XGBoost Algorithm to Predict a Patient's Risk of Stroke

Sada Anne and Amadou Dahirou Gueye(✉)

Alioune Diop University, Bambey, Senegal
{sada.anne,dahirou.gueye}@uadb.edu.sn

Abstract. The negative impact of stroke on society has led to a concerted effort to improve stroke management and diagnosis. As the synergy between technology and medical diagnostics grows, caregivers are creating opportunities for better patient care by systematically exploring and archiving patient records. The ubiquitous growth of artificial intelligence and its medical applications has improved the efficiency of healthcare systems for patients requiring long-term personal care. Today, chronic diseases such as stroke are the world's leading cause of death. Stroke can be caused by a number of factors. By measuring recorded values of patient characteristics such as heart rate, cholesterol levels, blood pressure, diabetes etc., this information can help doctors to make decisions about patient care, in order to predict a possible onset of the disease. Because most stroke diagnosis and prediction systems are image analysis tools such as CT or MRI, which are expensive and not available 24/7 in some African hospitals in general and Senegal in particular. We therefore use a dataset to predict stroke and compare its results with those of other models using the same data. We find that Xgboost, depending on the characteristics of the data, is the algorithm that can effectively predict stroke, and the results obtained are superior to those of other models.

Keywords: XGBoost (eXtreme Gradient Boosting) · Stroke Prediction · Machine Learning

1 Introduction

Artificial intelligence is revolutionizing the world, and its application in the medical field has increased the efficiency of medical care. With the help of technology, we have witnessed astonishing developments in the field of medicine [1]. Stroke, a cerebrovascular disease, is caused by damage to brain tissue due to abnormal blood supply to the brain as a result of cerebrovascular obstruction. Stroke is responsible for most deaths worldwide [2]. The mortality rate from stroke is very high. To be precise, between 2000 and 2019, stroke is responsible for around 11% of all deaths, making it the second leading cause of death [3]. The need for rapid diagnosis of stroke has become increasingly important and warrants great interest in both clinical and fundamental research, as this rapid management will enable adequate rehabilitation to minimize late sequelae and improve patients' quality of life. With this in mind, solutions are being considered, such

A. Seeam et al. (Eds.): InterSol 2023, LNICST 541, pp. 151–160, 2024.
https://doi.org/10.1007/978-3-031-51849-2_10

as the implementation of a stroke risk prediction system based on patient data, which would enable doctors to prescribe appropriate treatments and reduce the mortality rate. The use of artificial intelligence, and more specifically machine learning, will enable us to determine which characteristics have the greatest influence on stroke prediction. By identifying high-risk patients on the basis of these characteristics, doctors can prescribe preventive treatment to avoid stroke and save the patient's life. It is therefore important to collect sufficient data on patient characteristics to improve the accuracy of stroke prediction. The application of data mining techniques to medical records has had a considerable impact on the fields of healthcare and biomedicine [4, 5]. Several studies [6–9] have analyzed the importance of patients' lifestyle and medical records on their likelihood of suffering a stroke. In addition, machine learning models have also been used to predict stroke occurrence [10, 11]. Although the above models have achieved good results, the CNN, the K Neighbors Classifier, and SVMs, respectively, have drawbacks such as complexity, low accuracy, and difficulties in kernel selection. Xgboost itself is iterative learning, which means that the model will first predict something and then analyze its error on its own, giving more weight to the data point that made the wrong prediction in the next iteration. This process continues in a loop, so technically, if a prediction is made, it's at best certain that it didn't happen by chance, but was based on a thorough understanding of the data and models. In the present paper, we attempt to fill this gap by therefore using the Xgboost algorithm on various patient records with the aim of predicting stroke, the factors likely to cause stroke are identified, the stroke prediction model is established and the 3566 data set is used in the prediction. In the remainder of this paper, we organize ourselves as follows: in Sect. 2 we give an overview of related work, in Sect. 3 we focus on the methodology used. In Sect. 4, we highlight the results obtained. Section 5 concludes the paper.

