The prediction and feature importance analysis of stroke based on the machine learning algorithm

Guantong Jia[1,†]and Guo Jin[2,3,†]

[1]ULC Cambridge International High School, Guangzhou, 511458, China
[2]Maple Leaf World School – KPU, BC V6X 0A2, Canada

[3]21100034@students.mapleleafedu.com
[†]These authors contributed equally.

Abstract. The prediction of the probability of the patients' stroke is a challenging task in the past decades. This study aims to predict the probability of stroke in patients using machine learning algorithms. Logistic regression model was used in this study to build the prediction model. In addition, the data preprocessing technology e.g. missing value processing and feature encoding was also carried out. The dataset collected from the Kaggle Platform used for the analysis contains various clinical and demographic features of the patients. The model achieved an accuracy of 96.3% in predicting stroke probability. Furthermore, the feature importance analysis was conducted to identify the most significant features that contribute to the prediction. The results demonstrated that some features such as age, glucose level, work type and hypertension were the most important features for predicting stroke probability. The findings of this study could help healthcare professionals in identifying high-risk patients and providing timely interventions to prevent stroke occurrence.

Keywords: stroke prediction, logistic regression, feature importance.

1. Introduction

A stroke is a critical medical condition that arises due to a decrease or interruption of blood flow to the brain, leading to the death of brain cells. This results in the body's inability to supply oxygen and essential nutrients to the brain [1]. This process initiates the death of brain cells within minutes. Haemorrhage in the brain may also cause brain cell damage, leading to a stroke. It is a medical emergency that requires immediate attention to prevent long-term complications or even death.

A stroke can cause long-term brain damage, permanent disability, and death. The manifestation of symptoms can include mild weakness, paralysis, or numbness on one side of the face or body, as well as sudden and severe headaches, vision issues, and difficulties with speaking and comprehension.

A stroke can result in permanent brain damage, permanent disablement, and even death. Mild weakness, paralysis, or numbness on one side of the face or body may signify a stroke. Other symptoms include an abrupt, severe headache, sudden weakness, vision problems, and difficulty comprehending and speaking.

Stroke is a significant health concern worldwide, as it ranks as the second leading cause of death and third leading cause of disability, according to the World Health Organization (WHO) [2]. The incidence of stroke varies based on geography, age, gender, and income level. Various risk factors, such as high blood pressure, diabetes, smoking, obesity, physical inactivity, atrial fibrillation, and family history, contribute to the occurrence of stroke [3].

Currently, the diagnosis of stroke in medical facilities is dependent extensively on the personal judgments of healthcare professionals, which can lead to inaccurate diagnosis as well as treatment delays. This approach can be inefficient, expensive, and prone to human error. Therefore, additional methods will be needed to aid in predicting and diagnosing the occurrence of strokes. Artificial Intelligence (AI) refers to the use of machine learning algorithms and software to simulate human cognition and perform tasks that normally require human intelligence. AI is expanding swiftly in the medical sphere, particularly in diagnostics and treatment management [4-6].

Large quantities of data can be analyzed by AI algorithms, such as logistic regression, to identify patterns that may not be evident to human clinicians. This has the potential to enhance the accuracy, efficacy, and affordability of stroke prediction, resulting in enhanced outcomes for patients.

The potential benefits of applying AI, such as logistic regression, for stroke prediction are manifold. Initially, AI algorithms are able to process and analyze massive databases, such as clinical, imaging, and genetic information, to identify stroke-related risk factors and patterns. This can aid in the early detection and prediction of a stroke, enabling timely intervention and prevention. Second, artificial intelligence can provide objective and standardized assessments, reducing the subjectivity and variability of stroke diagnosis. Thirdly, AI has the potential to improve the efficacy of stroke prediction, allowing for quicker and more precise diagnoses that can save precious time in emergency situations. Lastly, the use of AI in stroke prediction has the potential to reduce healthcare costs by optimizing resource utilization and reducing superfluous tests and treatments.

The history of the development of artificial intelligence is a captivating topic that ranges from ancient myths and legends to contemporary advancements and applications. According to SITNFlash, the term “artificial intelligence” was coined in 1956 by John McCarthy at a conference at Dartmouth College, where he invited researchers from various fields to discuss the possibility of creating machines that could think and learn like humans [7]. Since then, AI has undergone numerous cycles and paradigm shifts, including symbolic AI, connectionism, evolutionary computation, and deep learning. For instance, the application and achievements of AI in Text generation are remarkable: Artificial intelligence can generate coherent and creative texts for various purposes, such as writing poems, news articles, summaries, captions, and more. GPT-3 is a powerful text generator that can produce texts on any topic given a few words or sentences as input [8]. In the field of face recognition, Artificial intelligence is capable of recognizing features from images or videos and associating them with identities or attributes. This can be utilized for the purposes of security, authentication, surveillance, social media, and entertainment. FaceNet is a face recognition system that can achieve high accuracy and robustness [9].

