Comparison of Logistic Regression and Support Vector Machine in Predicting Stroke Risk

Lensa Rosdiana Safitri [1], Nur Chamidah [2*], Toha Saifudin[2], Mochammad Firmansyah[3],and Gaos Tipki Alpandi[3]

[1]Mathematics Masters Study Program, Mathematics Department, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

[2]Mathematics Department, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

[2]Statistics Study Program, Mathematics Department, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

*Corresponding author: nur-c@fst.unair.ac.id

Received: 20 May 2024
Revised: 16 June 2024
Accepted: 19 June 2024

ABSTRACT –**The issue of health is the third goal of Indonesia's Sustainable Development Goals (SDGs) which is state to ensuring a healthy life and promoting prosperity for all people at all ages. One of the SDGs's concerns is deaths caused by non-communicable diseases (NCDs) including strokes. One prevention that can be done is by making a prediction of stroke for early detection. There are various methods available which are statistical methods and machine learning methods. In this research work, we aim to compare the two methods based on statistical method and machine learning method on stroke risk prediction. The data used in this research is primary data from Universitas Airlangga Hospital (RSUA) from June until August 2023. In this research, we compare the statistical method that is Logistic Regression (LR), and the machine learning method which is Support Vector Machine(SVM). We use Phyton to analyze all methods in this research. The results show that SVM with Radial Basis Kernel is better than LR in predicting stroke risk based on three goodness criteria namely sensitivity, F-1 score and aaccuracy where these three goodness criteria values of SVM are greater than those of LR.**

Keywords– **Stroke; Binary Logistic Regression; and Support Vector Machine**

I. INTRODUCTION

The health issue is one of the points of Indonesia's Sustainable Development Goals (SDGs) which are contained in goal number 3, That goal is to ensure a healthy life and promote prosperity for all people of all ages [1]. One of the points of concern to the SDGs in this sector is death from Non-communicable Diseases (NCD). World Health Organization (WHO) states that NCD including stroke causes 74% of all deaths worldwide. Stroke is a major source of disability and a major contributor to lost disability-adjusted life years, especially in low-income and middle-income countries [2]. Based on the World Stroke Organization report, there are more than 12.2 million new strokes every year. Globally, one in four people over the age of 25 will have a stroke in their lifetime [3]. WHO defines stroke as a symptom of a functional deficit of the nervous system caused by cerebrovascular disease.

The cause of stroke is due to changes in the nervous system caused by impaired blood circulation to parts of the brain that appear suddenly within seconds or symptoms and signs appear quickly within hours. The prevalence of stroke according to data from the World Stroke Organization shows that every year there are 13.7 million new cases of stroke, and around 5.5 million deaths occur due to stroke. WHO states that every year, there are more than 13.7 people worldwide who have a stroke. According to WHO, in 2018 there were 252,473 people or 14.83 percent of the total national death rate in Indonesia caused by stroke. Indonesia has the highest number 7th of deaths due to stroke in the world. Based on this fact, it is necessary to seriously prevent this disease. One of the preventions can be done with statistical and machine learning methods for early detection to prevent and reduce death from stroke by the SDGs target in the health sector.

There are many previous researches about predicting the risk of stroke. Comparison of SVM and binary logistic regression has been done by [4] that classify COVID-19 diagnostic data using data from the Kaggle website. This research results show that SVM accuracy is superior with an accuracy of 98.91% when compared to binary logistic regression with an accuracy of 95.64%. This research did not test bivariate and also did not produce the risk values needed for medical personnel and/or the government as a reference for making policies to overcome Covid-19.

In this research, we aim to compare two methods Logistic Regression and SVM, for predicting the risk of stroke. The comparison of those two methods is examined by the number of significant variables, accuracy, sensitivity, and specificity values. The results obtained from this study can be used to predict the chance of a person's risk of stroke as an effort to prevent stroke, especially in Indonesia.

