighted random forests for macro-level crash frequency prediction," *Accid. Anal. Prev.*, vol. 194, p. 107370, Jan. 2024, doi: 10.1016/j.aap.2023.107370.

[15] K. Kopczewska, "Spatial machine learning: new opportunities for regional science," *Ann. Reg. Sci.*, vol. 68, no. 3, pp. 713–755, Jun. 2022, doi: 10.1007/s00168-021-01101-x.

[16] Y. Zhou *et al.*, "Estimating Regional Forest Carbon Density Using Remote Sensing and Geographically Weighted Random Forest Models: A Case Study of Mid- to High-Latitude Forests in China," *Forests*, vol. 16, no. 1, p. 96, Jan. 2025, doi: 10.3390/f16010096.

[17] S. Bhattacharya *et al.*, "Correlation between visuo-cognitive tests and simulator performance of commercial drivers in the United States," *Accid. Anal. Prev.*, vol. 184, p. 106994, May 2023, doi: 10.1016/j.aap.2023.106994.

[18] Z. Li, "GeoShapley: A Game Theory Approach to Measuring Spatial Effects in Machine Learning Models," *Ann. Am. Assoc. Geogr.*, vol. 114, no. 7, pp. 1365–1385, Aug. 2024, doi: 10.1080/24694452.2024.2350982.

[19] Y. Li *et al.*, "STAR: A First-Ever Dataset and a Large-Scale Benchmark for Scene Graph Generation in Large-Size Satellite Imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1832–1849, Mar. 2025, doi: 10.1109/TPAMI.2024.3508072.

[20] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, "Nbias: A natural language processing framework for BIAS identification in text," *Expert Syst. Appl.*, vol. 237, p. 121542, Mar. 2024, doi: 10.1016/j.eswa.2023.121542.

[21] D. Koldasbayeva, P. Tregubova, M. Gasanov, A. Zaytsev, A. Petrovskaia, and E. Burnaev, "Challenges in data-driven geospatial modeling for environmental research and practice," *Nat. Commun.*, vol. 15, no. 1, p. 10700, Dec. 2024, doi: 10.1038/s41467-024-55240-8.

[22] L. Sherman, J. Proctor, H. Druckenmiller, H. Tapia, and S. Hsiang, "Global High-Resolution Estimates of the United Nations Human Development Index Using Satellite Imagery and Machine-learning," Cambridge, MA, Mar. 2023. doi: 10.3386/w31044.

[23] M. D. Ogah, J. Essien, M. Ogharandukun, and M. Abdullahi, "Machine Learning Models for Heterogenous Network Security Anomaly Detection," *J. Comput. Commun.*, vol. 12, no. 06, pp. 38–58, 2024, doi: 10.4236/jcc.2024.126004.

[24] B. Nikparvar and J.-C. Thill, "Machine Learning of Spatial Data," *ISPRS Int. J. Geo-Information*, vol. 10, no. 9, p. 600, Sep. 2021, doi: 10.3390/ijgi10090600.

[25] M. Geerts, S. vanden Broucke, and J. De Weerdt, "GeoRF: a geospatial random forest," *Data Min. Knowl. Discov.*, vol. 38, no. 6, pp. 3414–3448, Nov. 2024, doi: 10.1007/s10618-024-01046-7.

[26] F. Lu, G. Zhang, T. Wang, Y. Ye, and Q. Zhao, "Geographically Weighted Random Forest Based on Spatial Factor Optimization for the Assessment of Landslide Susceptibility," *Remote Sens.*, vol. 17, no. 9, p. 1608, May 2025, doi: 10.3390/rs17091608.

[27] S. Yulianti, Y. Widyanigsih, and S. Nurrohmah, "Spatial panel data model on human

Vol. 6, No. 4, August 2025, Page. 2756-2768
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.4.5204

development index at Central Java," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, p. 012090, Jan. 2021, doi: 10.1088/1742-6596/1722/1/012090.

[28] C. B. of Statistics, "IPM menurut Kabupaten Kota di Jawa Tengah," *Central Bureau of Statistics*, 2024. jateng.bps.go.id (accessed Jul. 21, 2025).

[29] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[30] D. A. Belsley, *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: John Wiley \& Sons, 1991.

[31] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," *Geogr. Anal.*, vol. 28, no. 4, pp. 281–298, Oct. 1996, doi: 10.1111/j.1538-4632.1996.tb00936.x.

[32] I. M. Putra, I. Tahyudin, H. A. A. Rozaq, A. Y. Syafa'At, R. Wahyudi, and E. Winarto, "Classification analysis of COVID19 patient data at government hospital of banyumas using machine learning," in *2021 2nd International Conference on Smart Computing and Electronic Enterprise: Ubiquitous, Adaptive, and Sustainable Computing Solutions for New Normal, ICSCEE 2021*, Jun. 2021, pp. 271–274, doi: 10.1109/ICSCEE50312.2021.9498020.

[33] B. P. Statistik, "Indeks Pembangunan Manusia," in *Badan Pusat Statistik*, Jakarta, 2011.

[34] N. H. Kim, S. G. Yu, S. E. Kim, and E. C. Lee, "Non-contact oxygen saturation measurement using ycgcr color space with an rgb camera," *Sensors*, vol. 21, no. 18, 2021, doi: 10.3390/s21186120.

[35] A. S. Fotheringham and T. M. Oshan, "Geographically weighted regression and multicollinearity: dispelling the myth," *J. Geogr. Syst.*, vol. 18, no. 4, pp. 303–329, Oct. 2016, doi: 10.1007/s10109-016-0239-5.

[36] S. Quiñones, A. Goyal, and Z. U. Ahmed, "Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA," *Sci. Rep.*, vol. 11, no. 1, p. 6955, Mar. 2021, doi: 10.1038/s41598-021-85381-5.

[37] S. S. Gokhale, V. Lebakula, and A. Peluso, "Explaining Health Risk Behaviors in the U.S. with Social Deprivation at Local and Regional Levels," in *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, Jul. 2024, pp. 1856–1864, doi: 10.1109/COMPSAC61105.2024.00294.

[38] B. Kurniati, "Perbandingan Metode Geographically Weighted Regression dan Geographically Weighted Random Forest pada Kasus Kriminalitas di Jawa Timur," Jember University, 2022.

[39] Y. S. Dewi, S. Hastuti, and M. Fatekurohman, "Analysis of stunting in East Java, Indonesia using random forest and geographically weighted random forest regression," *Brazilian J. Biometrics*, vol. 42, no. 3, pp. 213–224, Aug. 2024, doi: 10.28951/bjb.v42i3.679.