2 Related Work

Existing work in the literature has focused on different aspects of prediction. Chidozie Shamrock Nwosu et al. [8] put together an article that addresses stroke prediction from electronic health record. Shi Y et al. [12] provide a study to understand the different risk factors of stroke probability. Multivariate logistic regression (MLR) analyses were performed. The analysis showed that risk factors for stroke included hypertension, diabetes mellitus, high low-density lipoprotein, hypertriglyceridemia, and smoking compared to the control group. The analysis also showed that the risk factors for SVD cerebral stroke were hypertension, diabetes mellitus, high cholesterol, hypertriglyceridemia, and smoking. Hanifa and Raja [13] improved the accuracy of stroke risk prediction by using radial basis functions and polynomial functions in nonlinear applications applied in a support vector classification model. The risk factors identified in this work divided into four groups: demographic, lifestyle, medical/clinical, and functional. Benjamin B. et al. [14] Shows evidence of a decrease in stroke incidence emerged from population-based studies a study to identify risk factors associated with specific outcomes after stroke hospitalization, relative to stroke-specific clinical factors, using machine learning techniques [15]. Work has been explored in [16–18] to build an intelligent system to predict stroke from patient records. In the study conducted by Hong et al. [19] a comparison is

made between deep learning models and machine learning models for stroke prediction from the electronic health insurance database. Therefore, it is important for researchers to identify the different inputs. The factors in the electronic medical record are related to each other and how they are related to each other Impact on the accuracy of the final stroke prediction. Research in related areas [18] has shown that it is important to determine which features affect the final performance of a machine learning framework features affect the final performance of a machine learning framework. Thus, it is critical for practitioners of data mining in the healthcare field to determine how risk factors captured in electronic medical records relate to each other and how they affect patient health.

3 Methodology

This section presents the principle of XGBoost and its application to stroke prediction. We review the gradient tree reinforcement algorithm. We also highlight the mathematical foundations that govern it. We make minor improvements to the regularization objective, and explain how to move from the expression of a fairly generic objective function to the precise expression of the parameters required for the model.

3.1 The XGBoost Principle

XGBoost appeared in 2015 and quickly became one of the most efficient algorithms in machine learning. Its main functionalities: for regression tasks (prediction of continuous values) or classification/segmentation. It is a supervised learning algorithm that requires a set of training data to build a model that can be generalized, and is part of a so-called ensemble learning algorithm that involves the use of several decision trees to build predictions. It is particularly effective: in its ability to generalize, as it incorporates a very powerful and clever regularization mechanism into its construction, and the ability to handle missing data without degrading its performance, its speed of calculation on large volumes by making elegant approximations when constructing decision trees make XGBoost a very powerful tool. In this document, we will look at the principles that govern this algorithm and make it so powerful. It's called gradient boosting: the gradient boosting algorithmic model calls for multiple decision trees, to be used in a specific order. For an observation, each tree gives a result, and the final prediction is obtained by adding up each of the values obtained from the trees. Here's a diagram with N trees (Fig. 1).

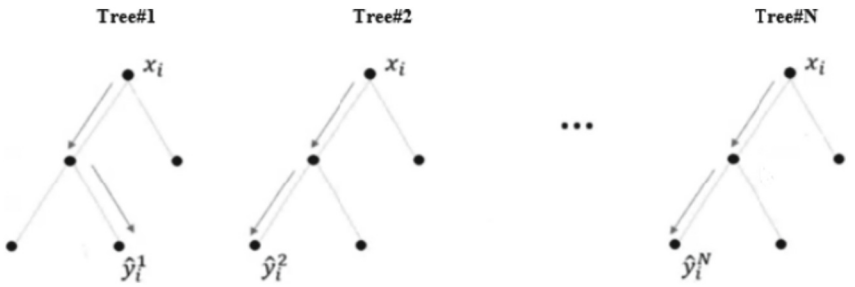

Fig. 1. Diagram with N trees in ullistration of the XGBoost function

3.2 Learning Phase and Objective Function

During the learning phase, XGBoost forces the global objective function to contain these two characteristics: a Loss function, measuring the deviation between predicted and target values, and a function to penalize model complexity, to avoid overfitting.

3.2.1 Loss Function

A Machine Learning classic for supervised learning.

The generic form is:

$$L = \sum_{i}^{\#instances} l(\hat{y}_i, y_i) \tag{1}$$

with $\hat{y}_i$: global prediction, y_i: target value and #instances: number of data used in the learning phase.