II. LITERATURE REVIEW

A. Support Vector Machine

A classification method commonly used to form a binary non-probabilistic classification is the Support Vector Machine. Principally, in this Support Vector Machine method, the results of the training model are used to determine the best hyperplane for classifying data [5]. In classifying using SVM, it is necessary to have both training and testing stages. The main purpose of this SVM method is to determine the optimal classifier function that can be used to separate two different datasets [6]. If a hyperplane is caught in the middle of two objects from both classes, then that hyperplane is the best hyperplane or separator function. The formula for the linear kernel function in a Support Vector Machine is presented as follows:

$$wx^T + \gamma = 0. \quad (1)$$

This SVM manipulates the model to allow linear domain division. SVM can be divided into linear and nonlinear models [7]. There are many techniques of ML or data mining developed under assumptions of linearity. This results in the resulting algorithm also only limited to linear cases. Generally, cases that occur in the real world are not non-linear cases. To overcome this non-linearity, kernel methods can be used [8]

A kernel function is a function given the original feature vector, returning a value equal to the dot product of the corresponding feature vectors that are mapped. The feature vectors cannot be hidden in a higher dimensional space explicitly by the kernel function. Also, the dot product of the mapped vectors cannot be calculated by kernel function. The kernel returns the equal value using an unequal operations set which can frequently be calculated more efficiently. The essential reason we use kernel functions is to remove the need for processing to obtain a vector space with higher dimensions than a defined underlying space of vectors, which allows data to be detached linearly in higher dimensions. The following are kernel functions commonly used in SVM [7], [9].

1) Linear.

The linear kernel function is expressed as follows:

$$K(x, y) = x^T y. \quad (2)$$

This function obviously does not change native representation and does not get over the linearity constraints of linear classification and linear regression models. However, this allows the linearity of dot product-based algorithms (such as linear support vector machines and linear support vector regression algorithms) to be considered as special cases of suitable kernel-based algorithms.

2) Polynomial

Polynomial kernel functions, especially those with a degree of two, are widely used for classification purposes. The following is the formula for the polynomial kernel function:

$$K(x, y) = (\alpha x^T + y), \alpha > 0 \quad (3)$$

3) Radial Basis Function (RBF) (also called Gaussian).

The RBF or Gaussian kernel is the best choice for problems requiring non-linear models. A decision limit in the feature space that are mapped, namely a hyperplane, is similar to a decision limit in the genuine space, namely a hypersphere. The space of feature generated by the RBF or Gaussian kernel can have dimensions with infinite numbers, a feat that would have been unlikely otherwise. The RBF or Gaussian kernel function follows a formula as follows:

$$K(x, y) = exp(\alpha \|x - y\|), \alpha > 0 \quad (4)$$

4) Sigmoid

The sigmoid kernel function can be written as follows:

$$K(x, y) = tanh(\alpha x^T y + h) \quad (5)$$

The sigmoid function has gained popularity for the kernel approach because it is often used as the activation function for neural networks (multilayer perceptions). If we use it correctly, it will be similar to the family of RBF kernels. It can describe complex nonlinear interactions with multiple parameters. In some parameter configurations, it resembles an RBF kernel. This sigmoid kernel, however, probably does not really represent a suitable kernel for some parameters because it is not completely positive.

B. Binary Logistic Regression

The logistic regression model from the linear regression model is that the outcome variable has binary or dichotomous type [10]. The response is also supposed to be a theoretical sample of an underlying probability distribution [11]. A model's purpose is to estimate the true parameters of the model's underlying PDF based on the response as adjusted by its predictors. The response in logistic regression is binary (0,1) and follows a Bernoulli probability distribution. Because the Bernoulli distribution is a subset of the more general binomial distribution, logistic regression is included in the binomial family of regression model [12]. The form of the binary logistic regression probability model can be written:

$$E(Y|\boldsymbol{x}) = \pi(\boldsymbol{x}) \quad (6)$$

With:

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (7)$$

$\pi(\boldsymbol{x})$ is a non-linear function so to get a linear function, it is necessary to carry out a logit transformation [12]. The logit transformation form is as follows:

$$g(\mathbf{x}) = \text{logit}(\pi(\mathbf{x})) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1\mathbf{x} \quad (8)$$

$$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = e^{\beta_0+\beta_1\mathbf{x}} \quad (9)$$

$\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} = e^{\beta_0+\beta_1\mathbf{x}}$ in equation (9) is called the odd values. The odds value explains the sample's risk of experiencing a certain event (y=1). The odds value can range from 0 to infinity, where an odds value that is close to 0 means very low probability/risk of experiencing a certain event (y=1), meaning that the greater the odds value, the greater the probability/risk of experiencing a certain event [13].

