Proceedings of the 2nd International Conference on Modern Medicine and Global Health
DOI: 10.54254/2753-8818/17/20240623

Predicting the risk of stroke based on machine learning

Jingyang Jiang

The High School Affiliated to Renmin University of China, 37 Zhonguancun Road, Beijing, 100080, China

jjiang62@jh.edu

Abstract. Stroke, the second largest leading cause of death among all chronic diseases, is affecting about 101 million people in the world currently. It is estimated that this number of stroke cases will increase by 2.25 by the year 2050. Considering the large number of potential patients with stroke, a mathematical model is designed to predict one's risk of having a stroke in the future based on one's basic health data using machinery methods. Using the algorithm of Logistic Regression, the model reaches an accuracy of 92.28% when predicting whether one has a stroke, the model also validates that hypertension is the leading cause of the incidence of stroke by finding out the highest correlation value among all the feature variables. People who would like to know their probability of having a stroke can use the model, then they can have some precautionary measures to lower the likelihood of happening of stroke based on the prediction given, which helps save the medical costs and overuse of medical resources. Governments can enact policies and allocate medical resources based on the predictions made by the model.

Keywords: Stroke, Prediction, Probability.

1. Introduction

According to the World Stroke Organization, it is estimated that around 101 million people are living with stroke worldwide. Based on a study by the World Health Organization, among all chronic diseases that cause death, stroke ranks second [1]. There are around 4.5 million deaths a year from stroke worldwide [2]. Moreover, the forecasts conducted by George Howard and David C. Goff point out that by the year 2050, the number of stroke cases will increase by 2.25 times [3]. The incurrence of stroke causes death and disabilities, such as impaired speech, restricted physical abilities, paralysis, and so forth. A specific example is Louis Victor Leborgne, who could only pronounce "Tan" due to the impairment of Broca's area, a part of the cerebral cortex, which regulates speech production caused by stroke [4]. Due to the large number of potential stroke patients and the severity of getting a stroke, it is essential to develop a mathematical model to predict people's risk of having a stroke in the future, to inform those who are likely to have a stroke to take some precautions to lower the probability of having stroke. Moreover, the predictive model can find out certain patterns and characteristics of stroke cases. In this way, people can prevent the incurrence of stroke more precisely and specifically by changing their lifestyles or other methods.

According to historical data analysis, it is confirmed that there are several risk factors for stroke. The World Stroke Organization has claimed that the ten main risk factors (in descending order) of stroke are high systolic blood pressure, high body index mass, high fasting glucose, air pollution, smoking,

DOI: 10.54254/2753-8818/17/20240623

unbalanced diet, high cholesterol level, kidney dysfunction, alcohol, and low physical activity. Among these risk factors, previous research conducted by the World Stroke Organization suggested that high systolic blood pressure is the most important risk factor for stroke [5]. In the mathematical model, it is expected to find the correlations between feature variables and stroke, to confirm the correctness of the leading stroke risk factors.

Even though the modern medical system has already developed advanced technologies that inspect whether one is suffering from stroke based on various symptoms and scanning [6], there is a lack of detection of stroke before the occurrence of symptoms. Therefore, a research gap exists in that people have few approaches to check whether they will have a stroke until the symptoms of stroke develop. The research and the model are helpful for potential patients of stroke to check the risk and probability of having a stroke in the future. It is both beneficial in economical and humanistic aspects because for the former, precautions for stroke will cost less than having treatment after symptoms develop; for the latter, people will be less likely to suffer from the symptoms of stroke, such as paralysis, restricted physical abilities and so forth, which can be mentally destructive to patients. Besides, governments can make medical planning and allocate medical resources previously based on the data reported by citizens.

