Stroke Classification Model using Logistic Regression

Stroke Classification Model using Logistic Regression

S Annas[1*], A Aswi[1], M Abdy[2], and B Poerwanto[1]
[1]Statistics Department, Universitas Negeri Makassar, Indonesia
[2]Mathematics Department, Universitas Negeri Makassar, Indonesia

*Email: suwardi_annas@unm.ac.id

Abstract. This study aims to determine the factors that significantly affect the classification of stroke. The response variable used is the type of stroke, namely non-hemorrhagic stroke and hemorrhagic stroke. The predictors used were cholesterol level, blood sugar level, temperature, length of stay, pulse rate, and gender. By using logistic regression, the results obtained modeling accuracy of 74.8% where the predictors that have a significant effect (alpha <0.05) are cholesterol level and length of stay.

Keywords: logistic regression, stroke, classification

1. Introduction

Stroke is the leading cause of disability in adults and is the third leading cause of death in the world [1]. In Indonesia, based on the results of the 2018 Basic Health Research, the prevalence of stroke based on diagnosis in the population aged 15 years in Indonesia increased from the previous 7 to 10.9 per mil compared to the prevalence in 2013, or in other words, there was an increase for all provinces in Indonesia [2].

Based on a study from the ASEAN Neurological Association (ASNA) [3], it was found that from 2065 patients spread across 28 hospitals in Indonesia, most of them arrived 6 hours late at the hospital since the stroke with the most dominant reason being not being aware of the symptoms. In addition, this disease is also often repeated to patients. From this study, it was found that 20% of patients had recurrent strokes.

One way to prevent or reduce the serious impact of stroke is to first know the risk factors that significantly affect it. To overcome these problems, it is necessary to conduct an analysis that aims to determine significantly the risk factors for stroke. This analysis can also classify the type of stroke between hemorrhagic and non-hemorrhagic (ischemic) based on data from the analyzed factors. Furthermore, based on the model formed, this analysis can predict a person's stroke status. The analysis used is logistic regression (LR) which is a statistical method to determine the relationship between the dependent variable (response) which is categorical and one or more independent variables (predictors) in the form of categorical data or continuous data [4].

LR is one of the machine learning methods that are currently widely used in many fields. LR method enables to perform classification analysis and also enables to provide information about variables that have a significant effect [5]. This method is also often used in health data, for example, research by Bustan and Poerwanto [6] who modeled breast cancer pathology diagnosis with metastasis. Another example is Josephus, et al. [7] who predicted death from Covid 19 using nonmedical features with logistic regression.

2. Material and method

2.1 Logistic regression

LR is the most frequently used linear prediction method for binary data. If the response used has two categories, then the regression analysis used is called binary LR [5]. The multiple binary logistic regression model can be written as follows:

$$\pi(x_i) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_iX_i)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_iX_i)} \quad (1)$$

The β_0, β_1, ... β_i are parameters of the model. Those parameters are estimated by using the maximum likelihood estimation (MLE) method. Basically, the MLE provides an estimated value of β to maximize the likelihood function [8]. Systematically, the likelihood function for the binary logistic regression model is as follows:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (2)$$

where :

y_i : observations on the i^{th} variables

$\pi(x_i)$: probability for the i^{th} predictor variable

Equation (2) is solved by using the log-likelihood approach, defined as follows:

$$l(\beta) = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\} \quad (3)$$

To get the values of $\hat{\beta}$, equation (3) is derived from β and then equalized to 0.

2.2 Stroke

Rapidly developing clinical signs due to local (or global) brain disorder with symptoms lasting 24 hours or more and can lead to death in the absence of other obvious causes other than vascular disease are signs of stroke [9]. Stroke is divided into two types, namely ischemic or non-hemorrhagic stroke (NHS) and hemorrhagic stroke (HS). Ischemic stroke is a category of stroke that can occur due to obstruction or clot in one or more large arteries in the cerebral circulation. Patients with ischemic stroke usually relapse because of seizures, migraines, and other conversion disorders that trigger patient relapse [10]. While hemorrhagic stroke is a category of stroke that can occur if the intracerebral vascular lesion ruptures, causing bleeding into the subarachnoid space or directly into brain tissue [11].