III. METHODOLOGY

A. *Data Source and Research Variables*

In this study, we use primary data from Universitas Airlangga Hospital (RSUA) that we get from June to August 2023. The total of our samples is 200 which consist of stroke patients and non-stroke patients. For validation accuracy, we divide the dataset into two parts for each validation, 75% for training and 25% for testing. We use phyton to analyze all methods in this research. The factor variables we used in this research are referred to [14]. Table 1 provides variables of the models used in this study

Table 1. Research Variable

Variable			Category
Response	Y	Stroke	0: Non-Stroke ; 1: Stroke
Predictors	x_1	Obesity	0: No; 1: Yes
	x_2	Hypertension	0: No; 1: Yes
	x_3	Diabetes	0: No; 1: Yes
	x_4	Smoking Status	0: No; 1: Yes

B. *Step of Analysis*

The following steps were taken in analyzing the data:

1. Perform a statistic descriptive of the predictor variables associated with the incidence of stroke.
2. Predict stroke risk factors by using Logistic Regression methods.
3. Predict stroke risk factors by using SVM based on kernel linear, RBF, sigmoid, and polynomial.
4. Constructing the diagnostic Confusion Matrix for Table 2X2 for each model as follows:

Table 2. Confusion Matrix

Actual	Prediction	
	Positive	**Negative**
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

5. Calculating the performance values based on accuracy, sensitivity, and specificity, respectively, by using the following formula [15]

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\%, \quad \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \times 100\%, \text{ and}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \times 100\%$$

6. Comparing the performance results of Logistic Regression and SVM methods.
7. Providing conclusions about which model approach has the best performance for predicting the risk of stroke.

IV. RESULTS AND DISCUSSIONS

A. Descriptive Statistics

In this section, we give the results of this study, i.e., a description of variables and the comparison of logistic regression and SVM for predicting stroke risk. Table 3 shows the crosstabulation of predictor variables toward stroke as the response variable.

Table 3. Crosstabulation of Stroke Risk Factor

		Stroke				Total
		Non Stroke		Stroke		
		n	%	n	%	
Obesity	No	52	44.07%	66	55.93%	118
	Yes	48	58.54%	34	41.46%	82
Hypertension	No	39	53.42%	34	46.58%	73
	Yes	61	48.03%	66	51.97%	127
DM	No	77	57.89%	56	42.11%	133
	Yes	23	34.33%	44	65.67%	67
Smoking Status	No	81	48.50%	86	51.50%	167
	Yes	19	57.58%	14	42.42%	33
Total		100	50.00%	100	50.00%	200

Based on Table 3, in general, out of 200 people who have a stroke or not have different risks depending on personal health history factors. These factors include a history of obesity, hypertension, diabetes mellitus (DM), and smoking status. Of all these factors, the highest number of patients were at risk of stroke and had a history of hypertension, namely 66 people, followed by a history of DM, 43 people, a history of obesity, 34 people, and the lowest had a history of smoking, 14 people. Therefore, from these results, it can be seen that someone has different risk factors depending on their other history so they cannot justify the risk factors.

B. Model Comparison

Below are the results of the comparison between logistic regression and SVM models for the four kernels.

