BAREKENG: Journal of Mathematics and Its Applications
December 2023 Volume 17 Issue 4 Page 2273–2280
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss4pp2273-2280

DETERMINING STUDENT GRADUATION BASED ON SCHOOL LOCATION USING GEOGRAPHICALLY WEIGHTED LOGISTIC REGRESSION

Hendra Perdana[1], Neva Satyahadewi[2*], Fritzgerald Muhammad Arsyi[3]

[1,2,3]*Department Mathematics, Faculty of Mathematics and Science, Tanjungpura University*
Prof. Dr. Hadari Nawawi Street, Pontianak, 78124, Indonesia

*Corresponding author's e-mail: *neva.satya@math.untan.ac.id*

Article History:

Received: 26th July 2023
Revised: 10th October 2023
Accepted: 9th November 2023

Keywords:

Classification Accuracy;
GWLR;
Kernel Function;
Logistic Regression;
Spatial Data;
Student Graduation.

ABSTRACT

Faculty of Mathematics and Natural Sciences (FMIPA) is one of the Faculties in Tanjungpura University with 9 Undergraduate Programs (S1). Based on the graduation data of the 2014 batch of FMIPA students, the number of students who did not complete their studies was 131 students or 29% of the total 445 students and 187 schools in Indonesia. If the study period of students can be predicted early, the study program can provide advice or recommendations so that students can complete their studies in/exactly 8 semesters. This study aims to determine the model for analyzing the factors that influence the graduation of FMIPA students using GWLR. Geographically Weighted Logistic Regression (GWLR) is a developing logistic regression model applied to spatial data. This model is used to predict data with binary dependent variables that consider the location characteristics of each observation. The units of observation in this study are the school location of 455 students spread across Indonesia. The variables used in this study were sourced from the Academic and Student Affairs Bureau UNTAN and divided into dependent variables (Y) and independent variables (X), i.e. Gender, college selection, Accreditation, School Type, School Location, and Name of Study Program. The dependent variable analyzed is the graduated status of FMIPA UNTAN students, i.e. completed and not completed their studies. The results showed that gender and the name of the study program are factors that affect the graduation of FMIPA UNTAN 2014 students with a classification accuracy of 72.6%.

How to cite this article:

H. Perdana, N. Satyahadewi and F. M. Arsyi., "DETERMINING STUDENT GRADUATION BASED ON SCHOOL LOCATION USING GEOGRAPHICALLY WEIGHTED LOGISTIC REGRESSION," *BAREKENG: J. Math. & App.*, vol. 17, iss. 4, pp. 2273-2280, December, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

The Faculty of Mathematics and Natural Sciences (FMIPA) is one of the Faculties at Tanjungpura University (UNTAN) which has 9 Undergraduate Programs (S1) namely Mathematics, Physics, Chemistry, Biology, Computer Systems, Geophysics, Marine Science, Statistics, and Information Systems. Undergraduate Program (S1) FMIPA Tanjungpura University has a study load of at least 144 credits with a minimum study period of 3.5 years and a maximum of 7 years (14 semesters). Based on the graduation data of the 2014 batch of FMIPA students, the largest percentage of students who did not complete their studies came from the Computer Systems Department, which was 51.81%. If the study period of students can be predicted early, the relevant Study Program can provide advice/recommendations so that students can graduate in exactly 8 semesters.

Regression analysis is a statistical method that can be used to explain the relationship between dependent variables and independent variables. In general, regression analysis for spatial data is used to analyze data with quantitative (continuous) dependent variables that have a normal distribution. However, in practice, there are often qualitative (categorical) dependent variables in fields such as education, social, economic, and health [1]. The logistic regression model is one regression model that can explain the relationship between categorical dependent variables and independent variables [2].

Logistic regression is one of the statistical methods that can be used to find the relationship of dichotomous or binary variables with one or more independent variables that are continuous. However, logistic regression analysis does not consider geographical factors that may affect each observation. Therefore, to see the influence factors of student graduation, an analysis that considers geographical factors is needed because each location of school origin can have different characteristics, such as Geographically Weighted Logistic Regression (GWLR) [3].

