Revista Científica Interdisciplinaria Investigación y Saberes
2024, Vol. 14, No. 1 e-ISSN: 1390-8146
Published by: Universidad Técnica Luis Vargas Torres

Predicción de accidente cerebrovascular utilizando regresión logística

Prediction of stroke using logistic regression

David Fernando Ramos Tomalá
Faculty of Industrial Engineering, University of Guayaquil, david.ramost@ug.edu.ec,
https://orcid.org/0009-0007-2702-8926

Mariuxi Del Carmen Toapanta Bernabé
Faculty of Industrial Engineering, University of Guayaquil, mariuxi.toapantab@ug.edu.ec,
https://orcid.org/0000-0002-4839-7452

Cesar Andrés Alcívar Aray
Faculty of Industrial Engineering, University of Guayaquil, cesar.alcivarar@ug.edu.ec,
https://orcid.org/0009-0004-0214-905X

Received 2023-11-18
Revised 2023-12-19
Published 2024-01-05
Corresponding Author
David Fernando Ramos Tomalá
david.ramost@ug.edu.ec
Pages: 158-177

Abstract

Open clinical trial data provide a valuable opportunity for researchers worldwide to evaluate new hypotheses, validate published results, and collaborate for scientific advances in medical research. Here we present a health dataset for noninvasive stroke prediction, containing data from 219 subjects. The dataset covers an age range of 20-89 years and disease records including hypertension and diabetes. Data acquisition was performed under the control of standard experimental conditions and specifications. The response variable used is whether or not stroke occurred, whether nonhemorrhagic or hemorrhagic. The predictors used were age, body mass index, blood sugar level, systolic and diastolic blood pressure, heart rate, and gender. Using logistic regression, good modeling accuracy was obtained, where the predictors with a significant effect (alpha <0.05) were age, body mass index, diastolic blood pressure, and type of diabetes.

How to cite this article (APA): Ramos, D., Toapanta, M., Alcivar, C. (2024) Prediction of stroke using logistic regression, *Revista Científica Interdisciplinaria Investigación y Saberes,* 14(1) 158-178

Keywords: stroke prediction, correlation matrix, ROC curve, sensitivity and specificity, confounding matrix, logistic regression, machine learning

Resumen

Los datos de ensayos clínicos abiertos proporcionan una valiosa oportunidad para que los investigadores de todo el mundo evalúen nuevas hipótesis, validen los resultados publicados y colaboren para obtener avances científicos en la investigación médica. Aquí se presenta un conjunto de datos de salud para la predicción no invasiva de accidentes cerebrovasculares, que contiene datos de 219 sujetos. El conjunto de datos cubre un rango de edad de 20-89 años y registros de enfermedades incluyendo hipertensión y diabetes. La adquisición de datos se llevó a cabo bajo el control de las condiciones y especificaciones experimentales estándar. La variable de respuesta utilizada es si se produjo o no el accidente cerebrovascular, sea este no hemorrágico o hemorrágico. Los predictores utilizados fueron la edad, el índice de masa corporal, el nivel de azúcar en la sangre, la presión arterial sistólica y diastólica, la frecuencia cardíaca y el género. Mediante el uso de la regresión logística, se obtuvo una buena precisión de modelado, donde los predictores que tienen un efecto significativo (alfa <0,05) son la edad, el índice de masa corporal, la presión arterial diastólica y el tipo de diabetes.

Palabras clave: predicción ictus, matriz de correlación, curva ROC, sensibilidad y especificidad, matriz de confusión, regresión logística, machine learning

Introduction

Stroke (CVA) is the leading cause of disability in adults and the third leading cause of death in the world (Abubakar & Isezuo, 2012). It is a disease that affects the arteries leading to and within the brain. The impact of stroke on people's lives represents a major challenge to

society. In addition to being a sudden event, stroke affects both the individual and family members who are not prepared to deal with the rehabilitation process or the disabilities that result from this condition. As a result, a large number of people are unable to work and receive financial assistance after suffering a stroke.

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 significantly determine the risk factors for stroke. Moreover, based on the model formed, this analysis can predict the occurrence or non-occurrence of stroke in a person. The analysis used is logistic regression (LR), which is a statistical method to determine the relationship between the dependent variable (response) that is categorical and one or more independent variables (predictors) in the form of categorical or continuous data (Agresti, 2002).