Table 4. Comparison of Performance of LR and SVM

	Specificity	**Sensitivity**	**Accuracy**
Logistic Regression (LR)	0.78	0.64	0.81
SVM (RBF)	0.75	0.82	0.84
SVM (Linear)	0.82	0.90	0.81
SVM (Poly)	0.88	0.70	0.74
SVM (Sigmoid)	0.67	0.60	0.55

Based on the accuracy level of the SVM model with various kernels, it is known that the RBF kernel is the best because it is able to produce an accuracy value of 0.84. Apart from that, the SVM model with a linear kernel can also be said to be quite good because it provides high specificity and sensitivity values of 0.82 and 0.90, respectively. Then, polynomial and sigmoid kernel SVMs do not provide good accuracy results when compared to RBF and linear kernel SVMs. This result is supported by the research of [16] that classify HDI using SVM that conclude that RBF kernel is the best kernel to solve HDI problem, with parameter combination cost= 1 and gamma=1 obtained classification accuracy of 98.1% which is the best classification accuracy. Furthermore, [17] also find that the accuracies of RBF-based SVM classifier for various crop types are relatively better than other two kernel functions.

For this reason, the SVM model with the RBF kernel is better than the logistic regression model because it has a higher accuracy and sensitivity value. Therefore, it can be concluded that the SVM model with the RBF kernel is the best compared to logistic regression and other SVM kernels. It concluded that the accuracy with SVM and GLM was relatively high.

C. Model Evaluation

The confusion matrix of SVM with RBF kernel that gives the best accuracy for predicting the stroke risk can be seen in Figure 1.

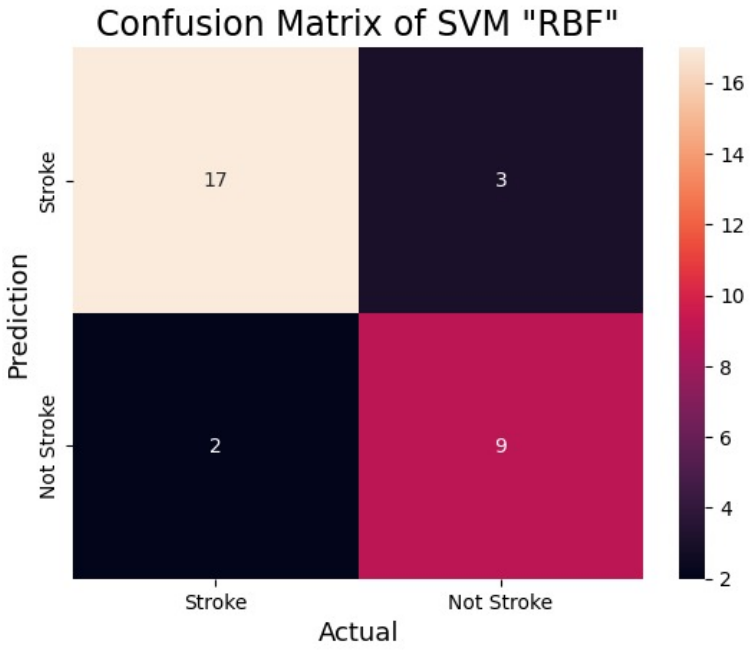

Figure 1. Confusion Matrix of SVM Based on RBF kernel

The confusion matrix of the SVM model based on the Radial Basis Function (RBF) kernel, as depicted in Figure 1, reveals the model's performance in the testing dataset. With a True Positive (TP) value of 17 and a False Negative (FN) value of 3, the model demonstrates its ability to accurately identify instances of successful payment by companies. However, it also misclassifies 2 instances of successful payments as failures (False Positives), while correctly identifying 9 instances of failed payments (True Negatives). This confusion matrix provides the basis for calculating key performance metrics such as specificity, sensitivity, and accuracy, which are essential for evaluating the model's effectiveness in distinguishing between payment outcomes. Overall, the analysis of the confusion matrix indicates a promising performance of the RBF-SVM model in predicting company payment outcomes, albeit with a slight tendency towards misclassification of successful payments.

V. CONCLUSIONS AND SUGGESTIONS

In summary, the study examined stroke risk factors among 200 individuals, highlighting hypertension as the most prevalent, followed by diabetes mellitus, obesity, hypertension, and smoking. The research underscored the multifaceted nature of stroke risk, emphasizing the importance of a comprehensive assessment approach considering various health history factors. Additionally, the study evaluated SVM models with different kernels for predicting stroke occurrence, with the Radial Basis Function (RBF) kernel emerging as the most effective, achieving an accuracy of 0.84. The study's findings contribute to advancing predictive modeling techniques in healthcare risk management, offering insights into enhancing risk assessment practices. In conclusion, the research underscores the intricate interplay of health history factors in determining stroke risk and highlights the superior predictive performance of the RBF kernel in SVM models.