GWLR models the relationship between a nominal dependent variable and independent variables that combine the GWR (Geographically Weighted Regression) model and the logistic regression model for the nominal dependent variable [4]. In this model, the coefficient value of each regression depends on the location of the observed data. The weights represent the location of the observed data from one another, the weights are used to provide parameter estimation results at different locations. The estimation of parameters at a point will be more influenced by points close to the location than points further away. Therefore, the selection of spatial weights used in estimating parameters is very important. The weight used is a kernel function. There are two types of kernel function weighting: Fixed Kernel and Adaptive Kernel [5]. Previous research on GWLR has been performed including GWLR Modeling on the Public Health Development Index (HDI) in Papua Province [1], the Use of GWLR models with Adaptive Gaussian Kernel weighting in Poverty Cases in East Nusa Tenggara Province [6], and factors affecting Population Growth Rate (PGR) of Semarang City using logistic regression and GWLR with Bisquare kernel and Gaussian kernel weighting functions [7].

Based on this explanation, the objective of the research is to determine the model for analyzing the factors that influence student graduation based on the school location using GWLR. The result of this study is a model of each school based on variables that have a significant effect on the graduation of FMIPA UNTAN students.

2. RESEARCH METHODS

2.1 Logistic Regression

The binary logistic regression model is used for data that is binary or dichotomous, namely each observation on the object is grouped as "failure" or "success" which is denoted 1 or 0 so that it follows the Bernoulli distribution as [8].

$$\boldsymbol{P(Y=y)} \quad = (\boldsymbol{p})^{\boldsymbol{y}}(\boldsymbol{1-p})^{(\boldsymbol{1-y})}; \boldsymbol{y=0,1} \tag{1}$$

$$\boldsymbol{\pi(x)} \quad = \frac{\boldsymbol{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}}{\boldsymbol{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}} \tag{2}$$

The function $\pi(x)$ is a nonlinear function that needs logit transformation to obtain a linear function. Thus the relationship between the dependent variable (y) and the independent variable (x) can be seen through logit transformation. The logit form of (x) is expressed as g(x)

$$\boldsymbol{g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p} \tag{3}$$

2.2 Multicollinearity Test

Multicollinearity is a condition that indicates a high linear intercorrelation among the explanatory variables in multiple regression models. This condition can lead to inappropriate regression analysis results. Diagnostic tools used to identify multicollinearity include the Variance Inflation Factor (VIF) [9]:

$$\boldsymbol{VIF_j = \frac{1}{1-R_j^2}} \tag{4}$$

VIF indicates how much the variance of an estimator is increased due to the presence of multicollinearity. If there is no collinearity between independent variables, VIF will be 1. The inverse of the VIF is called tolerance and both can be used interchangeably [10].

2.3 Heterogenity Test

Testing of the heterogenity spatial assumption is carried out to determine the variance of the residuals of different respond variables in each location or there is one observation location that has different residual variance [11]. Moran's Index for Spatial Dependence [12]. While, Breusch-Pagan is a statistical test that can detect spatial heterogeneity [13].

$$\boldsymbol{BP = \frac{1}{2}\left[f^T Z\left(Z^T Z\right)^{-1} Z^T f\right)\Big]} \tag{5}$$

where $\boldsymbol{Z}$ is a matrix of independent variables of size $n \times (k+1)$ and f is a vector of observations at $f_i = \left(\frac{\varepsilon_i^2}{\hat{\sigma}}\right) - 1$ of size $n \times 1$. The critical area rejects H_0 if the value of $BP > \chi^2_{\alpha,k}$ or $pvalue < \alpha$ [11].