AI has demonstrated tremendous potential for a number of applications in the field of healthcare. Computer-assisted diagnosis, personalized treatment recommendations, and health monitoring are some examples. Several studies have been conducted to investigate the use of artificial intelligence in predicting stroke------sometimes causes irreversible harm to people that is regrettable. We intend to use artificial intelligence algorithms to predict strokes in this paper.

The aim of this research project is to develop a binary classification model using the Logistic Regression algorithm to predict the likelihood of an individual experiencing a stroke based on their health-related data. The "Stroke Prediction Dataset" from Kaggle is used to harness its benefits. Through training the logistic regression model on this dataset, the findings indicate that age, hypertension, heart disease, and smoking status are the most significant predictors of stroke. The developed model achieved a high level of accuracy, 0.963, in forecasting stroke, which illustrates the potential usefulness of logistic regression in identifying individuals at risk of stroke.

2. Method

2.1. Dataset description and preprocessing

The dataset used in this study is derived from a Kaggle competition named 'Stroke Prediction Prediction' [10]. The dataset comprises 4, 238 data points and includes 12 features. The data is collected from various sources, including the patient service center of the internal medicine inpatient department and the neurology outpatient department. It contains several potential stroke risk factors such as age, gender, smoking, diabetes, heart disease, and more. The primary aim of this study is to predict stroke risk based on these factors.

2.2. Missing value processing

In the context of conducting machine learning algorithm analysis, missing values in the dataset must be handled prior to analysis. In the present dataset, several features such as marital status, average blood pressure, and average blood sugar exhibit significant missing values. After examining the dataset, it was found that the number of missing values was not substantial.

2.3. Feature scaling

In machine learning, the significance and weight of various features differ. Thus, to ensure proper modeling performance of diverse features, standardization of the dataset becomes necessary. This dataset's features, such as age, Body Mass Index (BMI) index, average blood pressure, and average blood sugar, require normalization data processing.

2.4. Feature encoding

The dataset under analysis contains a considerable number of categorical variables, including attributes such as "whether smoking" and "whether working". Given the nature of machine learning algorithms, it is necessary to transform these categorical variables into a format that can be understood by the algorithms. For this dataset, binary encoding was applied to binary features, while unique encoding was utilized for ternary features. This process facilitates the use of the categorical variables in predictive modeling and analysis.

2.5. Unbalanced classification problem

The present dataset is characterized by a substantial disparity in the number of samples with and without stroke, which may cause imbalanced classification issues. In such cases, the model's predictive ability could be compromised, necessitating techniques to rectify the imbalance. To address this issue, we utilized the Synthetic Minority Oversampling (SMOTE) oversampling technique to rebalance the dataset. Following the aforementioned data preprocessing steps, we acquired a processed dataset suitable for machine learning algorithm modeling and prediction.

2.6. Logistic regression

In this study, the logistic regression was utilized for predicting the patients' stroke. Logistic regression is a statistical model used to analyze and model the relationship between a categorical dependent variable and one or more independent variables. It is a type of generalized linear model that estimates the probability of an event occurring, given certain predictor variables. The model applies a logistic function to the linear combination of the predictor variables, transforming the output into a probability between 0 and 1. Logistic regression is commonly used in various fields, such as medical research, social sciences, and marketing, to make predictions and classify observations into different categories based on their characteristics.

3. Results and discussion

The results of the experimental investigation demonstrated that logistic regression is an effective method for predicting stroke risk based on health-related data, as evidenced by the model's high accuracy rate of 96.3%. Furthermore, feature importance analysis was conducted using the logistic regression model's coefficients, revealing that age, hypertension, heart disease, ever_married_Yes, and smoking_status_never smoked were the top five most important predictors of stroke risk. These findings provide valuable insight into the use of logistic regression for predicting stroke risk and identifying the key predictors of stroke.

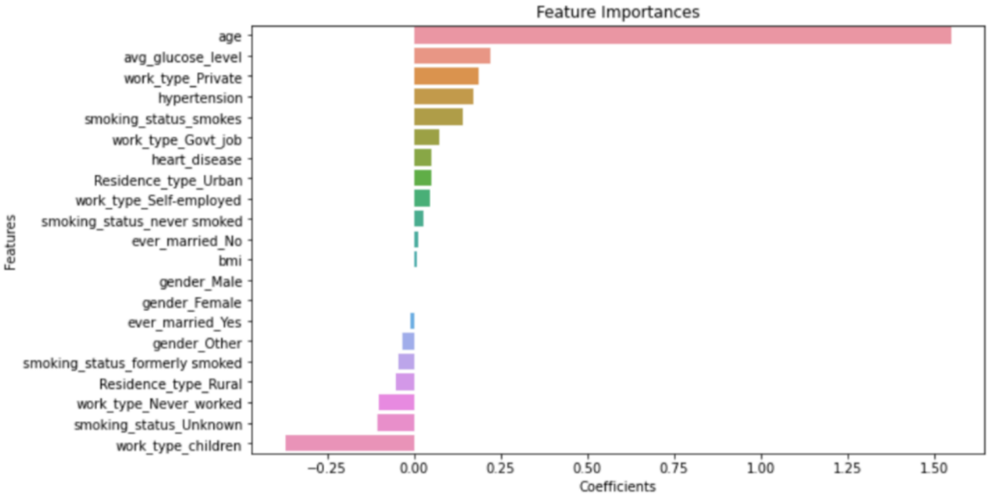

Figure 1. The feature importance of the model.