2. Methods

2.1. Data Sources

The dataset of stroke and its risk factors are acquired from Kaggle, which provides a large number of resources and datasets about stroke available for machine learning and data analysis [7]. The dataset contains over 40,000 cases of individuals who either are suffering from a stroke or are not being examined as having the symptoms of a stroke, which is recorded as 1 (have a stroke) or 0 (do not have a stroke) in the dataset. The stroke serves as the target variable in the model. The dataset also provides information on an individual's sex, age, whether having hypertension or heart disease, marital status (1 for married and 0 for not being married), work type (0 for never worked, 1 for children, 2 for government job, 3 for self-employed, 4 for private), residence type (1 for living in urban, 0 for living in rural area), average glucose level, BMI (Body Mass Index), and smoking status (1 for smokes, 0 for never smoked). The 10 indicators mentioned above are served as feature variables.

2.2. Statistical Analysis

Without manipulation of the dataset initially, the model cannot be accurate enough. It is learned that there is no feature scaling of the dataset that may cause a single variable to dominate the machine learning algorithm or introduce biases due to the difference in scales of each variable. To be more specific, the numeric values of age and average glucose level are relatively large, while the number representing variables such as hypertension, heart disease, and smoking status is 0 or 1, which is comparatively small. Therefore, feature scaling is introduced to avoid the magnitude differences between 10 feature variables. In this way, the numeric stability is improved in the model. Missing data is also excluded from the dataset and the sex variable is converted into a factor to represent it as a categorical value instead of a numerical one. By doing so, the model learned that 0 and 1 in sex represent different groups instead of continuous numerical values.

In the mathematical model, the method of Logistic Linear Regression is introduced. This method is appropriate when the target variable is binary [8], which just corresponds to the response of stroke (yes or no). In Logistic Regression, it can show the probability ρ_i of the happening of a certain thing. It is calculated that the equation for the logistic regression model with only 1 predictor X is:

$$\log\left(\frac{\rho_i}{1-\rho_i}\right) = \beta_0 + \beta_1 X$$

Similarly, by increasing the number of predictors to 10 in our model, comes out the equation:

$$\log\left(\frac{\rho_i}{1-\rho_i}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10}$$

DOI: 10.54254/2753-8818/17/20240623

Therefore, solving for ρ_i will give out:

$$\rho_i = \frac{e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10}}}{1 + e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10}}}$$

Because the dataset is relatively large, only 10% of the data is chosen to be used as the training dataset. With the input of different values, the model creates a best-fitted line to predict the probability of having a stroke. Then, a threshold is set which determines the final predicted outcome will be 0 or 1. If the goal is to lower the rate of missing cases that do have a stroke, the threshold should be set lower to make it more sensitive; if predicting a healthy person being sick is considered to avoid, the threshold can be set higher even though it may miss some patients who need medical assistance. All the analyses were undertaken in R version 4.2.1.

However, it is still not very possible to get ideal results without weighing each variable, because every feature variable has a distinct influence and correlation with the target variable—stroke. In order to find the correlation between feature variables and target variables, a corrplot—a visualization of a correlation matrix that can show the correlation between every single variable—is designed. The correlation coefficient of stroke and other feature variables will be shown in the corrplot. Based on the correlation coefficients, each variable in the Logistic Regression is weighted to reflect their importance to stroke, respectively.

2.3. Outcome Measures

As mentioned before, 10% of our data is chosen to be the train set; accordingly, the rest 90% of the data is being used as test data to see the accuracy of the predictive model. The confusion matrix is decided to be the indicator of accuracy. It is a 2*2 matrix in which rows represent actual values, while columns represent predicted values. It contains 4 elements, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), and they are shown in Table 1. TP and TN mean the model does the correct prediction. However, due to the sensitivity of the model, the model either has more FP or more FN. To be more specific, having more FP means that the model predicts a healthy individual having a stroke, which may cause waste in medical resources and it is a torment for the individual to be meaninglessly treated; having more FN means that the model predicts a patient of stroke to be healthy and let him go without any medical treatment. This is severe because it may cause life loss of the patient. To make the confusion matrix more detailed, the model performance index including accuracy, precision, recall, and F1-score is added.

Table 1. Confusion matrix. Actual values.

