Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2025, 2025:31

https://doi.org/10.28919/cmbn/9099

ISSN: 2052-2541

MODELING OF STROKE RISK USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE IN MULTIVARIATE ADAPTIVE REGRESSION SPLINE MODEL

LENSA ROSDIANA SAFITRI[1,2], NUR CHAMIDAH[2,*], TOHA SAIFUDIN[2]

[1]Mathematics Master Study Program, Department of Mathematic, Faculty of Science and Technology, Airlangga University, Surabaya 60115, Indonesia

[2]Mathematics Department, Faculty of Science and Technology, Airlangga Universitty, Surabaya 60115, Indonesia

Copyright © 2025 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: The health sector represents the third goal of Indonesia's Sustainable Development Goals (SDGs), which is focused on ensuring healthy lives and promoting well-being for individuals of all ages. Stroke is one of the leading causes of mortality globally and is classified as a non-communicable disease (NCD). Early detection and accurate risk prediction are essential to prevent stroke occurrences and reduce related mortality rates. This study evaluates the impact of the Minority Oversampling Technique (SMOTE) on Multivariate Adaptive Regression Splines (MARS) models in predicting stroke risk, particularly in addressing imbalanced datasets. The data used in this study was collected from Universitas Airlangga Hospital (RSUA) between June and August 2023. To assess the effectiveness of SMOTE, we compare the performance of MARS models with and without oversampling. The results show SMOTE- MARS achieves higher accuracy compared to MARS model (93.50% vs. 89.00%), sensitivity (97.70% vs. 94.50%), specificity (80.70% vs. 73.97%), and AUC (89.20% vs. 84.23%), These results underscore the importance of addressing class imbalance in stroke prediction datasets to achieve more accurate and reliable outcomes. Incorporating SMOTE into MARS models proves to be a highly effective approach for enhancing predictive performance, offering a valuable tool for early stroke risk detection and prevention strategies.

[*]Corresponding author

E-mail address: nur-c@fst.unair.ac.id

Received December 29, 2024

Keywords: Stroke; SMOTE; SMOTE-MARS.

2020 AMS Subject Classification: 62G08, 62P10, 65D07, 65D10.

1. INTRODUCTION

Health issue is one of the points of Indonesia's Sustainable Development Goals (SDGs) which are contained in goal number three, That goal is to ensuring a healthy life and promoting prosperity for all people at all ages. 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 cause 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 [1]. 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 [2]. 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 [1]. WHO states that every year, there are more than 13.7 people worldwide have a stroke. Based on WHO, in 2018 there were 252,473 people, or 14.83 percent of the total national death rate in Indonesia caused by stroke. Based on this fact, it is necessary to seriously prevent this disease. One of the preventions can be done with statistical and machine learning method for early detection to prevent and reduce death from stroke in accordance with the SDGs target in the health sector.

Statistical modeling is one approach for early detection of stroke and for analyzing the factors influencing stroke occurrence. In general, statistical modeling is a simplified concept derived from theory, commonly used in science and technology disciplines to study the relationships between real-life phenomena. One statistical modeling method with a categorical response variable is binary logistic regression. Research using logistic regression to model stroke risk has been widely conducted. There are many previous researches about modelling the risk of stroke using statistical model aproach. A research on modelling risk of stroke using logistic regression has been done by [3] and [4]. modeled stroke risk using logistic regression with 5,411 data points and analyzed 22 factors, nine of which were found to be significant:

hypertension, diabetes mellitus, atrial fibrillation, congestive heart failure, previous stroke, previous transient ischemic attack, hyperlipidemia, smoking habits, and snoring. This study achieved a classification accuracy of 94.1%.

However, binary logistic regression has limitations, particularly its inability to account for interactions between predictor variables, a factor often observed in real-world data. To address this issue, Multivariate Adaptive Regression Splines (MARS) offers an effective alternative. MARS is a nonparametric regression method capable of capturing complex interactions among independent variables, handling high-dimensional data, and accommodating intricate data structures [5]. Another research on modeling the risk of stroke using MARS has been done by [6] where the response variables were ischemic and hemoragic stroke.

