Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/18/20230961

The prediction of stroke and feature importance analysis based on multiple machine learning algorithms

Songhan Li

The Department of Software Engineering (Sino-Foreign Cooperation), Zhejiang University of Technology, Zhejiang, 310023, China

202003340208@zjut.edu.cn

Abstract. Stroke is a leading cause of death and disability worldwide, which requires the accurate and timely diagnosis for effective stroke management. Based on the Kaggle dataset, data preprocessing, which included addressing missing values, encoding categorical variables, and normalising numerical features, was done first in the study. Next, this paper implemented three commonly used machine learning models: logistic regression, decision tree, and random forest. To assess the performance of the models, the paper applied accuracy as the evaluation metric, which measures the proportion of correct predictions out of all predictions. This study also identified the most important features affecting stroke risk using feature importance analysis provided by the machine learning. All three models achieved accuracy rates, according to the experimental findings, albeit random forest outperformed the other two models. The reliability of the models for random forest, decision tree, and logistic regression were 0.963, 0.925, and 0.961, respectively. Feature importance analysis revealed that age, average glucose level, and work type were the most important predictors of stroke risk. Findings in this study suggest that machine learning algorithms, particularly the Logistic Regression model, can effectively predict the likelihood of stroke using the Stroke Prediction Dataset. These findings are in line with other research that also showed how machine learning has the potential to enhance stroke diagnosis. The identification of important features affecting stroke risk can provide valuable insights for clinicians and researchers in developing targeted interventions for stroke prevention and management.

Keywords: stroke prediction, machine learning, artificial intelligence.

1. Introduction

According to a study by the 21st Century, a vascular event that causes an acute focused injury to the central nervous system, such as a cerebral infarction, intracerebral haemorrhage, or subarachnoid haemorrhage, often causes a major global cause of disability and mortality. This condition has severe consequences for public health and leads to neurological impairments [1]. A major contributor to adult disability, stroke affects more than 17 million individuals annually [2]. After a stroke, many people have trouble moving around, taking care of themselves, and communicating. This can result in a loss of independence and a lower quality of life [3]. Additionally, stroke victims frequently have cognitive impairments, such as memory loss, concentration issues, and executive dysfunction, which can make it difficult for them to engage in social and professional activities [4]. Stroke is a serious, disabling

DOI: 10.54254/2755-2721/18/20230961

condition that can have a significant effect on both the person and society at large. Therefore, it is crucial to take precautions to avoid and guard against stroke. However, the existing techniques for diagnosing stroke are constrained by their manual labour's high cost, slow speed, and error-prone nature. The rapidly developing discipline of artificial intelligence (AI) has the potential to improve stroke diagnosis accuracy and efficacy. Machine learning algorithms may be trained on large datasets of stroke patient information to create models that accurately predict the likelihood of stroke, giving medical personnel the knowledge, they need to make treatment and care decisions. By allowing earlier and more precise diagnoses that can result in more timely treatment and a higher chance of recovery, the use of AI in stroke diagnosis has the potential to greatly improve patient outcomes. Through automating many of the tasks associated with stroke diagnosis, such as image analysis and data processing, AI can also lessen the strain on medical personnel [5].

The healthcare industry has undergone a significant transformation owing to the development of AI. Since the early days of rule-based systems, AI has advanced quickly to the current cutting-edge techniques like deep learning. The AlphaGo program, created by DeepMind Technologies in 2016, set a significant milestone in the annals of AI when it defeated a world champion in the challenging board game Go. Medical areas like medical imaging, disease diagnosis, and drug discovery have benefited from AI. For instance, deep learning has been used by some studies to analyze medical images and precisely identify conditions like skin cancer and breast cancer [6, 7]. In addition, AI is also useful in predicting disease outcomes and the efficacy of treatments, as evidenced by a recent study on lung cancer patients' overall survival forecast using deep learning [8]. But because the application of AI in healthcare is still in its infancy, there are a number of challenges and potential ethical problems that need to be addressed. These include possible overreliance on technology in healthcare decision-making, data privacy concerns, and algorithmic bias [9].