2.4 Geographically Weighted Logistic Regression (GWLR)

Logistic regression with added geographically weight is called GWLR [8] and should be applied in the case of binary response variables [14]. GWLR is a local statistical technique, assuming regression vary spatially across the locations of all the case in the study population [15], it can be expressed by

$$\boldsymbol{(x_i) = \frac{\exp\left(\beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i v_i) x_{ki}\right)}{1 + \exp\left(\beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i v_i) x_{ki}\right)}} \tag{6}$$

$$\boldsymbol{g(x_i) = \beta_0(u_i, v_i) + \sum_{k=1}^{p} \beta_k(u_i, v_i) x_{ki}} \tag{7}$$

where, x_{ki} represents a set of independent variables $(k = 1, \ldots, p)$ for the i-th location, (u_i, v_i) is considered as the (x, y) coordinates of the i-th, β_k (regression coefficient) was the estimated effect of independent variable k for the i-th.

GWLR model parameters can be obtained by estimating using the Maximum Likelihood Estimation (MLE) method [16] and the Fisher Scoring Algorithm [17]. The Adaptive Exponential Kernel is used in this study. The adaptive capability causes the Adaptive Kernel function to be adjusted to the condition of the observation points [18].

2.5 Evaluating Classification

Apparent Error Rate (APER) is described as the value of the number of observations misclassified by the classification function [19]. The error rates for the two groups can be seen in Table 1:

Table 1. Classification Error

Actual Group Membership	*Predicted Group Membership*		**Total**
	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	
π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

The classification error can express as

$$\boldsymbol{APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}} \tag{8}$$

Where,

n_{1C}: Number of π_1 item correctly classified as $\pi_1$items
n_{1M}: Number of π_1 item misclassified as $\pi_1$items
n_{2C}: Number of π_2 items correctly classified
n_{2M} Number of π_2 items misclassified

Then, the Actual Error Rate (AER) is expressed as

$$\boldsymbol{AER = 1 - APER} \tag{9}$$

2.6 Data

The data used in this study are the graduation data of FMIPA UNTAN students' batch 2014 which is secondary data and sourced from BAK (Academic and Student Affairs Bureau) UNTAN. The 2014 batch of students was the last batch to graduate with complete school location data and 445 FMIPA students were used as samples. The variables used in this study are divided into dependent variables (Y), i.e. completed and not completed their studies, and independent variables (X), i.e. Gender, college selection, Accreditation, School Type, School Location, and Department Name.

3. RESULTS AND DISCUSSION

3.1 Multicollinearity Test

Multicollinearity test is conducted on the independent variables to determine whether there is a correlation between the independent variables before using logistic regression and GWLR.

Table 2. Variance Inflation Factor

Description	Variable	VIF
Gender	X_1	1.326
College Selection	X_2	1.341
Accreditation	X_3	1.131
School Type	X_4	1.155
School Location	X_5	1.137
Name of Study Program	X_6	1.781

Based on the VIF value in Table 2, it can be concluded that there is no multicollinearity between independent variables. The VIF value is less than 10, so all variables can be used in the formation of the GWLR model.

3.2 Spatial Heterogenity Test

The spatial heterogeneity test uses the Breusch-Pagan test which is shown in **Table 3**.

Table 3. Breusch Pagan Test

Uji *Breusch-Pagan*	$pvalue$
41.756	0.000245

Based on **Table 2**. the value of the *Breusch-Pagan* test is 41.756 and $pvalue = 0.000245$, with $\alpha = 0.1$ and $\chi^2_{(0.1;15)} = 22.30713$. Therefore $BP > \chi^2_{(0.1;15)}$, meaning that it can be decided that H_0 rejected. This means that the variance between locations is different or heterogeneity occurs.

3.3 Parameter Estimation of GWLR Model

Testing the parameters of the GWLR model with Adaptive Exponential Kernel weighting is used to determine the factors that influence the graduation of FMIPA UNTAN students using the Wald test at (u_i, v_i).

