

Explore the application of machine learning algorithms in predicting stroke

Enming Zhu *

The Second High School Attached to Beijing Normal University, International Division, Beijing, China

* Corresponding author: enming_zhu@outlook.com

Abstract. Stroke is an acute cerebrovascular disease that seriously endangers human health and life safety. It occurs suddenly and is not easy to cure. It is easy to leave sequelae after the disease, and the age of onset tends to be younger, becoming a major disease that threatens human life and quality of life. With the advent of the era of big data, people can process and analyze the data obtained from a large number of patients with cardiovascular and cerebrovascular diseases. Through Machine Learning (ML), they can master the risk factor indicators leading to stroke, so that they can effectively predict the incidence of stroke, help cardiovascular and cerebrovascular patients carry out preventive treatment as soon as possible, reduce the incidence or the symptoms. This paper introduces several classic ML algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), Gaussian Naive Bayesian, k-NN, Logical Regression, etc. In order to explore the accuracy of these algorithms, this study obtained 15,000 patient data from the Kaggle website, including 22 characteristics, and used the above algorithms to predict. Through the evaluation of Accuracy Score indicators, we have a clear understanding of the application of various algorithms in predicting stroke.

Keywords: Machine Learning, Stroke, Predicting, Evaluation.

1. Introduction

With the rapid development of contemporary science and technology, people's lives are undergoing unprecedented technological changes. Among them, the rapid rise of computer science and information technology has laid the foundation for the development of many other industries and scientific research fields. In recent years, the medical field has achieved great success in disease prediction, auxiliary diagnosis, and health management of chronic diseases through machine learning under the guidance of big data and artificial intelligence.

Stroke is an acute cerebrovascular disease. It is a kind of cerebral blood circulation disorder characterized by sudden fainting, unconsciousness, accompanied by squinting corners of the mouth and unfavorable speech. Due to the characteristics of high incidence of stroke, high mortality, high disability rate, high recurrence rate and many complications, the age of onset also tends to be younger. Therefore, it is a major disease that threatens human life and quality of life. Therefore, the medical community has listed coronary heart disease and cancer as one of the three major diseases that threaten human health [1]. The importance of preventing stroke has attracted great attention from the medical community at home and abroad.

Although stroke is very sudden, from the patient's living habits, dietary habits, cardiovascular disease indicators and other health data, machine learning methods are used to explore the relationship between various medical indicators, grasp the impact of these risk factors on the incidence of stroke, provide diagnosis and prevention plans for doctors and patients with cardiovascular and cerebrovascular diseases, reduce the incidence of stroke, reduce the impact of the disease, and make it possible to predict stroke. At present, prediction models in many countries use gender, age, smoking, blood pressure, TC, diabetes, etc. as risk factors in machine learning [2], which can improve the accuracy of predicting stroke.

This paper obtains 15,000 pieces of data from the website, including 22 characteristic fields, and uses six algorithms: DTs, Random Forest, SVMs, Gaussian Naive Bayes, K-Neighbor Classifier and Logical Regression to predict stroke. And then uses the Accuracy method to evaluated the results.

Finally, the accuracy of each algorithm was obtained to explore the application of machine learning algorithms in predicting and evaluating stroke diseases.

2. Introduction to Machine Learning

2.1. Definition of ML

Machine Learning is a multi-domain and interdisciplinary discipline. It refers to using computers to simulate people's way of thinking to have the same learning ability as people, so that they can learn based on a large amount of training data without any clear programming, use the system itself to improve themselves, gradually improve performance, and then make complex decisions according to learning behavior. Depending on whether artificial markers are required for the processed data, machine learning can be broadly classified into three categories: supervised learning, semi-supervised learning, and unsupervised learning.

2.2. Classic algorithms for machine learning

2.2.1. Decision Trees (DTs)

DTs are a non-parametric supervised learning technique. To achieve the final classification, the algorithm utilizes a tree structure and layer-level inference [3]. The decision tree consists of the three elements: root node which is an entire collection of samples, internal node which is to match feature attribute examination, and leaf node which is a representation of the decision's outcome.