The goal of this study is to use machine learning algorithms to forecast a patient's risk of having a stroke based on other physical characteristic information. The Stroke Prediction Dataset from Kaggle was used as the primary data source in this study [10]. Three well-known machine learning models—logistic regression, decision tree, and random forest—were used to forecast the likelihood of stroke. In order to understand which variables affected stroke risk the most, the feature importance of each method was extracted. The effectiveness of these techniques in foretelling the incidence of stroke was demonstrated by experimental findings.

2. Methodology

2.1. Dataset preparation

The primary data source for this study is the Stroke Prediction Dataset from Kaggle [10]. The dataset contains 5110 observations and 12 features that provide information about various aspects of an individual's health and lifestyle, including age, gender, hypertension, heart disease, smoking status, and BMI. In addition, the dataset is balanced, with an equal number of stroke and non-stroke cases.

Any machine learning project must include data preparation because it can significantly affect the model's performance. The preparation of data prior to model training necessitates a cleaning process to ensure its quality. The following procedures were used in this study for data preprocessing: Initially, the missing values and outliers were identified. With the 'id' column excluded, matching rows with incomplete data were removed instead of having the mean or median of the feature be used to fill in the missing values. Next, one-hot encoding was used to transform the categorical variables to numerical variables because it better represents categorical data in machine learning models. For each category in the categorical variable, one-hot encoding entails establishing a binary column where a 1 denotes the category's presence and a 0 denotes its absence. After that, feature scaling was applied to ensure that all variables were on the same scale. The standard scaler from scikit-learn was used to scale the features. In the end, a data split of 70:30 ratio was employed, where 70% of the data was allocated to the training set for model training, and the remaining 30% was designated as the testing set to evaluate the performance of the machine learning models on unseen data.

DOI: 10.54254/2755-2721/18/20230961

2.2. Machine learning models

In this study, three machine learning models were employed to predict the likelihood of stroke, namely logistic regression, decision tree, and random forest.Logistic regression is a statistical method that models binary outcomes by predicting the probability of the outcome based on the values of independent variables [11]. Decision trees are a type of supervised learning technique that is non-parametric and commonly used for classification and regression tasks. In this approach, the tree structure is composed of internal nodes that contain tests on attributes, branches that represent the test's results, and leaf nodes that correspond to class labels, symbolizing the decisions and their possible outcomes in a tree-like fashion [12]. A sort of ensemble learning technique called random forests combines several decision trees to give predictions that are more accurate than those from a single decision tree. Each tree in the forest is built using a random subset of the data's features, and the final forecast is created by averaging all of the trees in the forest's predictions [13]. All models were implemented in this study using the scikit-learn library in Python. For each model, the dataset was randomly split into training and testing sets (70:30). After the training set was used to fit the model, the performance of the model was evaluated. The accuracy statistic, which gauges the percentage of accurate predictions among all those made by the model, was used to assess the performance of the models. Overall, logistic regression, decision tree, and random forest are all commonly used machine learning models for binary classification problems. By comparing their performance and feature importance, insights into which variables are most significant for predicting stroke and which model is the most effective for this particular dataset can be obtained.

3. Results and discussion

Table 1 displays the outcomes of the model comparison. It can be seen that all three models achieved high accuracy rates, with Logestic Regression having the highest accuracy of 96.3%. Random forest and decision tree achieved accuracy rates of 96.1% and 92.5%, respectively.

Table 1. The performance of different models.