There are also several previous studies that compare two methods namely logistic regression and MARS such as conducted by [7] and [8]. According to those several previous studies, MARS gives better performance than Logistic Regression in terms of accuracy. MARS as one of nonparametric regression approaches is able to provide greater flexibility because the form of the estimation of the regression function will adjust to the pattern of the data without being influenced by the subjectivity of the researcher. The research [9] comparing MARS with binary logistic regression (BLR) demonstrated that MARS outperformed BLR in stroke risk estimation. The MARS model achieved an accuracy of 93.5% on both training and testing datasets, whereas the BLR model produced accuracies of 91.3% for training data and 89% for testing data. This highlights the superior performance of MARS in capturing stroke risk more effectively.

The issue of imbalanced data is a critical challenge in predictive modeling, particularly when the proportion of one class far outweighs the other. The research [10] highlighted a disparity in child labour studies, where 96% of observations belonged to the non-child labor class, while only 4% represented child labour. This severe imbalance affected the analysis, resulting in a sensitivity of only 31.5% for identifying child labour cases. Such disparities exemplify the phenomenon of imbalanced data in data mining, where minority classes are often underrepresented, leading to biased model performance. To address this, future research has emphasized the need for pre-analysis techniques to handle imbalanced data effectively. One promising approach is the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic data for the minority class to achieve class balance. By augmenting the data, SMOTE improves the representation of minority classes and enhances the robustness of predictive analysis.

In this study, we aim to model stroke risk using the Multivariate Adaptive Regression Splines (MARS) technique in combination with the Synthetic Minority Oversampling Technique (SMOTE) to address the class imbalance in stroke data. This approach will be evaluated based on the number of significant variables, accuracy, sensitivity, and specificity. The results of this study will help predict an individual's stroke risk, supporting early detection efforts and contributing to stroke prevention, particularly in Indonesia.

2. MULTIVARIATE ADAPTIVE REGRESSION SPLINES

MARS is a type of nonparametric regression analysis aimed primarily at predicting a response variable influenced by one or several predictor variables, without assuming any underlying functional relationship between the response and predictor variables. One advantage of the MARS method, compared to others, is that it can examine all possible levels of interaction between independent variables. Therefore, the MARS model can handle high-dimensional datasets and complex structures observed in data points. MARS can estimate models with continuous or binary response variables [5].

Parameters in MARS must be carefully tuned to avoid over fitting. The degree parameter represents the maximum degree of interaction, while the nprune parameter represents the maximum number of terms allowed in the model. Lower degree interactions help interpret the final model, while higher degree interactions may occasionally lead to prediction inconsistencies. Three possible degree values are 3, 2, and 1. The guideline for nprune is that it should be $\geqslant 2$ and less than "nk", where "nk" is the "maximum number of model terms before pruning, calculated using the formula [11]:

$$nk \ = \ min(200, \ max(20, \ 2 * ncol\left(x\right))) \ + \ 1$$

The MARS model for a binary response variable can be expressed in the following equation:

$$\text{logit}\,\pi\left(x\right)=\ln\left(\frac{\pi\left(x\right)}{1-\pi\left(x\right)}\right)=\alpha_0+\sum_{m=1}^{M}\alpha_m\prod_{k=1}^{k_m}S_{Km}.\left(x_{v(k,m)}-t_{Km}\right)$$

That can be written in matrix form as follows:

$$\text{logit}\,\pi\left(x\right)=\boldsymbol{\beta\alpha}$$

With

$$\boldsymbol{\beta}=\begin{bmatrix} 1 & \prod_{k=1}^{K_1} S_{1m}(x_{1(1,m)}-t_{1m}) & \cdots & \prod_{k=1}^{K_M} S_{Mm}(x_{1(M,m)}-t_{Mm}) \\ 1 & \prod_{k=1}^{K_1} S_{1m}(x_{2(1,m)}-t_{1m}) & \cdots & \prod_{k=1}^{K_M} S_{Mm}(x_{2(M,m)}-t_{Mm}) \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 1 & \prod_{k=1}^{K_1} S_{1m}(x_{n(1,m)}-t_{1m}) & \cdots & \prod_{k=1}^{K_M} S_{Mm}(x_{1(M,m)}-t_{Mm}) \end{bmatrix},\quad \boldsymbol{\alpha}=(\alpha_0,\alpha_1...,\alpha_m)^T$$