Table 4. Parameter Estimation of GWLR Model with Adaptive Exponential Kernel at (u_1, v_1)

Variable	Estimate	*Standard Error*	W	$pvalue$	
$\beta_{Gender:\,Male}$	-1.334	0.735	-1.814	0.129	*
$\beta_{Gender:\,Female}$	-1.500	0.697	-2.152	0.084	*
$\beta_{College\ Selection:\,SBMPTN}$	0.240	0.342	0.701	0.515	
$\beta_{College\ Selection:\,SNMPTN}$	0.371	0.348	1.066	0.335	
$\beta_{Accreditation:\,Selain\ A}$	0.425	0.291	1.460	0.204	
$\beta_{School\ Type:\,SMA}$	-0.146	0.433	-0.146	0.889	
$\beta_{School\ Type:\,SMK}$	0.073	0.552	0.132	0.900	
$\beta_{School\ Location:\,City}$	0.053	0.278	0.191	0.856	
$\beta_{Department\ Name:\,Physic}$	-0.457	0.518	-0.881	0.419	
$\beta_{Department\ Name:\,Geophysic}$	-0.406	-0.406	-0.406	0.702	
$\beta_{Department\ Name:\,Marine\ Science}$	-1.114	0.645	-1.727	0.145	*
$\beta_{Department\ Name:\,Chemistry}$	-1.912	0.488	-1.912	0.114	*
$\beta_{Department\ Name:\,Mathematic}$	0.315	0.472	0.668	0.534	
$\beta_{Department\ Name:\,Information\ System}$	0.252	0.560	0.449	0.672	
$\beta_{Department\ Name:\,Computer\ System}$	0.810	0.471	1.720	0.146	*
$\beta_{Department\ Name:\,Statistic}$	-1.755	0.508	-3.452	0.018	*

Based on the results in **Tabel 4**. with $\alpha = 10\%$ obtained $Z_{(0.1/2)} = 1.64$. There are six parameters that have a significant effect on the model, * indicating the value $|W| \geq Z_{\alpha/2}$. Thus, the GWLR model formed in the case of determining the graduation of the batch 2014 of FMIPA UNTAN students is as follows:

$$\boldsymbol{\pi(x) = \frac{\exp(-1.334x_1 - 1.500x_2 - 1.114x_{11} - 1.912x_{12} + 0.810x_{15} - 1.755x_{16})}{1 + \exp(-1.334x_1 - 1.500x_2 - 1.114x_{11} - 1.912x_{12} + 0.810x_{15} - 1.755x_{16})}}$$

Then the logit function is:

$$\boldsymbol{g(x) = -1.334x_1 - 1.500x_2 - 1.114x_{11} - 1.912x_{12} + 0.810x_{15} - 1.755x_{16}}$$

Table 5. Parameter Estimation of GWLR Model with Adaptive Exponential Kernel At (u_2, v_2)

Variable	Estimate	*Standard Error*	*W*	*pvalue*	
$\beta_{Gender:\,Male}$	-1.456	0.736	-1.979	0.105	*
$\beta_{Gender:\,Female}$	-1.649	0.694	-2.376	0.063	*
$\beta_{College\ Selection:\,SBMPTN}$	0.284	0.343	0.829	0.445	
$\beta_{College\ Selection:\,SNMPTN}$	0.285	0.343	0.831	0.444	
$\beta_{Accreditation:\,Selain\ A}$	0.450	0.294	1.533	0.186	
$\beta_{School\ Type:\,SMA}$	-0.047	0.429	-0.047	0.965	
$\beta_{School\ Type:\,SMK}$	0.130	0.554	0.234	0.824	
$\beta_{School\ Location:\,City}$	0.081	0.282	0.286	0.786	
$\beta_{Study\ Program:\,Physic}$	-0.440	0.525	-0.837	0.441	
$\beta_{Study\ Program:\,Geophysic}$	-0.529	-0.529	-0.529	0.619	
$\beta_{Study\ Program:\,Marine\ Science}$	-0.845	0.648	-1.305	0.249	
$\beta_{Study\ Program:\,Chemistry}$	-1.844	0.494	-1.844	0.125	*
$\beta_{Study\ Program:\,Mathematic}$	0.314	0.475	0.662	0.537	
$\beta_{Study\ Program:\,Information\ System}$	0.323	0.572	0.565	0.596	
$\beta_{Study\ Program:\,Computer\ System}$	0.727	0.469	1.551	0.182	
$\beta_{Study\ Program:\,Statistic}$	-1.670	0.517	-3.229	0.023	*