Classic example for regression - mean-square:

$$L = \frac{1}{2} \times \sum_{i}^{\#instances} (\hat{y}_i - y_i)^2 \tag{2}$$

Example for classification - Log loss (cross-entropy):

$$L = -\sum_{i}^{\#instances} (y_i . \log(\hat{y}_i) + (1 - y_i) . \log(1 - \hat{y}_i)) \text{ with } y_i = 0 \text{ or } 1 \tag{3}$$

3.2.2 Penalization (Regularization)

Without regularization, there is a high risk of overlearning. The model will then be perfect on the training data, but mediocre during the inference phase, when it will have to generalize on unknown data. In simple terms, this means that the model must perform well, and that the accuracy of the model in the training phase must not deviate significantly from the accuracy in the test phase. The risk is all the greater as decision trees have the ability to be perfect on training data, if we develop them until there is only one instance per leaf. Here, we define a function that expresses the overall complexity of the system. To do this, we sum up the complexity of each tree in the model.

$$\sum_{k}^{\#trees} \Omega(tree_{\#k}) \text{ or } \Omega(tree_{\#k}) = \gamma . T + \frac{1}{2} \lambda . \|w\|^2 \tag{4}$$

The complexity of each tree depends on two variables: T: number of leaves in the tree, $\|w\|^2$: L2 norm of the values of each tree leaf γ and λ are hyper-parameters of the global model. The larger they are, the less complexity the model will contain. Their influence on model construction will be seen later.

3.2.3 Objective Function: Loss + Penalization

As seen above:

$$L = \sum_{i}^{\#instances} l(\widehat{Yi}, Y_I) + \sum_{k}^{\#treess} (\gamma.T_K + \frac{1}{2}\lambda.\|w_k\|^2) \tag{5}$$

The shape of this function is what makes this algorithm mysterious to grasp. There are several things that prevent us from easily optimizing this function: The loss function involves the value: which is obtained by using all the trees. Indeed, classically, we try to calculate the gradient of the function to move the parameters towards its minimization. To do this, we will rewrite the objective function, using the first principle of this model's design: that of adding a new tree at each iteration, step by step, to get closer to the final prediction by adding up the values obtained. At the first iteration (t = 0), we choose a simple constant as the prediction value.

For example, take 0:

$$\hat{y}_i^{(0)} = 0 \tag{6}$$

At the next iteration, we add the result obtained from the first tree:

$$\hat{Y}_i^{(1)} = \hat{Y}_i^{(0)} + \varphi_i(x_i) \tag{7}$$

At iteration t we have:

$$\hat{y}_i^{(1)} = \sum_{k=1}^{t} \varphi_k(x_i) = \hat{y}_i^{(t-1)} + \varphi_t(x_i) \tag{8}$$

And the form of the objective function is then, at this iteration t:

$$L_t = \sum_{i}^{\#instances} l\left(\hat{y}_i^{(t-1)} + \varphi_t(x_i),\ y_i\right) + \Omega(tree_{\#t}) \tag{9}$$

So, if we want to minimize the objective function at iteration t, all we have to do is find $\varphi_t(x_i)$ that minimizes L_t. Here, we rely on reasoning by recurrence, which assumes that at iteration t, the model is already optimized for iteration t-1, so there's no need to come back to it.

In other words:

$$\varphi_t = \underset{\varphi_t}{\operatorname{argmin}} \sum_{i}^{\#instances} l\left(\hat{Y}_i^{(t-1)} + \varphi_t(x_i),\ Y_i\right) + \Omega(tree_{\#t}) \tag{10}$$

This rewrite of the function will enable step-by-step optimization, which will be much simpler to implement.