LR is one of the machine learning methods that are currently widely used in many fields. The LR method allows performing classification analysis and also providing information about variables that have a significant effect (Fa'rifah & Poerwanto, 2019). This method is also often used in health data, for example, research that modeled the pathological diagnosis of metastatic breast cancer (Bustan & Poerwanto, 2021). In another example, death from COVID 19 was predicted using non-medical characteristics with logistic regression (Josephus et al., 2021).

Rapidly developing clinical signs due to a local (or global) brain disorder with symptoms that last 24 hours or more and may result in death in the absence of obvious causes other than vascular disease are signs of stroke (Bootkrajang & Kaban, 2014). Stroke, or ictus, is divided into two types, namely ischemic or nonhemorrhagic stroke (NHS) and hemorrhagic stroke (HS). Ischemic stroke is a category of stroke that can occur due to blockage or clot in one or more large arteries in the cerebral circulation. Patients with ischemic stroke often relapse due to seizures, migraines, and other conversion disorders that trigger relapse (Geyer et al., 2009). 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 the brain tissue (Garg et al., 2019).

In the last four decades, the incidence in low- and middle-income countries has doubled. This disease, of course, in addition to having an impact on socioeconomic, also causes permanent disability and productivity of the sick (Kementerian Kesehatan Republik Indonesia, 2018).

Data from the World Stroke Organization show that each year there are 13.7 million new cases of stroke and about 5.5 million deaths from stroke. About 70 % of these strokes occur in low- and middle-income countries (Bustan M. N., 2017).

Methodology

Many studies have been conducted using early diagnostic and noninvasive screening techniques for cardiovascular diseases (CVD) such as hypertension and coronary artery sclerosis in order to discover more convenient and effective methods for early identification of CVD. Of these methods, photoplethysmography (PPG) has been widely recognized as a low-cost noninvasive screening technology for CVD (Allen, 2007).

Data acquisition was conducted at the Guilin People's Hospital, China (Liang et al., 2017).

The information from the working dataset is summarized in Table 1. This dataset was collected from 219 subjects, aged 21 to 86 years, with a median age of 58 years. Males accounted for 48%. The dataset covers several diseases, including hypertension, diabetes, cerebral infarction, and cerebral insufficient blood supply.

The data collection program involved acquiring information on the basic physiology of the individuals, extracting cardiovascular disease information from the hospital's electronic medical records, collecting PPG waveform signals, and detecting instantaneous blood pressure at the same time.

The data set includes systolic and diastolic blood pressure information from subjects who were diagnosed with normotension, prehypertension, and stage I/II hypertension:

Table 1

Working dataset information

Article Site	US National Library of Medicine National Institutes of Health
Item name	A new short-log photoplethysmogram data set for blood pressure monitoring in China.
Article website	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5827692/
Dataset repository	https://figshare.com/articles/dataset/PPG-BP_Database_zip/5459299
Remarks	It has a total of 219 records, The dataset also covers several different cardiovascular (e.v.) diseases, including hypertension, cerebral infarction and insufficient blood supply to the brain and other related diseases, such as diabetes.
Variables	Sex(M/F) = sex (Male / Female)

	Age = Age
	Height(cm) = Height
	Weight(kg) = Weight
	Systolic Blood Pressure(mmHg) = Systolic Blood Pressure
	Diastolic Blood Pressure(mmHg) = Diastolic Blood Pressure
	Heart Rate(b/m) = Heart rate
	BMI (kg/m^2) = Body Mass Index
	(v.e) Hypertension = Hypertension
	(v.e) Diabetes
	(e.g.) cerebral infarction = cerebral infarction
	(e.g.) cerebrovascular disease = cerebrovascular accident

Note: This table explains where the information was obtained from, the sample size, and the variables that were finally considered relevant. Taken from ***PPG-BP Database***, by Y. Liang et al., 2017, Figshare.

Table 2 shows the data for all independent variables (x_{ij}) and those of the dependent variable (Y_i).