The parameter is estimated using the maximum likelihood method. The response variable, which has two categories, has a cutoff point of 0.5. The prediction rule is as follows: if $\pi(x)\geq$ 0,5 the predicted outcome is and if $\pi(x)<$ 0,5 the predicted outcome is 0.

3. The Synthetic Minority Oversampling Technique

An imbalanced dataset refers to a condition in a classification task where the proportion of labels is significantly skewed. The label with the larger proportion is called the majority class, while the other is referred to as the minority class [12]. SMOTE is an approach where the minority class is oversampled by creating synthetic data. Several studies have mentioned that SMOTE can improve the accuracy of classifiers for the minority class [13]. The SMOTE approach works by replicating minority class data, known as synthetic data. The method works by finding the k-nearest neighbors for each data point in the minority class, then generating synthetic data according to the desired percentage of oversampling, where the k-nearest neighbors are chosen randomly [14]. SMOTE creates synthetic data based on the distance between a minority data point and its nearest minority neighbor, so the new synthetic data is located between the two minority points. The nearest neighbors are selected based on the Euclidean distance between the data points. The Euclidean distance between two vectors is the square root of the sum of the squared differences between each element of the vectors. The formula to generate synthetic data using SMOTE is as follows [14]:

$$D_{\text{new}}=D_i+(\widehat{D}-D_i)\times\delta$$

Let D_{new} be synthetic data. D_i be the minority data to be replicated, $\widehat{D}$ be the data closest to D_i and δ be a random number between 0 and 1.

4. MAIN RESULTS

4.1 DESCRIPTIVE STATISTICS

Descriptive statistics summarize and present a dataset effectively. For categorical data, descriptive statistics typically include frequency and percentage [15]. The tables below describe each predictor variable in relation to the response variable using cross-tabulations. Cross-tabulation is a procedure to present the frequency distribution of two or more categorical variables simultaneously.

Table 1 Cross tabulation of Stroke Risk Factor

		Stroke				Total
		Non Stroke		Stroke		
		n	%	n	%	
Obesity	No	67	81.70%	15	18.30%	82
	Yes	58	85.30%	10	14.70%	68
Hipertension	No	47	78.60%	4	21.40%	51
	Yes	78	75.20%	21	24.80%	99
DM	No	98	90.74%	10	9.26%	108
	Yes	27	64.29%	15	35.71%	42
Smoking Status	No	101	84.90%	18	15.10%	119
	Yes	24	77.40%	7	22.60%	31
Total		125	83.33%	25	16.67%	150

The Table 1, provides an overview of the distribution of stroke occurrences based on several risk factors, including obesity, hypertension, diabetes mellitus (DM), and smoking status. Among individuals without obesity, 18.3% experienced a stroke, while 14.7% of those with obesity had a stroke, indicating a slightly lower stroke occurrence among those classified as obese. Regarding hypertension, individuals with hypertension showed a higher proportion of stroke occurrences (24.8%) compared to those without hypertension (21.4%). Diabetes mellitus showed a more pronounced association with stroke. Among individuals with DM, 35.71% experienced a stroke, significantly higher than the 9.26% observed in individuals without DM, emphasizing diabetes as a critical risk factor. Smoking status also appears to influence stroke occurrence, as 22.6% of smokers experienced a stroke compared to 15.1% of non-smokers. Factors such as hypertension, diabetes mellitus, and smoking status appear to have a stronger association with stroke, highlighting the importance of addressing these risk factors in stroke prevention efforts.

Overall, the total prevalence of stroke in the dataset is 16.67%. According [9] , datasets with a minority class proportion of less than 20% fall into the moderate imbalance category. Given the 17% minority class proportion, this dataset is classified as moderately imbalanced. Addressing this imbalance is crucial to prevent misclassification in subsequent modeling steps. To handle this issue, the Synthetic Minority Oversampling Technique (SMOTE) is employed to generate synthetic data for the minority class.

4.2 Comparison Between SMOTE MARS and MARS Models

The comparison of the accuracy, sensitivity, and specificity values for training and testing data between SMOTE MARS and MARS models can be seen in Table 2.

Table 2. Comparison of Performance of the SMOTE MARS and MARS Models

Performance Criteria	MARS	SMOTE MARS
Accuracy	89.00%	93.50%
Sensitivity	94.50%	97.70%
Specificity	73.97%	80.70%
AUC	84.23%	89.20%

Based on Table 2, SMOTE MARS demonstrates superior performance compared to the standard MARS model in all criteria, both for training and testing datasets. SMOTE MARS achieves higher accuracy, sensitivity, and specificity, which highlights its effectiveness in handling imbalanced datasets. Furthermore, the application of SMOTE enhances the ability of the MARS model to detect minority classes (strokes) while maintaining robust predictive performance for the majority class. This improvement is particularly evident in the increased specificity values, indicating better classification of negative stroke cases without compromising the sensitivity. In addition, SMOTE MARS retains the advantages of the MARS model, including its capacity to explore interactions between predictor variables. The improved performance of SMOTE MARS confirms that addressing class imbalance is a crucial step in enhancing the predictive accuracy of stroke risk models.

4.3 The Modeling of Stroke Risk using SMOTE-MARS model

All analyses in this study were conducted using R software. Stroke risk modeling using MARS began with parameter tuning. The MARS model adjustment involved selecting appropriate values for parameters such as degree and nprune. The results showed that the combination yielding the minimum RMSE was degree 2 and nprune 8. The next step identified the best SMOTE-MARS model

$$\hat{\pi}(\mathbf{x}) = \frac{e^{\hat{f}(\mathrm{x})}}{1+e^{\hat{f}(\mathrm{x})}}$$

where $\hat{f}(x) = \text{logit}\left[\hat{\pi}(\mathbf{x})\right] = -1.06 + 2.446\left(\text{Hipertensi* DM}\right)$

A significant interaction was found between Hypertension and Diabetes Mellitus (DM), with a coefficient value of 2.446. This means that the interaction between the hypertension and diabetes variables contributes a relatively large positive coefficient. It indicates that the simultaneous presence of hypertension and diabetes significantly increases the log-odds of stroke. In other words, the combined effect of hypertension and diabetes is greater than the individual effects of each condition separately. The SMOTE-MARS model provides a more flexible approach to capturing nonlinear patterns and interactions among predictor variables in explaining stroke risk variability. The interpretation of coefficients helps in understanding the relative contributions of each variable and their interactions to the prediction of stroke risk. Based on the MARS output, the most important variables in the model, in order, are Hypertension, DM, and Obesity, while smoking status is considered insignificant.

CONCLUSION

In In conclusion, addressing the data imbalance is essential to avoid misclassification in the next modeling steps. SMOTE MARS demonstrates superior performance compared to the standard MARS model. SMOTE MARS achieves higher accuracy (93.50% vs. 89.00%), sensitivity (97.70% vs. 94.50%), specificity (80.70% vs. 73.97%), and AUC (89.20% vs. 84.23%), which highlights its effectiveness in handling imbalanced datasets. Furthermore, the application of SMOTE enhances the ability of the MARS model to detect minority classes (strokes) while maintaining robust predictive performance for the majority class. This improvement is particularly evident in the increased specificity values, indicating better classification of negative stroke cases without compromising the sensitivity.

In addition, SMOTE MARS retains the advantages of the MARS model, including its capacity to explore interactions between predictor variables. The improved performance of SMOTE MARS confirms that addressing class imbalance is a crucial step in enhancing the predictive accuracy of stroke risk models.