	Negative	**Positive**
Negative	TN	FN
Positive	FP	TP

3. Results

Before multiple handling of the dataset and model was given, the accuracy and overall performance of the predictive model were not quite ideal. With simple logistic linear regression, the initial result of the confusion matrix is shown in Table 2. There is only about 66% accuracy for the model, and there are 8720 cases of FN and 4451 cases of FP, which means that the model misses a large number of patients and fails to cure them of stroke; it is also asking 4451 patients to be careful that they may get stroke in the future, the precautionary measures may cause meaningless financial loss and mental distress. To avoid the happening of this situation and improve the accuracy of the model, a lot of measures of amelioration for the model have been done.

DOI: 10.54254/2753-8818/17/20240623

Table 2. Confusion Matrix of Initial Model.

	Negative	**Positive**
Negative	15019	4451
Positive	8720	10626

A corrplot is introduced to weigh each feature variable of the model. The corrplot gives our numeric values that can be used as a reference when weighing the variables [9]. According to the outcome, stroke has a negative correlation with sex. The negative correlation means women are more likely to suffer from stroke than men. As for all other 9 feature variables, stroke has a positive correlation with them, which indicates that aging, hypertension, cardiovascular diseases, urban living lifestyle, average blood sugar level, high BMI level, and smoking, may promote the occurrence of stroke. However, some of the feature variables present a relatively weak correlation with stroke, such as marital status, work type, and residence type; their correlation coefficients are only around 0.1, so lower weights were given to them. Other variables such as hypertension, heart disease, average glucose level, and smoking status show a moderately strong correlation with the target variable; correlation coefficients are about 0.6 to 0.7 in these feature variables, thus higher weights were given to them in the logistic regression model. After preprocessing the dataset and weighting feature variables, the performance of the predictive model becomes. The accuracy, precision, recall, and f1-score all increase to around 90%.

Table 3. Corrplot of Correlation Coefficient between stroke and feature variables.

Corrplot	**Stroke**
sex	-0.29406
age	0.2701
hypertension	0.73423
ever_married	0.18432
work_type	0.06903
residence_type	0.06323
avg_glucose_level	0.59497
BMI	0.29483
smoking_status	0.53401
heart_disease	0.65091
stroke	1

The model's threshold of deciding whether one is suffering from a stroke is also adjusted slightly several times to see how accuracy and the confusion matrix change with the adjustment of that threshold. Specifically, the line is modified to 0.45, 0.5, 0,55, and 0.6. It is finally found that 0.5 is the best value for the threshold because it has the highest accuracy among others, and the number of FN is also the lowest. Therefore, the threshold is changed from previously 0.4 to now 0.5. The final accuracy reaches 92.28% under the threshold of 0.5. (Table 4)

Table 4. Final Confusion Matrix Result**.**

	Negative	**Positive**
Negative	21491	1019
Positive	1986	14320

DOI: 10.54254/2753-8818/17/20240623

4. Discussion

The model reaches a prediction accuracy of 92.28% and validates that hypertension is the most important leading cause of stroke. The strength of the model is that it can provide reliable predictions for people who want to be aware of their future health situation. The model is also easy to use since people can get their results immediately by providing some of their health data, for instance, whether they have hypertension or heart disease, their age, BMI, and so forth, the model will calculate their probability of having a stroke automatically and then report to them in a second. For people who have a higher probability of having a stroke, they can stay alert and try to take some precautionary measures like exercising regularly and eating more balanced diets to lower their BMI [10]. After a period of time of taking precautionary measures, they can come back and input their current indexes again to see whether their probability of having a stroke becomes lower.

Clearly, there are limitations to this model. First and foremost, the accuracy of the model is not perfect and people cannot rely on the model's prediction completely. To further increase the accuracy, several methods can be conducted. Currently, the model contains 10 feature variables and some of them are found loosely correlate with stroke, like work type, marital status, and residence type because their correlation coefficients are lower than 0.2. Therefore, it is reasonable to eliminate those variables and add some variables that are considered more related to stroke, such as diet, circumstance of air pollution, body cholesterol level, and so forth. Future studies are expected to improve the accuracy of the model by trying different algorithms, like K-Nearest Neighbors, and Random Forest [11].

The future expectation is to create an easy website based on the predictive model for potential stroke patients in the future. The main role of the website is for people to check their probability of having a stroke based on the data they report. For those who have higher probabilities, the website can give them advice on reducing the risk of having a stroke and recommend proper hospitals for people to have a physical examination. The website can remind potential stroke patients to do some tasks every day to keep fit and ask them to report their data every week to see when their probability of having a stroke gets lower. With close interactions with people, the website can help people to stay away from stroke. Besides, the information collected and the predictions made can be useful sources for governments. If people are willing to give their geographical areas, it is better to gain a dataset of stroke prediction. By separating people into different age groups, governments can get to know previously about how the trend of stroke will be in their location. They can see which age group needs assistance with medical care to the greatest extent, so governments can allocate more funds to them specifically, thus reducing the issues of lack of medical resources.

5. Conclusion

In the face of the large burden of potential patients with stroke, the mathematical model provides people with an easy and effective way to estimate their probability of having a stroke in the future. The model is effective because it reaches an accuracy of 92.28% using the test data from the dataset. It is also simple for users because they can get to know the outcome by inputting some basic health data or filling out a questionnaire. Based on the model, the incidence of hypertension is verified as the most crucial leading cause of stroke. Future improvements of the model can be done by adding related feature variables and modifying the algorithm of the predictive model. The study's objective is to trace the health data of people to keep them away from stroke and other chronic diseases. Policymakers can acquire our predictive data to have an overview of the disease burden of stroke in the future, thus making pre-arranged planning before the explosion of the stroke population.

References

[1] Wolfe C. D. (2000). The impact of stroke. *British Medical Bulletin*, *56*(2), 275–286. https://doi.org/10.1258/0007142001903120

[2] Howard, G., & Goff, D. C. (2012). Population shifts and the future of stroke: forecasts of the future burden of stroke. *Annals of the New York Academy of Sciences*, *1268*, 14–20. https://doi.org/10.1111/j.1749-6632.2012.06665.x

[3] Mohammed, N., Narayan, V., Patra, D. P., & Nanda, A. (2018). Louis Victor Leborgne ("Tan"). *World neurosurgery*, *114*, 121–125. https://doi.org/10.1016/j.wneu.2018.02.021
[4] World Health Organization. (2020). The Top 10 Causes of Death.
[5] World Stroke Organization. (2019). Global Stroke Fact Sheet.
[6] Salerno, A., Strambo, D., Nannoni, S., Dunet, V., & Michel, P. (2022). Patterns of ischemic posterior circulation strokes: A clinical, anatomical, and radiological review. *International journal of stroke: official journal of the International Stroke Society*, *17*(7), 714–722. https://doi.org/10.1177/17474930211046758
[7] Tolkachev, A., Sirazitdinov, I., Kholiavchenko, M., Mustafaev, T., & Ibragimov, B. (2021). Deep Learning for Diagnosis and Segmentation of Pneumothorax: The Results on the Kaggle Competition and Validation Against Radiologists. *IEEE journal of biomedical and health informatics*, *25*(5), 1660–1672. https://doi.org/10.1109/JBHI.2020.3023476
[8] Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Methods in molecular biology (Clifton, N.J.)*, *404*, 273–301. https://doi.org/10.1007/978-1-59745-530-5_14
[9] Liu, Z., Wang, L., Xing, Q., Liu, X., Hu, Y., Li, W., Yan, Q., Liu, R., & Huang, N. (2022). Identification of GLS as a cuproptosis-related diagnosis gene in acute myocardial infarction. *Frontiers in cardiovascular medicine*, *9*, 1016081. https://doi.org/10.3389/fcvm.2022.1016081
[10] Marzolini, S., Robertson, A. D., Oh, P., Goodman, J. M., Corbett, D., Du, X., & MacIntosh, B. J. (2019). Aerobic Training and Mobilization Early Post-stroke: Cautions and Considerations. *Frontiers in neurology*, *10*, 1187. https://doi.org/10.3389/fneur.2019.01187
[11] Collin, F. D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J. M., & Estoup, A. (2021). Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. *Molecular ecology resources*, *21*(8), 2598–2613. https://doi.org/10.1111/1755-0998.13413