Table 2

Sample data

Sexo	Edad	IMCo	FrCa	Sist	Diast	TD	EnCe
1	45	27,27	97	161	89	0	0
1	50	20,28	76	160	93	0	0
1	47	20,89	79	101	71	0	0
0	45	21,97	87	136	93	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1	68	19,63	65	142	90	0	0
0	52	23,03	80	154	88	0	0
0	66	24,77	93	173	107	0	0
0	65	22,58	73	111	62	0	0
0	66	19,03	84	107	63	0	0
0	47	23,44	71	128	66	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1	24	20,7	74	108	65	0	0
1	25	20,81	64	84	56	0	0
0	25	23,46	63	104	70	0	0
1	24	19,49	87	109	68	0	0
0	24	21,6	77	111	70	0	0
1	25	19,31	79	93	57	0	0
0	25	17,76	72	120	69	0	0
0	25	21,05	67	106	69	0	0
0	24	18,94	65	108	68	0	0

Note. In total there are 219 rows and 8 columns: grouped in 7 independent variables (x_j) (Sex, Body Mass Index, Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, and Type of Diabetes), of which 2 (Sex and Type of Diabetes) are binary (0 or 1); and all predict

1 dependent variable (Y) (Brain Disease) which is also binary (0 or 1).

Taken from *PPG-BP Database*, by Y. Liang et al., 2017, Figshare.

Prediction of stroke occurrence in subjects:

The data set used for this study can be evaluated here with responses from 219 patients of both sexes who had a stroke (1) and who did not have a stroke (0).

The logistic regression model assumes that each response Yi is an independent random variable with Bernoulli distribution (pi), where the log-odds corresponding to pi are modeled as a linear combination of the covariates plus a possible intercept term:

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

The intercept **β0** represents the "baseline" log-odds of the subject who will suffer a stroke, if all covariates take the value 0. Each coefficient **βj represents** the amount of increase or decrease in log-odds, if the value of covariate xij is increased by 1 unit.

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

$$\mathbb{P}[Y_i = 1] = p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}. \qquad (1)$$

The **β0**, **β1**, ... **βi** are parameters of the model. These parameters are estimated using the maximum likelihood estimation (MLE) method. Basically, the MLE provides an estimated value of **β** to maximize the likelihood function (Tampil et al., 2017). Systematically, the likelihood function for the binary logistic regression model is as follows:

$$\text{lik}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n}(1-p_i)\left(\frac{p_i}{1-p_i}\right)^{Y_i},$$

Given that the responses Y1, ..., Yn are independent Bernoulli random variables, the probability for the logistic regression model is given by:

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} Y_i \log\frac{p_i}{1-p_i} + \log(1-p_i) = \sum_{i=1}^{n}\left(Y_i \sum_{j=0}^{p}\beta_j x_{ij} - \log\left(1+e^{\sum_{j=0}^{p}\beta_j x_{ij}}\right)\right).$$

To obtain the values of β, the equation is derived from β and then equals 0:

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^{n} x_{im}\left(Y_i - \frac{e^{\sum_{j=0}^{p}\beta_j x_{ij}}}{1+e^{\sum_{j=0}^{p}\beta_j x_{ij}}}\right).$$

Logistic regression is still often used as a tool for binary classification problems, even if the model does not give an extremely accurate fit to the data, as long as the model has good classification accuracy. In such fits, the **βˆ** MLE represents the "logistic regression model" closest (on the given covariates) to the true distribution of Y1, . . ., Yn. The standard error for **βˆj** can be robustly estimated using a sandwich estimator or the nonparametric bootstrap. For the logistic regression model, the sandwich estimation of the covariance matrix of **βˆ** is given by:

$$(X^T\hat{W}X)^{-1}(X^T\tilde{W}X)(X^T\hat{W}X)^{-1},$$

where W = diag ((Y1 - pˆ1)2, ... (Yn - pˆn)2) and pˆi is the probability of fit for the ith observation, defined by the right-hand side of formula (1) with **βˆ** instead of **β**. The (j, j) element of this matrix gives an estimate of the variance of **βˆj**. Alternatively, one can use the pairwise bootstrap, which pairs the covariates and the response for each ith observation into a single data vector (xi1, ..., xip, Yi), and then draws bootstrap samples by randomly selecting, with replacement, n from these vectors. The logistic regression model is fit to each bootstrap sample to produce an MLE **βˆˆ**, and the standard error of **βˆj** is estimated by the empirical standard deviation of **βˆ** across bootstrap samples.

ROC curves (receiver operating characteristic curve) are a very useful tool in health sciences, basically under two circumstances. The first is to be able to determine the best cut-off point for a diagnostic test; that is, in this case, what will be the criterion for predicting whether or not a subject will have a stroke. The second is to be able to determine, from among two tests, with which one can obtain better results in establishing a diagnosis; that is, in this case, with which one will be able to better predict whether or not a subject will have a stroke.