Based on the results in Table 5. with $\alpha = 10\%$ obtained $Z_{(0.1/2)} = 1.64$. There are four parameters that have a significant effect on the model, * indicating the value $|W| \geq Z_{\alpha/2}$. Thus, the GWLR model formed in the case of determining the graduation of the batch 2014 of FMIPA UNTAN students is as follows:

$$\boldsymbol{\pi(x) = \frac{\exp(-1.456x_1 - 1.649x_2 - 1.844x_{12} + 0.727x_{16})}{1 + \exp(-1.456x_1 - 1.649x_2 - 1.844x_{12} + 0.727x_{16})}}$$

Then the logit function is:

$$\boldsymbol{g(x) = -1.456x_1 - 1.649x_2 - 1.844x_{12} + 0.727x_{16}}$$

Location (u_1, v_2) and (u_2, v_2) have different parameters that significantly affect the model. This indicates that there is an influence of location on student graduation. The parameter testing process is repeated at each location, from (u_3, v_3) until (u_{455}, v_{455}) or until the last observation location.

3.4 Evaluating Classification of GWLR Model

The model classification accuracy test is a way to state the feasibility of the model, namely how much the percentage of observations is classified correctly. Model classification can be seen based on the classification results between observations and predictions using Table 1. The classification of the GWLR model with Adaptive Exponential Kernel weights can be seen in Table 6.

Table 6. Classification Error GWLR Model

Actual	Predicted		Total
	Complete	Not Complete	
Complete	291	23	314
Not Complete	99	32	131
Total	**390**	**55**	**445**

The calculation of Apparent Error Rate (APER) and Actual Error Rate (AER) values is as follows

$$\boldsymbol{APER = \left(\frac{99 + 23}{445}\right) = 0.274}$$

$$\boldsymbol{AER = 1 - 0,274 = 0.726}$$

Based on the APPER value obtained, the GWLR model formed has a classification error rate of 27.4%, which means that the GWLR model is correctly classified with a value of 72.6%.

4. CONCLUSIONS

Based on the results of the analysis, the factors that affect the graduation of FMIPA UNTAN 2014 students based using the GWLR model with Adaptive Exponential Kernel weighting are gender and name of study program with classification accuracy of 72.6%. The geographical location factor of this study has a significant influence because the research variables contain spatial heterogeneity.

REFERENCES

[1] M. Fathurahman, Purhadi, Sutikno, and V. Ratnasari, "Pemodelan Geographically Weighted Logistic Regression pada Indeks Pembangunan Kesehatan Masyarakat di Provinsi Papua," *Pros. Semin. Nas. MIPA 2016*, no. April, pp. 34–42, 2016.

[2] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression: Third Edition*. 2013. doi: 10.1002/9781118548387.

[3] Desriwendi, A. Hoyyi, and T. Wuryandari, "Pemodelan Geographically Weighted Logistic Regression (GWLR) dengan Fungsi Pembobot Fixed Gaussian Kernel dan Adaptive Gaussian Kernel," *Concept Commun.*, vol. 4, no. 2, pp. 193–204, 2015.

[4] M. Rifada and Purhadi, "Pemodelan Tingkat Kerawanan Demam Berdarah Dengue di Kabupaten Lamongan dengan Pendekatan Geographically Weighted Ordinal Logistic Regression," 2011.

[5] S. M. Soemartojo, R. D. Ghaisani, T. Siswantining, M. R. Shahab, and M. M. Ariyanto, "Parameter Estimation of Geographically Weighted Regression (GWR) Model Using Weighted Least Square and Its Application," *AIP Conf. Proc.*, 2018, doi: 10.1063/1.5054485.