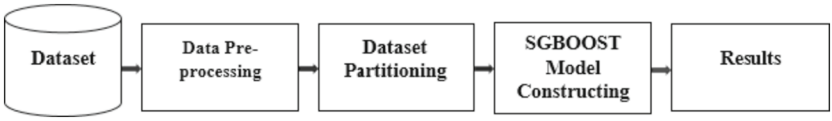

Fig. 2. Proposed methodology

3.3 The Application of XGBoost in Stroke Prediction

The first phase consists of acquiring the data to be used later. In phase 2, we pre-process the data, i.e. data cleansing, conversion of object-type data to float or int, feature selection and extraction, etc. In phase 3, the data set is divided in two respectively for training and testing. In phase 4, an XGBoost model is built. In this phase, we already have the pre-processed data, which enables us to build an XGBoost model. Here are the steps we follow: a) we start by importing the required libraries: from sklearn.model_selection import train_test_split, from sklearn.pipeline import Pipeline, from category_encoders.target_encoder import TargetEncoder, from xgboost import XGBClassifier. b) Installing the category_encoders library: this step is designed to perform the encoding, transforming the categorical data into a numerical form understood by the XGBoost algorithm. c) Definition of pipeline steps: We define the pipeline steps, creating a sequence of operations that will be applied sequentially during modeling. d) Optimizing the hyperparameters of the XGBoost model: we use a hyperparameter optimization process for an XGBoost model using the scikit-optimize library, as we explained mathematically the larger the hyperparameters, the less complexity the model will contain, so we need to find the best values for the hyperparameters to improve its predictive performance. To do this, we create a BayesSearchCV object that explores different combinations of hyperparameters to determine which give the best results according to the specified score metric. e) BayesSearchCV object fit: we fit the BayesSearchCV object to the training data (X_train and y_train) using the opt.fit(X_train, y_train) statement. Once optimization is complete, this attribute returns the best estimator (model) found by optimization, i.e. the model with the best hyperparameter values. You can obtain the best estimator using the opt.best_estimator_ statement. f) Model training: finally, we train the model using 50% of the randomly selected data to form the training set, with the remaining 50% of the data used as a verification set to adjust the model's prediction parameters. In phase 5, we obtain the results (prediction, most influential features in prediction (feature importance)) (Fig. 2).

3.4 Data Description

The dataset is available on Kaggle [20], a public data repository for datasets. The database contains the records of 5110 patients. It has a total of 11 input attributes and one output feature. The output response is a binary state indicating 1 if the patient has suffered a stroke and 0 means they have not. The other 11 input attributes are described in the table below (Table 1):

Table 1. Representative example of the dataset

Feature	Description
Id	the patient's identifier
Gender	The sex of the patient is indicated by "Male", "Female"
age	the patient's age
hypertension	the binary state (whether the patient has (1) or not (0) hypertension (HT)
heart_disease	Binary status (whether the patient has heart disease (1) or not (0))
ever_married	"No" to not get married and "Yes" to get married
work_type	"children" for the patient who has children, "Govt_job" for the patient who works in the administration the patient works in the administration, "Never_worked" for the patient has never worked, "Private" for the patient working in the private private work, or "Self-employed" for the patient who works for themselves
Residence_type	"Rural" if the patient lives in a rural area and "Urban" if the patient lives in an urban area
avg_glucose_level	the patient's average blood glucose level
bmi	the patient's BMI
smoking_status	"Formerly smoked" represents the patient who used to smoke, "never smoked" means the patient has never smoked, "smoked" represents the observer who currently smokes
stroke	1 or 0, i.e. 1 for a stroke, 0 for no stroke

4 Result

XGBoost can predict whether patients are having a stroke based on relevant information about them in the dataset. In this section, we perform a comparative analysis of three popular classification algorithms: CNN, SVM and XGBoost on our patient dataset. The accuracy rate is 0.83, which means that almost 8 out of 10 people can actually be predicted when they are sent to hospital and can be detected and receive the real treatment. We use classification accuracy as a measure of the performance of machine learning models. Table 2 presents the average classification for the three benchmark tests.

Table 2. Average binary classification accuracy of XGBoost, SVM and CNN over 3565 experiments on our dataset.

Approach	Average accuracy
XGBoost	82.8%
SVM	79.2%
CNN	72.16%

The following figures, Fig. 3 and Fig. 4, show the confusion matrix of our XGBoost model and the most important features in stroke prediction.