A diagnostic test is basically a classifier that attempts to distinguish individuals who have a pathology or condition from those who do not; that is, in this case, individuals who will or will not suffer a stroke in the future. However, like any detection system, it is susceptible to errors. In this case, the possible errors would be: predicting that a person will not have a stroke when he or she will; or predicting that a person will have a stroke when he or she will not. In this sense we can say that a diagnostic test can present four scenarios for each individual:

Individual without stroke classifies as without stroke (True negative: TN). Individual with stroke classifies as having stroke (True positive: TP). Individual with stroke classifies as without stroke (False negative: FN). And, finally, an individual without stroke is classified as having a stroke (False positive: FP).

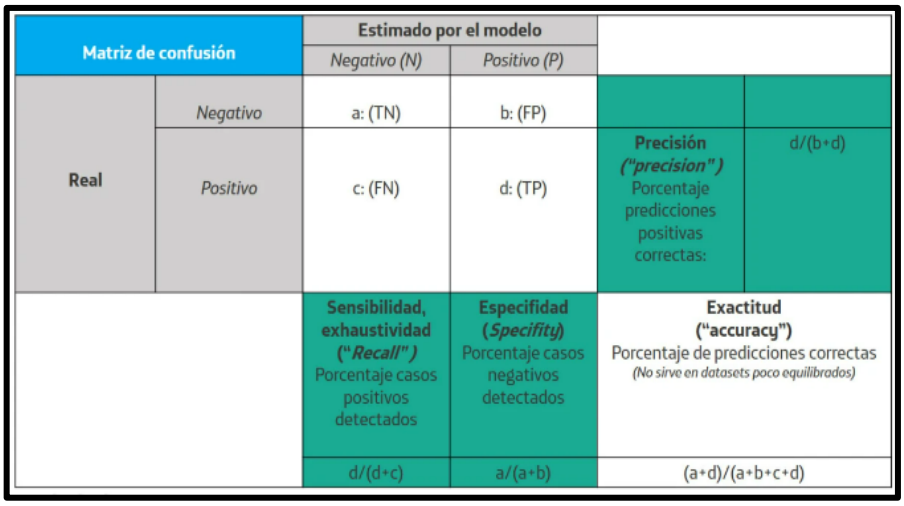
Clearly the first pair is correct and the second pair is wrong. Thus, diagnostic tests can be evaluated by quantifying how many times a correct result is obtained with respect to the total number of tests. The parameters that establish how accurate the diagnostic test is are sensitivity and specificity, which are defined as follows:

Sensitivity: The probability of classifying an individual with stroke as having a stroke. Specificity: The probability of classifying an individual without stroke as not having a stroke.

For the calculation of these parameters it is very practical to organize the information in a contingency matrix of size 2×2, as shown in Figure 1, where the columns are generated from the LCA variable, and the rows from the Test variable:

Figure 1

Confusion Matrix



Matriz de confusión		Estimado por el modelo		
		Negativo (N)	*Positivo (P)*	
Real	*Negativo*	a: (TN)	b: (FP)	
	Positivo	c: (FN)	d: (TP)	**Precisión** ***("precision")*** Porcentaje predicciones positivas correctas: d/(b+d)
		Sensibilidad, exhaustividad ***("Recall")*** Porcentaje casos positivos detectados	**Especifidad** ***(Specifity)*** Porcentaje casos negativos detectados	**Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)*
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)

Note: Where: TP: True Positive; TN: True Negative; FP: False Positive; and FN: False Negative. Taken from *Cómo interpretar la matriz de confusión: ejemplo práctico*, by P. Recuero de los Santos, 2021, Telefónica Tech.

In terms of conditional probability, sensitivity is understood as the probability that the test predicts stroke given that the individual will actually have stroke (P (predict stroke | will have stroke)), and specificity is the probability that the test predicts no stroke given that the individual will not actually have stroke (P (predict no stroke | will not have stroke)). Applying Bayes' Rule and the Law of Total Probability we can arrive at the following equations for sensitivity and specificity (Recuero de los Santos, 2021).

$$Sensibilidad = \frac{TP}{TP + FN} \qquad Especificidad = \frac{TN}{TN + FP}$$

It is important to clarify that in order to consider a person as having or not having a stroke, it is necessary to have a reference test that indicates in the first instance the state of the individual in question.