When making predictions, utilize an attribute value at the tree's internal node to make a judgment. Based on the judgment results, determine which branch node to enter until it reaches the leaf node, at which point the classification results are produced. Among the benefits of DTs are:

- 1) Easy to comprehend and interpret, visual analysis, easy to extract rules.
- 2) Categorical and numerical data can be processed simultaneously.
- 3) When handling samples that have missing properties, it works better.
- 4) Able to deal with irrelevant features.
- 5) The operation speed is rather quick when testing the dataset.
- 6) For big data sources, it can generate workable and efficient results in a reasonable amount of time.
- 7) The model performs well despite some assumptions being slightly impacted by the actual data model.

Among DTs' drawbacks are:

- 1) Learners of decision trees may produce overly intricate trees called overfitting, so that the data cannot be well generalized.
- 2) Decision trees are unstable because little changes in the data could result in a completely different tree.
- 3) Decision trees are not good at extrapolation.
- 4) Decision trees make it hard to represent several concepts, such as XOR, parity, and multiplexer issues, which makes them hard to learn.
- 5) Decision tree learners create biased trees if certain classes predominate. It is therefore recommended to balance the dataset prior to decision tree fitting.

2.2.2. Random Forest

Random Forest is a comprehensive learning method composed of multiple decision trees for classification, regression and other tasks [4]. Different decision trees do not share any correlation. Each decision tree in the forest will be judged and classified separately when fresh input samples are added while the task of categorization is ongoing. Each decision tree's classification outcome will be distinct. The Random Forest will choose the result with the most classification from the decision tree as its final result. The problem of decision trees overfitting their training sets can be corrected by forests.

Among the benefits of Random Forest are:

- 1) It is not necessary to lower dimensions or choose features when working with high-dimensional (many feature) data to generate a random forest.
- 2) It has the ability to assess feature importance.
- 3) The interaction between different characteristics can be judged.
- 4) The speed of training is relatively fast and parallel methods can be easily implemented.
- 5) It is relatively easy to implement.
- 6) It can balance faults in uneven datasets.
- 7) The accuracy can still be upheld even if a significant portion of the feature is gone.

Among Random Forest' drawbacks are:

- 1) In certain noisy classification or regression problems, random forests have been shown to overfit.
- 2) Utilize an attribute value at the internal node of the tree to judge when making predictions. Based on the judgment findings, determine which branch node to enter until it reaches the leaf node, at which point the classification outcomes are acquired.

2.2.3. Support vector machines (SVMs)

SVMs is a supervised learning algorithm [5]. It can solve classification and regression problems. The idea was first put forward by Corinna Cortes and Vapnik in 1995. Solving small sample, nonlinear, and high-dimensional pattern recognition can be achieved with its unique advantages, and it can be applied to other machine learning problems like function fitting.

Among the benefits of SVMs are:

- 1) It can solve high-dimensional problems, that is, large feature spaces.
- 2) It can resolve issues with machine learning with limited sample sizes.
- 3) Has the ability to handle nonlinear feature interaction.
- 4) The local minimal problem does not exist.
- 5) Does not have to depend on all of the facts (compared to algorithms such as neural networks).
- 6) The generalization ability is relatively strong.

Among SVMs' drawbacks are:

- 1) Efficiency decreases with the increase in the number of observed samples.
- 2) Nonlinear problems have no universal solution, therefore it can occasionally be challenging to choose an appropriate kernel function.
- 3) The interpretation of kernel functions via high-dimensional mapping is weak.
- 4) Conventional SVMs only support two classifications.
- 5) Sensitive to missing data.

2.2.4. K-Neighbor Classifier (KNN)

One of the most straightforward techniques in data mining classification technology is the K-Nearest Neighbor classification algorithm [6]. The so-called K nearest neighbor means that each sample can be represented by its nearest neighbor K. The KNN methodology lacks an explicit learning process and a training step. The data set is pre-classified and eigenvalued, and it is designed to handle fresh samples immediately upon receipt. So, it has been announced as "lazy learning".