Based on a release from the Indonesian Ministry of Health [12], World Stoke Organization data shows that every year there are 13.7 million new cases of stroke and about 5.5 million deaths from this disease. About 70% of these strokes occur in low- and middle-income countries, including Indonesia [13]. Over the last 4 decades, the incidence in low- and middle-income countries has doubled. This disease of course, in addition to having an impact on the socioeconomic, also causes permanent disability and productivity of sufferers. Based on data from BPJS Health, stroke is one of the diseases with the highest cost, namely 2.56 trillion rupiahs in 2018 and increasing every year [12].

2.3 Data and variables

The data used was 261 medical record data consisting of 161 non-hemorrhagic stroke and hemorrhagic stroke 101 from Dadi Hospital, Makassar. The description of the variables used can be seen in Table 1.

Table 1. Description of variables

Variables	Description
Stroke Type (Y)	0: Non-hemorrhagic 1: hemorrhagic
Cholesterol Level (X_1)	mg/dl
Blood sugar level (X_2)	mg/dl
Temperature (X_3)	Celcius
Lenght of stay (X_4)	Days
Pulse rate (X_5)	x/min
Gender (X_6)	0: Male 1: Female

3. Result and discussion

3.1 Simultaneous significance test results

Simultaneous testing is carried out to see the effect of the overall predictor variables on the response variable. The hypothesis for this test is as follows:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_i = 0$
H_1: there is at least one parameter $\neq 0$

According to David & Stanley [14] the test statistics used for the likelihood ratio test is as follows:

$$G = -2log\frac{l_0}{l_1} \quad (4)$$

By using an alpha (α) of 0.05, it means that the test criterion is to reject H_0 if $G > \chi^2_{(df,\alpha)}$. The result G is 303.068 while $\chi^2_{(df,\alpha)}$ is 298.611. Therefore, this step concludes that there is at least one parameter that is not equal to zero.

3.2 Partial test results

This partial test was conducted to identify the predictors which had a significant effect on the dependent variable. The results for each predictor can be seen in Table 2.

Table 2. Wald test

Predictors	$\boldsymbol{\beta}$	**Standard Error**	**Wald**	**P-Value**	**Exp($\boldsymbol{\beta}$)**
Cholesterol level	-0.006	0.003	4.038	0.044	0.994
Blood sugar level	-0.002	0.003	0.746	0.388	0.998
Temperature	-0.419	0.218	3.695	0.055	0.657
Length of stay	0.175	0.031	31.515	0.000	1.192
Pulse rate	0.014	0.010	2.254	0.133	1.015
Gender(1)	-0.448	0.290	2.376	0.123	0.639
Constant	13.694				

The hypothesis used for each variable is as follows:
$H_0: \beta_i = 0$ (the logit coefficient is not significant to the model)
$H_1: \beta_i \neq 0$ (the logit coefficient is significant to the model)

In the partial test, the test statistics used is the Wald test which can be counted by using the equation (5).

$$W = \frac{\widehat{\beta}_i}{S.E(\widehat{\beta}_i)} \quad (5)$$

The *W* value follows a Chi-square distribution with df = 1. The criterion for rejecting H_0 is if $W > \chi^2_{(1,\alpha)}$ or *p-value* < 0.05. By using *p-value*, there are 2 predictors having p-value < α, namely cholesterol level, and length of stay. According to significant predictors from Table 2, the logistic regression model can be seen in equation (6).

$$\pi(x) = \frac{exp(13.694-0.006X_1+0.175X_4)}{1+exp(13.694-0.006X_1+0.175X_4)} \quad (6)$$

3.3 Goodness of fit test

The test statistic used to measure the goodness of fit is the Hosmer and Lemeshow test given in equation (7).

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i-N_i\overline{\pi}_i)^2}{N_i\overline{\pi}_i(1-\overline{\pi}_i)} \quad (7)$$

The criterion for rejecting H_0 is if $\chi^2_{HL} \geq \chi^2_{(g-2,\alpha)}$ or *p-value* < 0.05 while the hypothesis is:

H_0: The model has sufficiently explained the data
H_1: The model does not adequately explain the data

The result for the Hosmer and Lemeshow test can be seen in Table 3.

Table 3. Hosmer and Lemeshow test

Chi-Square	P-value
15.084	0.0058

It can be seen from Table 3 that the p-value is more than α, therefore the conclusion is to accept H_0 or the model has sufficiently explained the data.

4. Discussion

In this study, logistic regression was used to determine the factors that significantly influence stroke so that people can prevent stroke as early as possible. Two predictors have a significant effect on the model, namely cholesterol level and length of stay. In general, the accuracy of the model is around 75% and it can be seen in Table 4.