Suppose that the value of a biomarker X of a person is known. This biomarker could have the property of defining the person as having or not having stroke if it is above (or below) a certain cut-off point C at first unknown.

An ROC curve is a graphical representation of the variation of sensitivity versus (1-specificity) when the cut-off point changes. In other words, for each possible cut-off point, a sensitivity and specificity must be calculated and plotted.

One parameter to determine the capacity of a diagnostic test to discriminate between possible or not strokes is the area under the curve; that is, the area delimited by the (1-specificity) axis, the vertical line passing through (1-specificity) =1, and the ROC curve. The maximum value of this area is less than or equal to unity and a test is said to be non-discriminative if it coincides with the straight line joining the point (0,0) with (1,1) which defines an area equal to 0.5 (50%). The larger the area, the more discriminative it is considered to be. Thus, the area under the curve can be a good parameter to compare two diagnostic tests, considering that the best one will be the one that covers the largest area.

Another aspect of interest is to determine the cut-off point that best identifies those who will have a stroke from those who will not. In terms of sensitivity and specificity, the best cut-off point would be the one that classifies all those who will have a stroke as positive and all those who will not have a stroke as negative. In practice, it is very difficult to ensure that a diagnostic test is infallible, so obtaining perfect sensitivity and specificity would be very complicated; however, the cut-off point that best fits these conditions can be sought.

One criterion for finding the best cut-off point is to identify the point P for which the distance from point P to point (0,1) is minimum.

Results

Table 3 shows that systolic blood pressure (Sist) is directly related to diastolic blood pressure (Diast) ($\boldsymbol{r = 0.72}$); which was to be expected.

Table 3

Pearson correlation matrix

	Sexo	Edad	IMCo	FrCa	Sist	Diast	TD
Edad	-0,07						
IMCo	0,04	0,02					
FrCa	-0,01	-0,09	-0,11				
Sist	-0,09	0,41	0,23	0,14			
Diast	-0,08	0,00	0,23	0,19	0,72		
TD	0,00	0,06	0,13	-0,02	0,07	0,06	
EnCe	0,01	0,38	0,08	-0,04	0,14	-0,11	-0,23

Note. The correlation, in absolute value, can be: null (0.00); very weak (0.01-0.20); weak (0.21-0.40); moderate (0.41-0.60); strong (0.61-0.80); very strong (0.81-0.99); or perfect (1.00).

4.2 Simultaneous significance test results

Simultaneous tests are conducted to see the effect of the global predictor variables on the response variable. The hypothesis for this test is as follows:

H0: $\beta 1 = \beta 2 = = = \beta \mathrm{i} = 0$

H1: there is at least one parameter 0

The test statistics used for the likelihood ratio test are as follows (David & Stanley, 2000):

$$-2\log\Lambda = -2\log\frac{\mathrm{lik}(\hat{\beta}_{0,0},\ldots,\hat{\beta}_{0,p-1},0)}{\mathrm{lik}(\hat{\beta}_0,\ldots,\hat{\beta}_p)}.$$

As Table 4 shows, the result D is 69.85 while the p-Value is 1.59E-12. Therefore, this step concludes that there is at least one parameter that is not equal to zero.

Table 4

Simultaneous significance test result

$pseudo\text{-}R^2$	D	-111,231275	df	p-Value$_\beta$
0,313966578	69,84580529	-76,3083721	7	1,58732E-12

Note. By using an alpha (α) of 0.05, it means that the test criterion is to reject H0 if D > χ2.

Partial test results

This partial test was performed to identify predictors that had a significant effect on the dependent variable (Y). Table 5 shows the individual (partial) results for each of the 7 predictors, and for the constant term.