Model	Accuracy
Logistic Regression	96.3%
Decision Tree	92.5%
Random Forest	96.1%

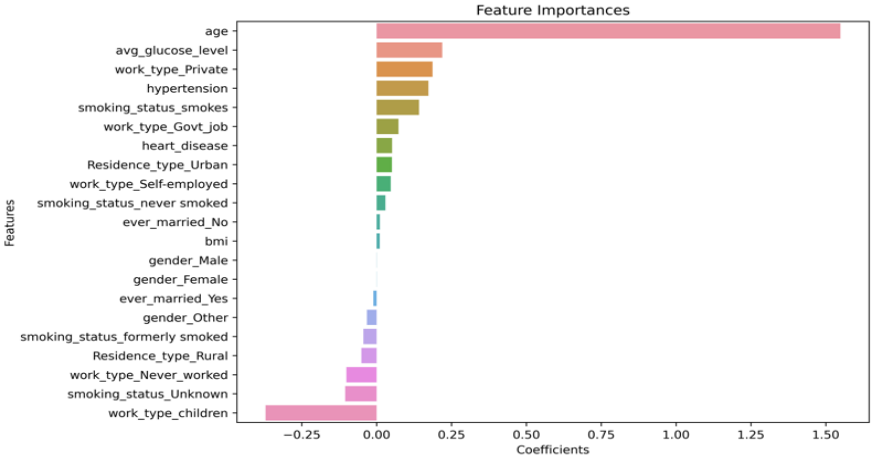

Figure 1. Feature importance of logistic regression.

DOI: 10.54254/2755-2721/18/20230961

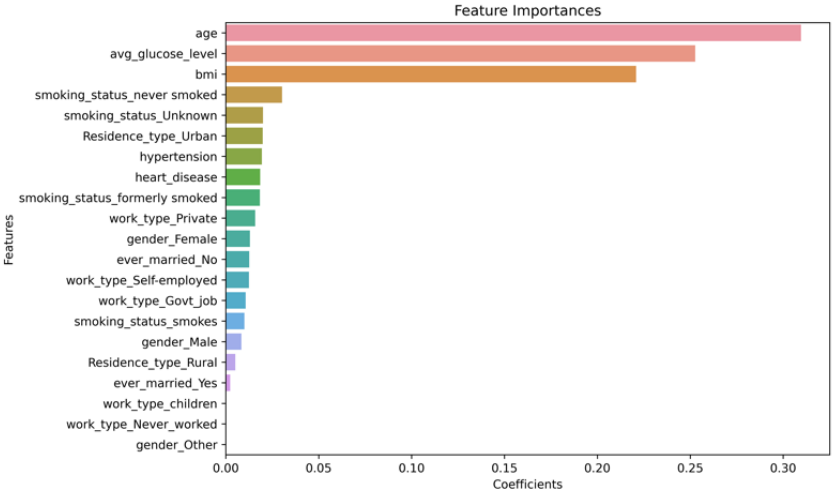

Figure 2. Feature importance of decision tree.

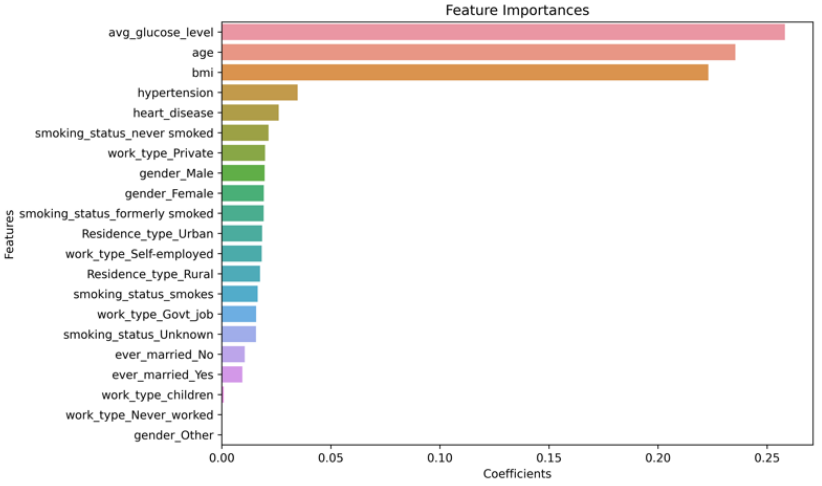

Figure 3. Feature importance of random forest.

Figure 1, Figure 2 and Figure 3 illustrate the feature importance of the variables obtained from three different model. The variables are ranked based on their importance scores, with higher scores indicating greater importance in predicting stroke. It is noteworthy that age and average glucose level are the top two most important variables in predicting stroke, followed by body mass index (BMI) and heart disease etc.