2.2.5. Gaussian Naive Bayesian (GNB)

GNB used probability methods and Gaussian distribution to solve classification problems [7].

The advantages are:

- 1) Stable classification efficiency.
- 2) Small-scale data is well-served by it and it can handle multiple classification tasks, making it suitable for incremental training.
- 3) The algorithm is simple and unaffected by missing data, making it commonly used for text classification.

The disadvantages are:

- 1) The performance is the best when the attribute correlation is small, whereas it is not good when the correlation is large.
- 2) It is necessary to know the prior probability, and the prior probability often depends on the hypothesis, so in some cases, the prediction results will be poor due to the hypothetical prior model.
- 3) There is a certain error rate in classification decision-making.

2.2.6. Logical Regression

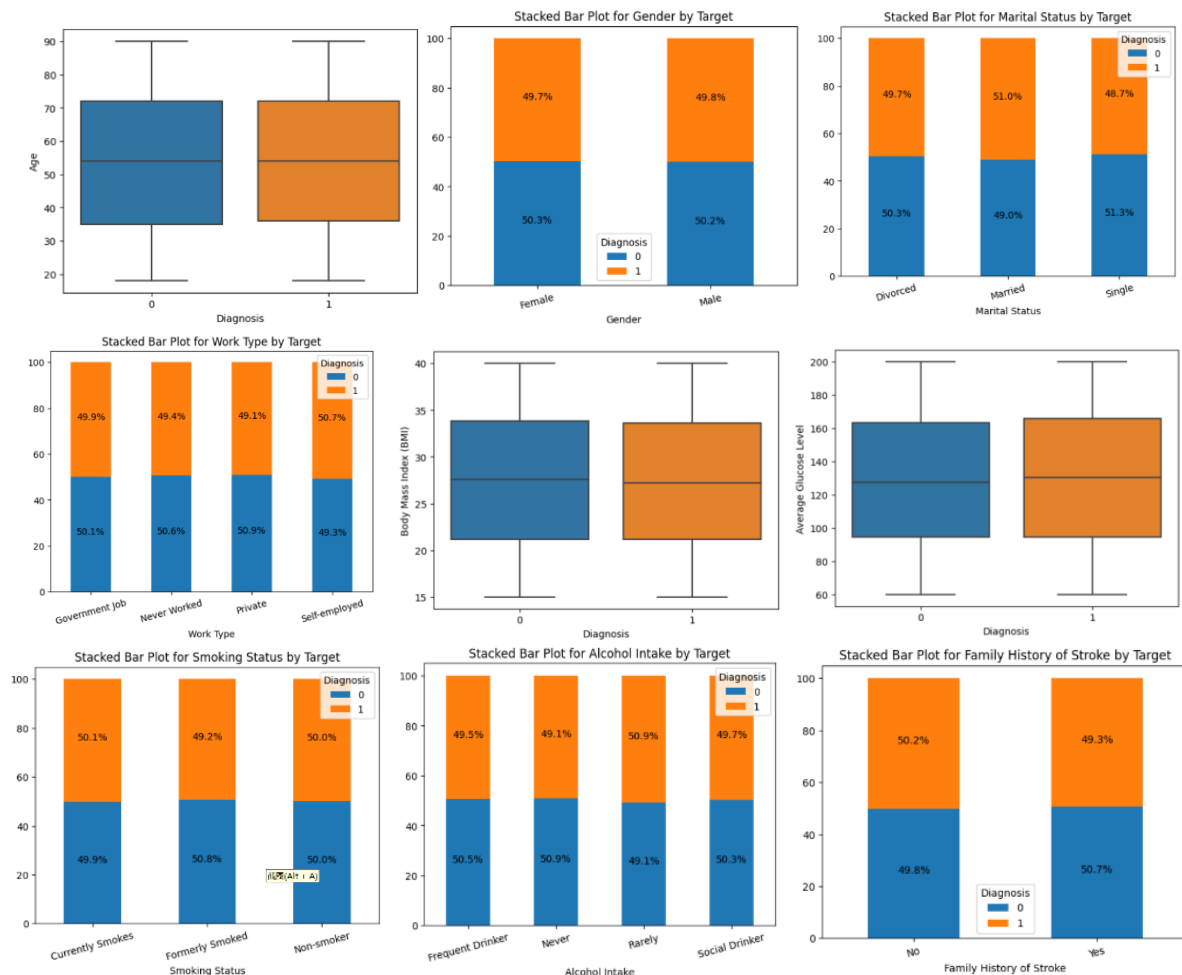
Logical Regression is a type of linear algorithm which used to solve binary classification problems to indicate the possibility of something happening. Logical Regression applies a logistic function instead of hyper-plane or a linear line to place the linear equation output between 0 and 1. The algorithm is simple, the sample probability score is easy to observe, the computation amount is extremely tiny, the speed is rapid, and the storage resources are minimal.

The disadvantages are:

- 1) Only two classification problems can be dealt with, and they must be linearly separated.
- 2) The performance of logical regression suffers when the feature space is large.
- 3) Easy to overfitting.

3. Data source research and analysis

In order to explore data on health indications, living habits of stroked people, this study obtained datasets of 15,000 record & 22 features on the Kaggle website [8], which contain 'Patient ID', 'Patient Name', 'Age', 'Gender', 'Hypertension', 'Heart Disease', 'Marital Status', 'Work Type', 'Residence Type', 'Average Glucose Level', 'Body Mass Index (BMI)', 'Smoking Status', 'Alcohol Intake', 'Physical Activity', 'Stroke History', 'Family History of Stroke', 'Dietary Habits', 'Stress Levels', 'Blood Pressure Levels', 'Cholesterol Levels', 'Symptoms' and 'Diagnosis'. Fig.1 shows the statistics of datasets.



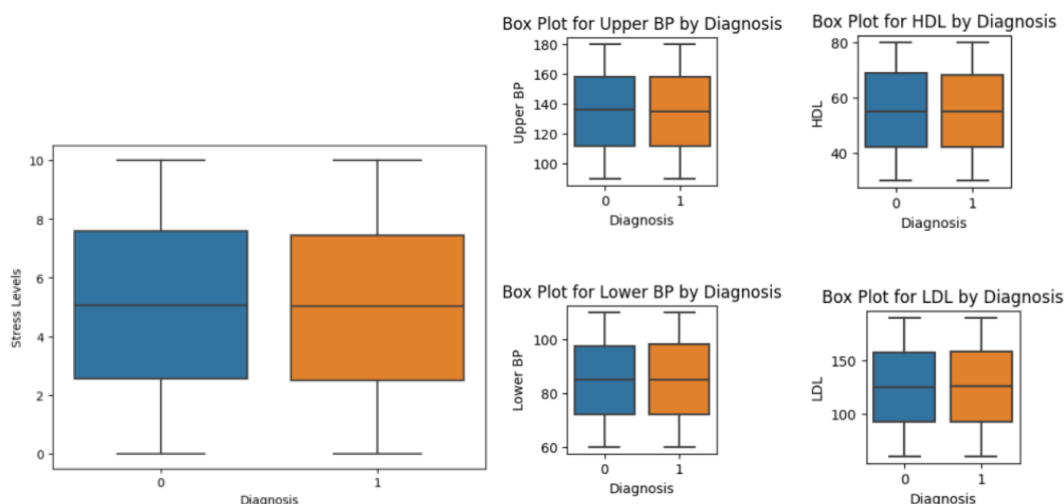


Figure 1. Bar plot an box plot for features ‘Age’, ‘Gender’, ‘Marital Status’, ‘Work Type’, ‘BMI’, ‘Average Glucose Level’, ‘Smoking Status’, ‘Alcohol Intake’, ‘Family History of Stroke’, ‘Stress Levels’, ‘Blood Pressure Levels’, ‘Cholesterol Levels’

4. Prediction and evaluation of several machine learning algorithms

This study uses Python to run algorithm programs. All 22 features are taken into account to achieve the goal, as these data are input into the six ML algorithms mentioned above. In this experiment, Accuracy Score was selected as the evaluation parameter of this study in Recall Score, Accuracy Score, Precision Score (PPV), MAE, MSE or RMSE [9], NPV and other evaluation parameters.

Among the parameters of these evaluations, Recall Score, Precision Score and Accuracy Score are more applicable to the evaluation of the results of classic questions. MSE and MAE are more suitable for evaluating the results of regression estimation problems.

Table 1. Confusion matrix for stroke disease prediction

True Class			
Positive (With disease)	Negative (Without disease)		
TP	FP	Positive (With disease)	Predicted Class
FN	TN	Negative (Without disease)	

The confusion matrix is as Table 1 shows.

1) $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$: Divide all positive samples with accurate predictions by the sum of all positive samples with accurate predictions and false positive samples that predict errors.

2) $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$: Use all positive samples with accurate predictions and the sum of all negative samples with accurate predictions divided by the sum of all samples. This method is relatively comprehensive, but it cannot be directly used for target detection problems, otherwise there may be misjudgment, but the accuracy value is very high.

3) $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$: Use all positives with correct predictions divided by all positive samples. Precision indicates that the useful part of the whole test result accounts for the useful proportion of the whole result.

4) Negative predict value (NPV) = $\text{TN} / (\text{TN} + \text{FN})$: Negative prediction value is also known as retrospective accuracy. Negative prediction value refers to the proportion that the test results can show that all negative samples are truly negative.

After evaluating different algorithms, the accuracy results of different prediction algorithms can be calculated. As can be seen from Table 2, the accuracy of Gaussian Naïve Bayes has reached 0.96,

which is the best overall effect. Decision Tree Classifier and Random Forest Classifier are slightly less accurate, and there is room for further improvement.

Table 2. Comparison of different algorithms use Accuracy Score

	Random Forest Classifier	SVM	Decision Tree Classifier	K-Neighbors Classifier	Gaussian Naïve Bayes	Logistic Regression
Accuracy	0.936	0.941	0.922	0.940	0.960	0.940

5. Conclusion

This paper mainly explores the application characteristics of various ML algorithms in stroke prediction, aiming to find the most suitable model for predicting stroke diseases. Random Forest, SVM, DTs, K-NN, Gaussian Naïve Bayes, Logistic Regression have achieved a certain accuracy in predicting stroke, among which Gaussian Naïve Bayes has a more accurate prediction ability for these test data. In the future, through continuous in-depth research on machine learning algorithms, they can more accurately and effectively predict stroke diseases and relieve pain for human beings.

References

- [1] Apoplexia. Medical encyclopedia. 2023.<https://www.yixue.com>.
- [2] Zhao Zhongsheng. Risk assessment of cardiovascular health. International Journal of Nursing, 2006, 11 (11): 949.
- [3] Decision Trees. scikit-learn <https://scikit-learn.org/stable/modules/tree.html>.
- [4] Random forest. Easyai.2023.<https://easyai.tech/ai-definition/random-forest/>.
- [5] Support Vector Machine.Easyai2023.<https://easyai.tech/ai-definition/svm/#wiki>.
- [6] k-nearest. Easyai.2023. <https://easyai.tech/ai-definition/k-nearest-neighbors/>.
- [7] Sklearn Machine learning: The Gaussian Naive Bayes GaussianNB_sklearn gaussiannb-CSDN blog. 2023.7.
- [8] Stroke Prediction.2023. <https://www.kaggle.com/datasets/teamincrito/stroke-prediction>.
. Regression evaluation index.2023.https://blog.csdn.net/weixin_40651515/article/details/105930868.
- [9] Pothala Ramya, Ashapu Bhavani and Sangeeta Viswanathan. Review: heart diseases detection by machine learning classification algorithms. Journal of Harbin Institute of Technology (New Series), 2022, 29 (4): 81 – 92. DOI: 10. 11916 /j. issn. 1005 - 9113. 2021144.