ACKNOWLEDGMENTS

The authors thank the directorate of research, technology, and community service (direktorat riset, teknologi, dan pengabdian kepada masyarakat–DRPTM), the ministry of education, culture, research and technology, the republic of indonesia through the Masters Thesis Research Grant (hibah Penelitian Tesis Magister–PTM) with a master contract number: 1748/B/UN3.LPPM/PT.01.03/2024.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] V.L. Feigin, M. Brainin, B. Norrving, et al. World Stroke Organization (WSO): Global Stroke Fact Sheet 2022, Int. J. Stroke 17 (2022), 18–29. https://doi.org/10.1177/17474930211065917.

[2] WHO, World Stroke Day, 2021. https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day.

[3] R.D.L.N. Karisma, S. Harini, Multivariate Adaptive Regression Spline in Ischemic and Hemorrhagic Patient (Case Study), AIP Conf. Proc. 2084 (2019), 020003. https://doi.org/10.1063/1.5094267.

[4] M. Gholam Azad, J. Pourmahmoud, A.R. Atashi, et al. Predicting of Stroke Risk Based on Clinical Symptoms Using the Logistic Regression Method, Int. J. Ind. Math. 14 (2022), 209-218. https://doi.org/10.30495/ijim.2022.64325.1559.

[5] A.H. Naser, A.H. Badr, S.N. Henedy, K.A. Ostrowski, H. Imran, Application of Multivariate Adaptive Regression Splines (MARS) Approach in Prediction of Compressive Strength of Eco-Friendly Concrete, Case Stud. Constr. Mater. 17 (2022), e01262. https://doi.org/10.1016/j.cscm.2022.e01262.

[6] R. Lu, T. Duan, M. Wang, et al. The Application of Multivariate Adaptive Regression Splines in Exploring the Influencing Factors and Predicting the Prevalence of HbA1c Improvement, Ann. Palliat. Med. 10 (2021), 1296–1303. https://doi.org/10.21037/apm-19-406.

[7] S. Park, S.Y. Hamm, H.T. Jeon, J. Kim, Evaluation of Logistic Regression and Multivariate Adaptive Regression Spline Models for Groundwater Potential Mapping Using R and GIS, Sustainability 9 (2017), 1157. https://doi.org/10.3390/su9071157.

[8] A. Wibowo, M.R.Ridha, Comparison of Logistic Regression Model and MARS Using Multicollinearity Data Simulation, J. Teor. Apl. Mat. 4(2020), 39-48.

[9] L.R. Safitri, N. Chamidah, T. Saifudin, Modeling Risk of Stroke Using Binary Logistic Regression and Multivariate Adaptive Regression Splines, AIP Conf. Proc. 3201 (2024), 060008. https://doi.org/10.1063/5.0230694.

[10] D. Adiangga, H. Wijayanto, B. Sartono, Multivariate Adaptive Regression Spline (MARS) for Modelling of Child Labor in Jakarta, in: Proceeding of International Conference on Research, Implementation and Education of Mathematics and Sciences, pp. 183-190, (2015).

[11] M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer, New York, 2013. https://doi.org/10.1007/978-1-4614-6849-3.

[12] V. Kumar, G.S. Lalotra, P. Sasikala, et al. Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques, Healthcare 10 (2022), 1293. https://doi.org/10.3390/healthcare10071293.

[13] N.V. Chawla, K.W. Bowyer, L.O. Hall, et al. SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res. 16 (2002), 321–357. https://doi.org/10.1613/jair.953.

[14] B. Santoso, H. Wijayanto, K.A. Notodiputro, B. Sartono, Synthetic Over Sampling Methods for Handling Class Imbalanced Problems: A Review, IOP Conf. Ser.: Earth Environ. Sci. 58 (2017), 012031. https://doi.org/10.1088/1755-1315/58/1/012031.

[15] T.K. Naab, Statistics, Descriptive, in: J. Matthes, C.S. Davis, R.F. Potter (Eds.), The International Encyclopedia of Communication Research Methods, 1st ed., Wiley, 2017: pp. 1–5. https://doi.org/10.1002/9781118901731.iecrm0241.