The results suggest that all three models attained considerable accuracy in predicting stroke, with Logestic Regression exhibiting slightly superior performance compared to the other models. Furthermore, the feature importance analysis revealed that age, average glucose level, and work type were the most influential variables in predicting stroke. This is consistent with previous studies that have identified these variables as risk factors for stroke [14].

The high accuracy of the models in predicting stroke suggests that machine learning algorithms have the potential to be efficacious tool for early detection and prediction of stroke risk, which could aidhealthcare professionals in identifying individuals at high risk of stroke and implementing preventive measures in a timely manner.

However, it is essential to acknowledge some limitations of this study. First, the dataset used in this study may not be completely representative of the general population, as it was obtained from Kaggle

DOI: 10.54254/2755-2721/18/20230961

and may be subject to selection bias. Additionally, the models developed in this study may not be generalizable to various populations or healthcare settings, and further validation on diverse datasets is warranted. Moreover, the interpretation of feature importance should be done with caution, as it may be influenced by various factors such as sample size and model complexity.

4. Conclusion

The study's ultimate goal was to assess how well different machine learning algorithms could foretell the chance of a stroke. The Stroke Prediction Dataset from Kaggle was utilized, and three commonly used models, namely logistic regression, decision tree, and random forest were implemented. After extensive data preprocessing, the accuracy was chosen as the evaluation metric for evaluating the model's performance. All three models obtained good accuracy rates, according to the experimental findings, albeit random forest outperformed the other two models. The study also identified the most significant features affecting stroke risk, which can be used as a reference for future research. It is possible to increase the precision and effectiveness of stroke diagnosis and treatment through the application of machine learning techniques, which will ultimately result in better patient outcomes. Future research should, however, also take into account the ethical and societal ramifications of employing machine learning in healthcare, as this study primarily focused on the models' accuracy. Overall, the study shows how machine learning could help in stroke detection and underlines the need for more research in this field.

References

[1] Sacco R L et al. 2013 An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association Stroke vol 44 7 2064-89
[2] Feigin V L et al 2014 Global and regional burden of stroke during 1990–2010: findings from the Global Burden of Disease Study 2010 The Lancet 383(9913) 245-255
[3] Barker-Collo S Bennett D A Krishnamurthi R V et al. 2015 Sex differences in stroke incidence, prevalence, mortality and disability-adjusted life years: results from the Global Burden of Disease Study 2013 Neuroepidemiology 45(3): 203-214
[4] Bhogal S K Teasell R Foley N et al. 2004 Lesion location and poststroke depression: systematic review of the methodological limitations in the literature Stroke 35(3): 794-802
[5] Jiang F Jiang Y Zhi H et al. 2017 Artificial intelligence in healthcare: past, present and future Stroke and vascular neurology 2(4)
[6] Esteva A Kuprel B Novoa R A 2017 et al. Dermatologist-level classification of skin cancer with deep neural networks nature 542(7639): 115-118
[7] Wang X Peng Y Lu L et al. 2017 Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases Proceedings of the IEEE conference on computer vision and pattern recognition 2097-2106
[8] Huang C Yin C 2021 DEEP LEARNING SURVIVAL PREDICTION FOR LUNG CANCER PATIENTS Biomedical Engineering: Applications, Basis and Communications
[9] Topol E J 2019 High-performance medicine: the convergence of human and artificial intelligence Nature medicine 25(1): 44-56
[10] Stroke Prediction Dataset Kaggle 2021 https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
[11] Hosmer Jr D W Lemeshow S Sturdivant R X 2013 Applied logistic regression John Wiley & Sons
[12] Breiman L Friedman J Olshen R et al 1993 Classification and regression trees, wadsworth international group, belmont, ca Case Description Feature Subset Correct Missed FA Misclass 1: 1-3
[13] Breiman L 2001 Random forests Machine learning 45: 5-32
[14] Newer risk factors for stroke 2001 Neurology 57(suppl 2): S31-S34