To enhance prediction accuracy, it is advised to broaden data coverage by incorporating external factors like age, dietary habits, ancestry, and others, which can influence stroke risk. Furthermore, evaluating prediction models using additional metrics like Receiver Operating Characteristic (ROC) curves is essential. Exploring alternative classification models such as gradient boosting, random forest, or decision trees can aid in identifying the most effective model. Furthermore, the next research can implement widen data source to be used for worldwide or national scale society. Collaboration with healthcare professionals and statistical analysts is recommended for deeper insights into prediction outcomes, facilitating the development of a more effective early warning system for stroke risk.

ACKNOWLEDGMENT

The authors would like to thank Universitas Airlangga for supporting the process of paper writing by providing funds so that the researcher can take part in the research activity. Appreciation is also given to Universitas Airlangga Hospital (RSUA) for providing research data, as well as all those who have provided support in the research process and this publication.

REFERENCES

[1] Bappenas. Tujuan Pembangunan Berkelanjutan [Internet]. 2022. Available from: https://sdgs.bappenas.go.id/tujuan-3/
[2] WHO. World Stroke Day. 2021. Available from: https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day
[3] Feigin VL, Brainin M, Norrving B, Martins S, Sacco RL, Hacke W, et al. World Stroke Organization (WSO): global stroke fact sheet 2022. International Journal of Stroke. 2022;17(1):18–29.
[4] Rismawati Y, Tirta IM, Dewi YS. Klasifikasi Data Diagnosis Covid-19 Menggunakan Metode Support Vector Machine (Svm) Dan Generalized Linear Model (Glm). UNEJ e-Proceeding. 2022;246–52.
[5] Tripathy A, Agrawal A, Rath SK. Classification of Sentimental Reviews Using Machine Learning Techniques. Procedia Computer Science. 2015;57:821–9.
[6] Dangeti P. Statistics for machine learning. Packt Publishing Ltd; 2017.
[7] Suthaharan S. Support Vector Machine-Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems. 2016;36.
[8] Schölkopf B, Smola A. Support vector machines and kernel algorithms. In: Encyclopedia of Biostatistics. Wiley; 2005. p. 5328–35.
[9] Kernel methods. In: Data Mining Algorithms [Internet]. Chichester, UK: John Wiley & Sons, Ltd; 2015 [cited 2023 Feb 13]. p. 454–97. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781118950951.ch16
[10] Hosmer DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Third edition. Hoboken, New Jersey: Wiley; 2013. 1 p. (Wiley series in probability and statistics).
[11] Stoltzfus JC. Logistic Regression: A Brief Primer: LOGISTIC REGRESSION: A BRIEF PRIMER. Academic Emergency Medicine. 2011 Oct;18(10):1099–104.
[12] Hilbe JM. Practical guide to logistic regression. crc Press; 2016.
[13] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. Vol. 112. Springer; 2013.
[14] Joundi RA, Patten SB, Williams JV, Smith EE. Vascular risk factors and stroke risk across the life span: A population-representative study of half a million people. International Journal of Stroke. 2022;17474930211070682.
[15] Piegorsch WW. Confusion Matrix. Wiley StatsRef: Statistics Reference Online. 2014;1–4.
[16] Al Azies H, Trishnanti D, Mustikawati P.H E. Comparison of Kernel Support Vector Machine (SVM) in Classification of Human Development Index (HDI). IJPS. 2019 Dec 30;0(6):53.
[17] Yekkehkhany B, Safari A, Homayouni S, Hasanlou M. A Comparison Study Of Different Kernel Functions For Svm-Based Classification Of Multi-Temporal Polarimetry Sar Data. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2014 Oct 22;XL-2/W3:281–5.

© 2024 by the authors. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (http://creativecommons.org/licenses/by-sa/4.0/).