Table 5

Results of partial significance tests

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
-8,26444356	0,101862291	0,084110733	0,13140241	0,008189016	0,019807976	-0,06116992	-85,2100284
$se_{\beta 0}$	$se_{\beta 1}$	$se_{\beta 2}$	$se_{\beta 3}$	$se_{\beta 4}$	$se_{\beta 5}$	$se_{\beta 6}$	$se_{\beta 7}$
2,275298657	0,417872836	0,018041966	0,049060926	0,017092958	0,014433266	0,02983895	0,506265124
-81,9379589	-76,3393417	-87,0478724	-79,5971918	-76,3959504	-77,1174776	-78,4107112	-87,8653917
$Z_{\beta 0}$	$Z_{\beta 1}$	$Z_{\beta 2}$	$Z_{\beta 3}$	$Z_{\beta 4}$	$Z_{\beta 5}$	$Z_{\beta 6}$	$Z_{\beta 7}$
-3,63224561	0,243763848	4,661949427	2,678351627	0,479087092	1,372383495	-2,05000244	-168,311077
11,25917372	0,061939157	21,47900071	6,577639462	0,175156536	1,618211034	4,20467829	23,1140393
p-Value$_{\beta 0}$	p-Value$_{\beta 1}$	p-Value$_{\beta 2}$	p-Value$_{\beta 3}$	p-Value$_{\beta 4}$	p-Value$_{\beta 5}$	p-Value$_{\beta 6}$	p-Value$_{\beta 7}$
0,000280966	0,807413721	3,13228E-06	0,00739855	0,631876672	0,169944086	0,04036419	0
0,000792305	0,803456866	3,57725E-06	0,010326775	0,67556911	0,203341414	0,04031262	1,52672E-06

Note. In the partial test, the test statistics used are the Wald test (W).

The criterion for rejecting H0 is if W > χ2 or p-value < 0.05.

The hypothesis used for each variable is as follows:

H0: βi = 0 (the logit coefficient is not significant for the model).

H1: βi 0 (the logit coefficient is significant for the model).

To perform the Wald test, the following equation is used:

$$W = \frac{\widehat{\beta_\iota}}{se_{\widehat{\beta_\iota}}}$$

The W value follows a Chi-square distribution with df = 1.

Using p-value, there are 4 predictors that have p-value < α; i.e., age (x_2), body mass index (x_3), diastolic blood pressure (x_6) and type of diabetes (x_7) (in addition to the constant term, which is also significant

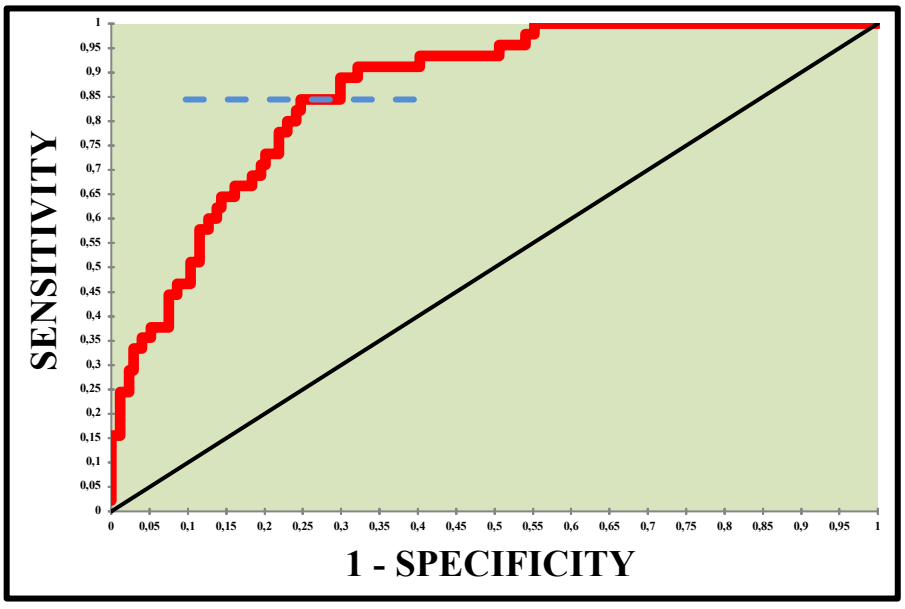
at 5%). According to those significant predictors, the logistic regression model can be seen in eq:

$$P[Y_i = 1] = p_i = \frac{e^{-8{,}264+0{,}084x_2+0{,}131x_3-0{,}0612x_6-85{,}21x_7}}{1+e^{-8{,}264+0{,}084x_2+0{,}131x_3-0{,}0612x_6-85{,}21x_7}}$$

The ROC curve (receiver operating characteristic curve) is shown in Figure 2. Based on this curve, it is possible to determine the best cut-off point for a diagnostic test for stroke, i.e., the criterion for predicting whether or not a subject will have a stroke.