[6] N. A. Solekha and M. F. Qudratullah, "Pemodelan Geographically Weighted Logistic Regression dengan Fungsi Adaptive Gaussian Kernel Terhadap Kemiskinan di Provinsi NTT," *Jambura J. Math.*, vol. 4, no. 1, pp. 17–32, 2022, doi: 10.34312/jjom.v4i1.11452.

[7] C. A. W. Aji, M. A. Mukid, and H. Yasin, "Analisis Faktor-Faktor Yang Mempengaruhi Laju Pertumbuhan Penduduk Kota Semarang Tahun 2011 Menggunakan Geographically Weighted Logistic Regression," *Gaussian*, vol. 3, no. 23, pp. 161–171, 2014.

[8] C. Zhang and Y. Yang, "Modeling the spatial variations in anthropogenic factors of soil heavy metal accumulation by geographically weighted logistic regression," *Sci. Total Environ.*, vol. 717, p. 137096, 2020, doi: 10.1016/j.scitotenv.2020.137096.

[9] J. H. Kim, "Multicollinearity and misleading statistical results," *Korean J. Anesthesiol.*, vol. 72, no. 6, pp. 558–569, 2019, doi: 10.4097/kja.19087.

[10] D. N. Gujarati, *Basic Econometrics*, 4th ed. Gary Burke, 2004.

[11] A. Yulia and S. Astutikand U Sa'adah, "Modeling Spatial Variation of Money Laundering Crime in Indonesia Using Geographically Weighted Multinomial Logistic Regression," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1115, no. 1, p. 012065, 2021, doi: 10.1088/1757-899x/1115/1/012065.

[12] A. M. Yolanda, K. Yunitaningtyas, and Indahwati, "Spatial Data Panel Analysis for Poverty in East Java Province 2012-2017," *J. Phys. Conf. Ser.*, vol. 1265, no. 1, 2019, doi: 10.1088/1742-6596/1265/1/012027.

[13] P. G. Widayaka, M. Mustafid, and R. Rahmawati, "Pendekatan Mixed Geographically Weighted Regression untuk Pemodelan Pertumbuhan Ekonomi Menurut Kabupaten/Kota di Jawa Tengah," *J. Gaussian*, vol. 5, no. 4, pp. 727–736, 2016, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[14] V. N. Mishra, V. Kumar, R. Prasad, and M. Punia, "Geographically Weighted Method Integrated with Logistic Regression for Analyzing Spatially Varying Accuracy Measures of Remote Sensing Image Classification," *J. Indian Soc. Remote Sens.*, vol. 49, no. 5, pp. 1189–1199, 2021, doi: 10.1007/s12524-020-01286-2.

[15] J. Tu and W. Tu, "How the relationships between preterm birth and ambient air pollution vary over space: A case study in Georgia, USA using geographically weighted logistic regression," *Appl. Geogr.*, vol. 92, no. January, pp. 31–40, 2018, doi: 10.1016/j.apgeog.2018.01.007.

[16] 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," *J. Phys. Conf. Ser.*, vol. 1776, no. 1, 2021, doi: 10.1088/1742-6596/1776/1/012048.

[17] M. Xu, C. L. Mei, and S. J. Hou, "Local-linear likelihood estimation of geographically weighted generalized linear models," *J. Spat. Sci.*, vol. 61, no. 1, pp. 99–117, 2016, doi: 10.1080/14498596.2016.1138245.

[18] A. R. Tizona, R. Goejantoro, and Wasono, "Pemodelan Geographically Weighted Regression (Gwr) dengan Fungsi Pembobot Adaptive Kernel Bisquare untuk Angka Kesakitan Demam Berdarah di Kalimantan Timur Tahun 2015," *J. Eksponensial*, vol. 8, no. 1, 2017.
[19] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis.: Pearson Prentice Hall*. Pearson Prentice Hall, 2007.