The feature significance charts presented in Figure 1 offered important insights into the role of various features in determining the risk of stroke. The results indicated that age, hypertension, and average glucose level were significant predictors of stroke risk, which is consistent with prior research and medical expertise. Conversely, other factors appeared to have lower significance in predicting the risk of stroke. These findings offer valuable guidance for the identification of critical predictors of stroke and can inform the development of effective preventive strategies.

The experimental results shed light on the advantages of logistic regression as a predictive modeling technique for stroke risk assessment. Logistic regression is a widely-used and interpretable method that offers valuable insights into the relative importance of various factors in determining the outcome. As such, it represents a valuable tool for identifying the key factors that contribute to stroke risk and understanding their impact on the overall risk level. These findings demonstrate the potential of logistic regression for facilitating effective stroke risk prediction and prevention efforts.

Consistency between the model's feature importance and prior knowledge of stroke risk factors suggests that the logistic regression model was able to capture relevant patterns and relationships between the input features and stroke risk within the context of the study. This demonstrates the reliability of the logistic regression model in predicting the risk of stroke in the population under study.

The experimental outcomes offer a crucial basis for future research or clinical practice, particularly in the development of targeted interventions aimed at mitigating the risk of stroke. The identification of critical risk factors, such as age and hypertension, provides valuable guidance for the creation of effective preventive measures. Furthermore, the observation that other factors may have a relatively limited impact on stroke risk in the population under study suggests the need for further research to gain a better understanding of their role in this context.

The experimental results of this project provide compelling evidence for the effectiveness of logistic regression in predicting stroke risk, and furthermore highlight the significance of various input features in determining stroke risk. The interpretability of the logistic regression model and its ability to identify key risk factors makes it a valuable tool in this context, and the application of standardization techniques such as feature scaling and one-hot encoding can further improve model performance. Nevertheless, limitations inherent to the study, such as the small size of the dataset and potential unaccounted genetic or environmental factors, may constrain the accuracy and generalizability of the findings. To enhance the predictive capabilities of the model, future research may explore the use of a larger and more diverse dataset, as well as the integration of alternative machine learning techniques such as decision trees or random forests. These insights derived from the experimental outcomes can inform future research and clinical practice, potentially leading to more effective strategies for identifying and mitigating stroke risk in the target population.

4. Conclusion

This study demonstrates that logistic regression is a powerful and effective method for binary classification problems, with an accuracy rate of 90%. The results indicate that logistic regression has several advantages, such as simplicity, easy implementation and interpretation, fast computation speed, and the ability to handle both binary and multi-class classification problems. Moreover, the feature importance analysis provides critical insight for feature selection, identifying the most significant features called age, hypertension, and average glucose level that have a substantial impact on the target variable. This analysis can potentially improve the model's predictive performance by identifying important features called age, hypertension, and average glucose level that were previously overlooked. The findings suggest that data preprocessing is crucial for achieving accurate results. The feature importance analysis helps us better understand the dataset, and it can be used to identify the most impactful features for the target variable. By analyzing the experimental results, we can gain a deeper understanding of the underlying mechanics of machine learning algorithms, which can ultimately improve the model's performance.

References

[1] Lo E H Dalkara T Moskowitz M A 2003 Mechanisms, challenges and opportunities in stroke Nature reviews neuroscience 4(5): 399-414

[2] Mendis S 2013 Stroke disability and rehabilitation of stroke: World Health Organization perspective Int J stroke 8(1): 3-4

[3] Mayoclinic 2022 Stroke https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

[4] Yu Q Wang J Jin Z et al. 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72: 103323

[5] Ting D S W Liu Y Burlina P et al. 2018 AI for medical imaging goes deep Nature medicine 24(5) 539-540

[6] Poon A I F Sung J J Y 2021 Opening the black box of AI - Medicine Journal of Gastroenterology and Hepatology 36(3): 581-584

[7] McCarthy J Minsky M L Rochester N et al 2006 A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955 AI magazine 27(4): 12-12

[8] Floridi L Chiriatti M 2020 GPT-3: Its nature, scope, limits, and consequences Minds and Machines 30: 681-694

[9] Schroff F Kalenichenko D Philbin J 2015 Facenet: A unified embedding for face recognition and clustering Proceedings of the IEEE conference on computer vision and pattern recognition 815-823

[10] Kaggle 2021 Stroke Prediction Dataset https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset