Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/16/20230877

Stroke risk early prediction leveraging machine learning algorithms

Chi Zhang
School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 610072, China

2020090908026@std.uestc.edu.cn

Abstract. When the blood supply is suddenly blocked in a part of the brain, stroke could be triggered. With blood supply faulted, brain cells would die bit by bit, and handicap happens according to the affected region of the brain. It could cause paralysis or even death to stroke patients. Early symptom recognition can be very helpful in predicting stroke and fostering a healthy lifestyle. Considering conducting large-scale manually early diagnosis is almost impossible, machine learning is always regarded as the first choice for automatic risk estimation. In this research, multiple models of machine learning are developed and trained to predict long-term of the risk of suffering from stroke. This study's main contribution is a randomized forest which presents a high yield which is verified by different kinds of parameters, for instance, F-measure, recall, AUC, accuracy and precision. The outcome of the paper presents that other models are surpassed by random forest model, with 97.6% AUC.

Keywords: random forest, machine learning, stroke prediction.

1. Introduction

As the major cause of the global death, stroke could lead to serious impact on the normal and healthy life of human beings [1,2]. Not only stroke patients are negatively influenced, but also their family and work environments. According to the World Organisation for stroke, ever year, around 13 million patients will get stroke, and about 5.5 million patients dead [3.4]. Moreover, different from common cognition, anyone may have stroke, at any period of life, no matter what physical condition or gender. Stroke, an acute nervous system disease could be triggered when the block happens in the blood vessels of human brain. Given the absence of sufficient blood supply, affected brain cells cannot receive enough neuration and will die in a short period. It could be broken down into haemorrhagic and ischemic [5]. It can be slight or extremely hard, accompanied by temporary and permanent destroy. Haemorrhages are seldom observed. When it happens the blood vessels would break in s specific brain region. The most common, ischaemic strokes, include blocking the blood transfer to a certain area of the brain and cause blocking artery or narrowing [6]. Whether there has been a stroke similarly, previously, the existence of heart diseases, for instance, heart failure, atrial fibrillation, and age, other diseases caused by bad lifestyles (atherosclerosis, estrogenic treatment, high blood cholesterol diabetes, obesity, sedentary lifestyle, smoking, drinking, coagulation disorders, and the use of cocaine and amphetamines and other euphoric substances) are included in the factors that increase the risk of stroke [7]. In this paper, a methodology to develop efficacious machine learning (ML) binary classification models for the

DOI: 10.54254/2755-2721/16/20230877

germination of a stroke is unveiled. Since class balance is essential for designing effective stroke prediction methods, a synthetic few oversampling (SMOTE) method was used. Next, different models are grown and evaluated data set [8]. For the aim of this paper, K-NN, multilayer perception, logistic regression, decision trees, naive bayes, random forests, and stochastic gradient descent (SGD) were assessed [9,10]. Moreover, cumulation methods and majority voting were used, the latter one being this paper's main contribution. Tests demonstrated effectiveness of what is known as stacking method relative to single model and accuracy, making AUC highly, accuracy, recall and also F-measure. Rest of the document arranges in the following manner. Section 2 provides a description of the data set as well as a description of these methods used. What's more, for Section 3, this work addresses the configuration and the research outcomes obtained. At last, Sections 4 and 5 describe the coming guidance.

2. Method

2.1. Dataset

This research was supported by a Kaggle data set. From this data set, this paper focused on participants over the age of 18. There are 3254 attendees, and all attributes are named as follows (1 for the target class and 10 for ML models): (1) Age (years): Refers to participants' age older than 18. (2) Gender: member's gender. Number for men is 1260, while 1994 for women. (3) Hypertension: whether the participant is non-hypertensive or hypertensive. Participants suffering from high blood pressure is about 13.45% of all. (4) Heart Diseases: This characteristic refers to whether the participant has heart related diseases. Participants with heart disease account for 6.33%. (5) Married: This reflects whether participants are married or not, of which 79.84% are married. (6) Kind of work: The character shows what the participant's work status is. It is divided into four kinds. 63.02% of all is "private", 21.21% of all is "self-employed", 14.67% of all is "government employment" and 1.1% of all is "never worked". (7) Kind of Residence: The characteristic shows how the participant lives and consists of two kinds. 50.14% of them live in urban, and 49.86% of them live rurally. (8) BMI (kg/m2): This characteristic captures a participant's body mass index. (9) Status of smoking: The characteristic whether the participant smoke or not and includes three kinds. 22.37% of them smoke, 52.64% of them never smoke and 24.99% of them smoke previously. (10) Stroke: This is an indicator of whether the participant has ever suffered a stroke or never gets stroke. 6.63% of them has had stroke before.

2.2. Data preprocessing

To estimate the risk of suffering stroke long-term, original dataset usually was divided into two parts, a testing package and training. In dataset, c, a variable with two options, represents the class beacon in the I instance. Two possible states should be in the class variable, for instance, c could be "Stroke" or "Non-Stroke". Considering there are corelations between stroke and the possibility of the happening of these risks, ML models could be leveraged to estimate the characteristics of novel instance classes. An instance I is specified for character vector for as if = (fi1, fi2, fin).

The next analysis is intended to develop machine learning models which allow for otherwise sensitivity (or high recall) and a place below the ensuring, curve correct calculate of stroke cases prediction. The proposed method for predicting stroke includes the following steps.

Basic data quality can decide the final quality of forecasts, resulted from noisy data and/or missing values. As a result, necessarily data pre-processing is important, including reducing redundancy, selecting features, and discretizing data to let it become more proper for analysis and exploration. Furthermore, using a resampling method, data pre-processing is class balancing in some parts. In this paper, what is known as SMOTE is used to correct the imbalance in the participants' distribution between these two classes. Particularly, minority of the class, "stroke", was oversampled, so participants were evenly distributed. Furthermore, there were no missing or zero values, so no deletion or imputation of the data was applied.

DOI: 10.54254/2755-2721/16/20230877

2.3. Machine learning models (ML)

For section, the machine learning models to be used in the classification framework for the germination of a stroke is outlined. To this end, different kinds of graders are used.

2.3.1. Naive Bayes (NB). Initially, it was taken into account, which guarantees the most of probabilities if characteristics are very independent. A new subject i with vector characteristics fi assign to this class c for which $P(f_{i1}, \dots, f_{in})$ is generally maximized. Probability is described as follows conditionally:

$$P(c|f_{i1}, \dots, f_{in}) = \frac{P(f_{i1}, \dots, f_{in}|c)P(c)}{P(f_{i1}, \dots, f_{in})} \tag{1}$$

$P(f_{iI}, \dots, f_{in})$ is earlier possibility of characteristics. In addition, earlier probability of class is P(c). To optimize the equation (1), its numerator should be maximized, which further lead to the following optimisation:

$$\hat{c} = argmax\ \ P(c) \prod_{j=1}^{n} P(f_{ij}|c) \tag{2}$$

where$c \in \{Stroke, Non - sroke\}$

2.3.2. Regression (LR). This's a method of statistical classification, which firstly designed to binary tasks. It has also been expanded to meticulous tasks. A binary variable is outputted by this model where p=P(Y=1) represents the possibility that a sample belongs to the class "Stroke", so 1-p=P(Y=0) gets the possibility that the example belongs to this class called "Non-Stroke". There is a relationship linearly between logarithmic fraction and basic parameter b and model β_i, which is as follows:

$$log_b\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 f_{i1} + \cdots + \beta_n f_{in} \tag{3}$$

*2.3.3. K-nearest neighbours (KNN). T*he K-NNs is a kind of method distance-based (i.e., Euclidean, Manhattan) which calculates the difference or similarity between two aspect in the data set under review with the most common and easiest The Euclidian distance. Whether f_{new} is the vector of characteristics to be categorized as whether non-stroke or stroke. KNN decides which K (neighbouring) vectors are closest to f_{new}. Next, the category of f_{new} could be determined by the majority classes of its closest samples.

2.3.4. Random forest (RF). The RF sets contain a lot of independent decision trees. Besides, After resampling, generate various instance subsets for performing regression and classification tasks. After the last class is obtained by a majority vote, own classification results were derived by each decision.

2.3.5. Evaluation matrices. As part of the ML models' the evaluation process under consideration, a number of performance measures have been recorded. In this discussion, the most widely adopted in the recent paper is considered. Recall rate, or response rate, is the percentage of "Stroke" and was suggested positive for all "Stroke". Alerting and precision are better suited for presenting model errors when it comes to unbalanced data. The accuracy presents the number of those who have had suffered a stroke, fall into the category. Reminder shows the number of "Stroke" are adequately anticipated. The F-measurement is the average of accuracy and call back, summarizes a model's predictive performance.

DOI: 10.54254/2755-2721/16/20230877

$$Recall = \frac{TP}{TP + FN}, \qquad Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F - Measure = 2 * \frac{Precision \cdot Recall}{Precission + Recall}, \qquad Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5}$$

Note that FP: negative false and FN: positive false, TP: negative true, TN: positive true.

The AUC is a helpful measurement with rate in the range in [0.1]. As getting close, the performance of the Machine Learning model differs from uninterrupted instances. Perfect discrimination between aspects of these two classes signifies the curve surface (AUC) is equal to 1.

3. Result

Figures 1 and 2 illustrate performances of the aforementioned models, especially for the Stroke class, recall, AUC, F-measure in the field of accuracy. Additionally, Table 1 summarizes selected models' return in average.

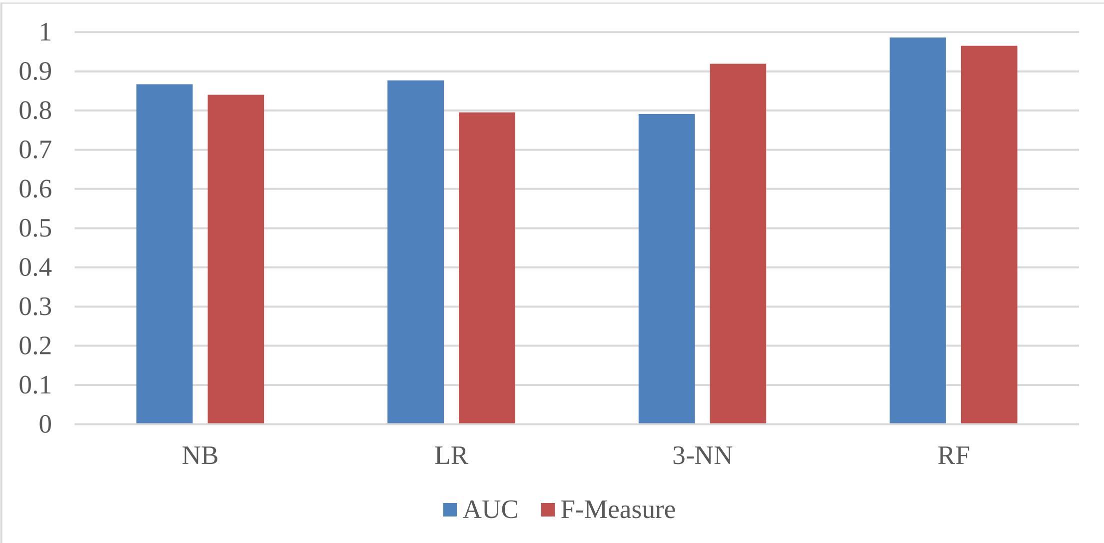

Figure 1. F-measure performances evaluation and corresponding AUC value.

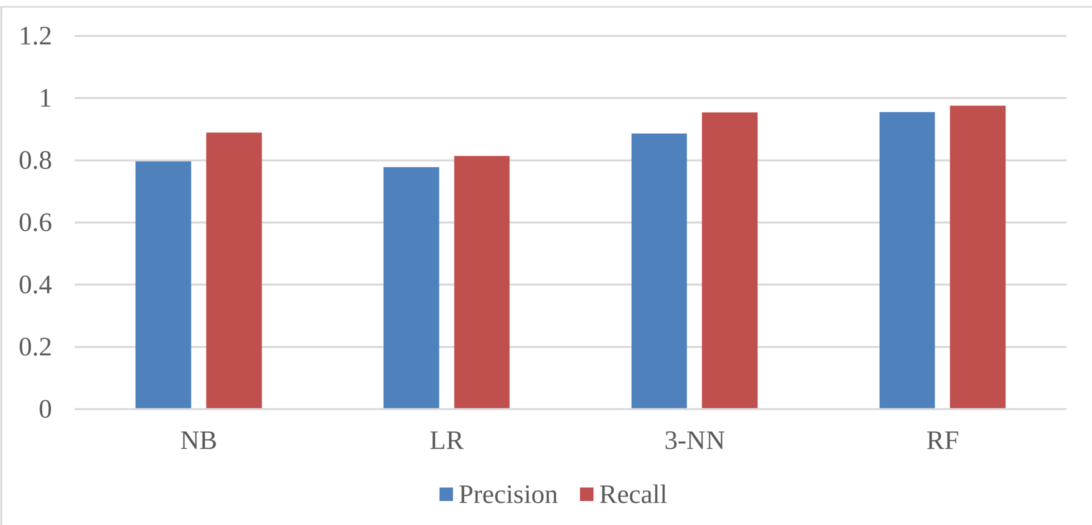

Figure 2. Precision and recall performances.

DOI: 10.54254/2755-2721/16/20230877

Table 1. Performance comparisons.

	Precision	Recall	F-Measure	AUC	Accuracy
NB	0.812	0.860	0.835	0.867	0.84
LR	0.791	0.791	0.791	0.877	0.79
K-NN	0.918	0.916	0.915	0.953	0.81
RF	0.966	0.966	0.966	0.986	0.97

The most effective of all the parameters under study was the stacking model in the selected basic models. As well, RF classifiers reached high values. By concentrating on the measurement of AUC, RF models have shown that, with a high probability of 98.6%, it is possible to successfully identify the course from uninterrupted instances. Apart from RF, the 3-NN model follows, with a substantially higher AUC to 95.3%.

In addition, a comparison of Figure 1 and Table 1 shows that the average behaviour followed by the AUC values for the race class. Note that this stacking grader made a recall which is better than the rest. And in both cases, because of the balanced dataset, the F-measurement shows an appropriate data that may present the outcome of the Machine Learning models. At standpoint of F-measure, respectively, the cumulation goes 5.9% higher than 3-NN and 0.8% higher than the RF.

4. Conclusion

A stroke is a life-threatening event. Unexpected complications can be avoided by preventing and treating. With AI/ML developing rapidly nowadays, medical experts, policy makers, and clinical providers can identify the characteristics (or risk factors) most associated with stroke happening and can assess their probabilities and risks by utilizing established models. In this area, ML/AI can help predicting early of strokes and reduce serious results. The report looks at the ability of different kinds of Machine Learning models in predicting stroke according to participants' data. Assessment of how grader performs with the help of F-measure, and accuracy is, in essence, appropriate for explanation of the models, presenting their classification performance. The objective in future of this paper aims to improve the Machine Learning shell frame by employing deep learning. At last, a full of challenge but up-and-coming field shall be assessing the predictive capacity of in-depth learning models in stroke by gathering CT scans' image data.

References

[1] Yew, K. S., & Cheng, E. M. (2015). Diagnosis of acute stroke. American family physician, 91(8), 528-536.

[2] Mirzaei, H., Momeni, F., Saadatpour, L., Sahebkar, A., Goodarzi, M., et al. (2018). MicroRNA: relevance to stroke diagnosis, prognosis, and therapy. Journal of cellular physiology, 233(2), 856-865.

[3] Benjamin, E. J., Muntner, P., Alonso, A., Bittencourt, M. S., Callaway, C. W., et al. (2019). Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation, 139(10), e56-e528.

[4] Teasell, R., Salbach, N. M., Foley, N., Mountain, A., Cameron, J. I., et al. (2020). Canadian stroke best practice recommendations: rehabilitation, recovery, and community participation following stroke. Part one: rehabilitation and recovery following stroke; update 2019. International Journal of Stroke, 15(7), 763-788.

[5] Popkirov, S., Stone, J., & Buchan, A. M. (2020). Functional neurological disorder: a common and treatable stroke mimic. Stroke, 51(5), 1629-1635.

[6] Hata, J., & Kiyohara, Y. (2013). Epidemiology of stroke and coronary artery disease in Asia. Circulation Journal, 77(8), 1923-1932.

[7] Carvalho, M., Carmo, H., Costa, V. M., Capela, J. P., Pontes, H., Remião, F., et al. (2012). Toxicity of amphetamines: an update. Archives of toxicology, 86, 1167-1231.

DOI: 10.54254/2755-2721/16/20230877

[8] Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., & Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. PloS one, 12(7), e0179805.
[9] Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263-1265.
[10] Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., et al. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, 15(6), e0234722.