Table 4. The accuracy of the model

		Predicted		
		Stroke Type		Percentage
Observed		NHS	HS	Correct
Stroke Type	NHS	141	20	87.6
	HS	46	55	54.5
Overall Percentage				74.8

As can be seen in Table 2 that an increase in cholesterol level will increase the risk of NHS by less than 1 time compared to HS. Cholesterol level was positively associated with stroke mortality [15], therefore it can be an early warning to prevent stroke. On the other hand, an increase of 3 days of the length of stay will increase the risk of NHS by almost twice compared to HS.

5. Conclusion

Based on the analysis that has been done, it can be concluded that cholesterol level and length of stay are two predictors significantly affecting stroke type. The logistic regression model had 75% of accuracy in classifying 261 medical record stroke data.

6. Acknowledgement

The authors give great thanks to those who have provided funding assistance, Kemdikbudristek, through the PDUPT research grant scheme.

7. References

[1] Abubakar, S. A., & Isezuo, S. A. 2012. Health related quality of life of stroke survivors: experience of a stroke unit. *International journal of biomedical science: IJBS*, 8(3), 183.

[2] Kementerian Kesehatan Republik Indonesia. 2018. *Hasil Utama Riskesdas 2018*. Badan Penelitian dan Pengembangan Kesehatan Kementerian Kesehatan Republik Indonesia.

[3] Usrin, I. 2013. Pengaruh hipertensi terhadap kejadian stroke iskemik dan stroke hemoragik di ruang Neurologi di Rumah Sakit Stroke Nasional (RSSN) Bukittinggi tahun 2011. *Kebijakan, Promosi Kesehatan dan Biostatiskik*, 2(2).

[4] Agresti, Alan. 2002. *Categorical Data Analysis, 3rd edition, John Wiley & Sons, Inc.*, New York.

[5] Fa'rifah, R. Y. dan Poerwanto, B. 2019. Penerapan Regresi Logistik dalam Menganalisis Faktor Penyebab Peningkatan Angka Kematian Bayi. *Jurnal Ilmiah d'ComPutarE*, 9(1), 52-55.

[6] Bustan, M. N., & Poerwanto, B. 2021. Logistic Regression Model of Relationship between Breast Cancer Pathology Diagnosis with Metastasis. *Journal of Physics: Conference Series*, 1752(1), 1–5. https://doi.org/10.1088/1742-6596/1752/1/012026

[7] Josephus, B. O., Nawir, A. H., Wijaya, E., Moniaga, J. V., & Ohyver, M. 2021. Predict Mortality in Patients Infected with COVID-19 Virus Based on Observed Characteristics of the Patient using Logistic Regression. *Procedia Computer Science*, 179(2019), 871–877. https://doi.org/10.1016/j.procs.2021.01.0760

[8] Tampil, Y., Komaliq, H., & Langi, Y. 2017. Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado. *D'CARTESIAN*, 6(2), 56–62. https://doi.org/10.35799/dc.6.2.2017.17023

[9] Bootkrajang, J., & Kabán, A. 2014. Learning kernel logistic regression in the presence of class label noise. *Pattern Recognition*, 1–15. https://doi.org/10.1016/j.patcog.2014.05.007

[10] Geyer, J. D., Faasm, M. D., & Gomes, C. R. 2009. *Stroke: A Practical Approach. (F. DeStefano & L. McMillan, Eds.) (1st ed.)*. Philadelphia USA: Lippincott Williams & Wilkins, a Wolters Kluwer Business.

[11] Garg, R., Rech, M. A., & Schneck, M. 2019. Stroke Mimics: An Important Source of Bias in Acute Ischemic Stroke Research. *Journal of Stroke and Cerebrovascular Diseases*., 1–6. https://doi.org/10.1016/j.jstrokecerebrovasdis.

[12] Kementerian Kesehatan Republik Indonesia. 2018. *Stroke Don't Be The One (2019)*. Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia

[13] Bustan, M. N. 2015. *Manajemen Pengendalian Penyakit Tidak Menular*. Jakarta: Rineka Cipta.

[14] David, W. . H., & Stanley, L. 2000. *Applied logistic regression*. Willey

[15] Yi, S. W., Shin, D. H., Kim, H., Yi, J. J., & Ohrr, H. 2018. Total cholesterol and stroke mortality in middle-aged and elderly adults: A prospective cohort study. *Atherosclerosis*, *270*, 211–217. https://doi.org/10.1016/j.atherosclerosis.2017.12.003