Figure 2

ROC curve



Note: The intersection of the dotted line (blue) with the ROC curve (red) gives the cut-off point of the diagnostic test for stroke (the

threshold). If the area under the ROC curve is close to 1.00, then this test has a very good predictive ability for stroke.

Where the threshold is:

Threshold
0,2241074

Therefore, under a probability of 0.224 it can be assumed that a stroke will not occur, while above 0.224 it will be assumed that a stroke will occur (Fawcett, 2006).

And the area under the curve (AUC):

AUC
0,859

Therefore, this test does have a very good discriminatory ability to predict whether or not a person will suffer a stroke (Hanley & McNeil, 1982).

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. Four predictors have a significant effect in the model, namely age, body mass index, diastolic blood pressure, and type of diabetes.

It can be seen that an increase in age and body mass index will increase the risk of stroke. Age and body mass index was positively associated with stroke mortality (Yi et al., 2018), so it may be an early warning to prevent stroke.

Conclusions

Based on the analysis that has been done, it can be concluded that age and body mass index are two predictors that significantly affect the occurrence of stroke.

Diastolic blood pressure and the person's type of diabetes were also shown to have an impact on the prognosis of a future stroke.

The logistic regression model was fairly accurate in classifying 219 medical record data.

References

Abubakar, S. A., & Isezuo, S. A. (2012). Health-related quality of life of stroke survivors: experience from a stroke unit. *International Journal of Biomedical Science.* Kota Samarahan, Malaysia: IJBS. 8(3), 183.

Agresti, A. (2002). *Categorical data analysis, 3rd edition.* New York, NY, USA: John Wiley & Sons, Inc.

Allen, J. (2007). Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement.* Bristol, England: IOP Publishing Ltd. 28, R1-39.

Bootkrajang, J., & Kabán, A. (2014). Learning kernel logistic regression in the presence of class label noise. *Pattern recognition.* Amsterdam, The Netherlands: Elsevier B.V. 1-15. doi:10.1016/j.patcog.2014.05.007.

Bustan, M. N. (2017). *Manajemen Pengendalian Penyakit Tidak Menular.* Jakarta, Indonesia: Rineka Cipta.

Bustan, M. N., & Poerwanto, B. (2021). Logistic regression model of the relationship between breast cancer pathology diagnosis with metastasis. *Journal of Physics.*

Makassar, Indonesia: Conference Series. 1752(1), 1-5. doi:10.1088/1742-6596/1752/1/012026.

David, W. H., & Stanley, L. (2000). *Applied Logistic Regression.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Fa'rifah, R. Y., & Poerwanto, B. (2019). Penerapan Regresi Logistik dalam Menganalisisis Faktor Penyebab Peningkatan Angka Kematian. *Jurnal Ilmiah d'ComPutarE.* Kota Palopo, Indonesia. 9(1), 52-55.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters.* Amsterdam, The Netherlands: Elsevier B.V. 27(8), 861-874.

Garg, R., Rech, M. A., & Schneck, M. (2019). Stroke mimics: a major source of bias in acute ischemic stroke research. *Journal of Stroke and Cerebrovascular Diseases.* Amsterdam, The Netherlands: Elsevier B.V. 1-6.

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

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* Oak Brook, IL, USA: Radiological Society of North America. 143(1), 29-36.

Josephus, B. O., Nawir, A. H., Wijaya, E., Moniaga, J. V., & Ohyver, M. (2021). Predicting mortality in patients infected with COVID-19 virus based on observed patient characteristics using logistic regression. *Procedia Computer Science.* Amsterdam, The Netherlands: Elsevier B.V. 179(2019), 871-877.

Kementerian Kesehatan Republik Indonesia (2018). *Stroke Don't Be the One (2019).* Pusat Data dan Information Kementerian Kesehatan Republik Indonesia. Jakarta, Indonesia.

Liang, Y., Liu, G., Chen, Z., & Elgendi, M. (2017). *PPG-BP Database.* London, UK: Figshare. doi:10.6084/m9.figshare.5459299.

Recuero de los Santos, P. (December 13, 2021). *Blog: Telefónica Tech: How to interpret the confusion matrix: practical example.* Retrieved from Telefónica Tech Web site: https://telefonicatech.com/blog/como-interpretar-la-matriz-de-confusion-ejemplo-practico

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. *D'CARTESIAN.* Manado, Indonesia. 6(2), 56-62.

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.* Amsterdam, The Netherlands: Elsevier B.V. 270, 211-217.