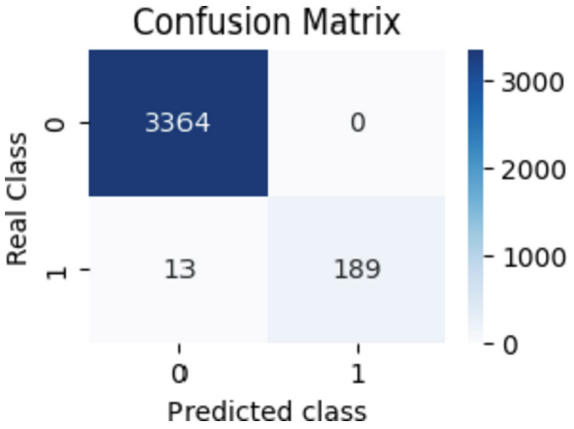

Fig. 3. Result measured by confusion matrix

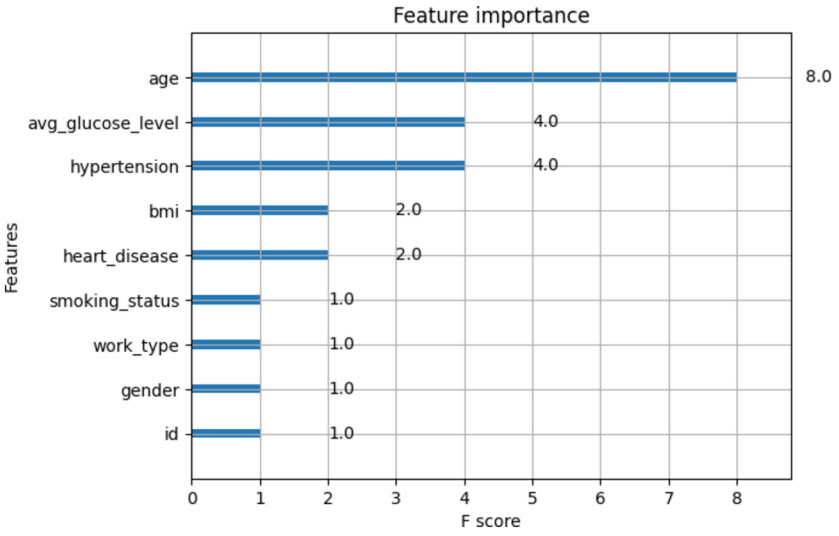

Fig. 4. Importance of patient attributes in predicting stroke occurrence using Xgboost

Figure 4 shows the importance of individual patient attributes in predicting stroke. The relative importance of a patient attribute is measured by the increase in the model's prediction error due to that attribute. We use the XGBoost estimator to calculate the relative importance of features. Patient age (A) is the most important feature for predicting stroke. Other very important characteristics are the patient's average blood sugar level and whether he or she suffers from high blood pressure.

5 Conclusion

Early detection and treatment of cerebrovascular disease is essential to reduce morbidity and mortality. Current applications of artificial intelligence in this field offer enormous potential for improving treatment choice and clinical outcomes at all stages of diagnosis and treatment, including outcome prediction. In this paper to solve the problem of stroke prediction, the XGBoost method is proposed. Two experiments are carried out in this paper to verify the above proposal: the first shows that, with regard to stroke prediction, the accuracy of XGBoost is 0.83, so we can conclude that XGBoost is capable of effectively predicting stroke, and that it performs very well and even better than other machine learning methods. The second is that there are a large number of characteristics, so in order to simplify the work of doctors and be efficient, we had to focus on those characteristics that had a strong impact on prediction. We found from Fig. 4 that characteristics such as age, hypertension, cholesterol level, body mass index and heart disease were more relevant, and in the same order, by way of verification, we created a new dataset considering only these 5 characteristics, and we obtained an accuracy with XGBoost of 0.825, roughly equal to that with all characteristics. This study has certain limitations in certain situations, as it is important to point out that other factors may be at the origin of a stroke; for example, many patients may even inherit from their parents. Furthermore, patients can only be correctly predicted if their health statistics are exactly the same as those of their parents. What's more, the data we've used comes from another part of the world, where certain characteristics differ from one place to another in Africa, for example, climate can influence hypertention, quality of life and lifestyle can influence stress, and so on. What's more, the majority of hospitals in Africa keep their data in physical rather than electronic form. Consequently, it will take some time to implement this predictive model in reality for some countries. However, although there are still limitations, stroke prediction remains very useful. In short, stroke prediction based on XGBoost can be applied to real-life problems. In future projects, the main contribution will be to be able to confirm the results of this article with local data from Senegal, which we have begun to collect from hospital facilities.

References

1. Sivapalan, G., Nundy, K., Dev, S., Cardiff, B., John, D.: ANNet: a lightweight neural network for ECG anomaly detection in IoT edge sensors. IEEE Trans. Biomed. Circ. Syst. **16**(1), 24–35 (2022)
2. Pastore, D., Pacifici, F., Capuani, B., et al.: Sex-genetic interaction in the risk for cerebrovascular disease. Curr. Med. Chem. **24**, 2687–2699 (2017)
3. The top 10 causes of death. https://www.who.int/news-room/factsheets/detail/the-top-10-causes-of-death. Accessed 22 June 2023
4. Koh, H.C., Tan, G.: Data mining applications in healthcare. J. Healthc. Inf. Manag. **19**(2), 64–72 (2011)
5. Yoo, I., et al.: Data mining in healthcare and biomedicine: a survey of the literature. J. Med. Syst. **36**(4), 2431–2448 (2012). https://doi.org/10.1007/s10916-011-9710-5
6. Meschia, J.F., et al.: Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke **45**(12), 3754–3832 (2014)

7. Harmsen, P., Lappas, G., Rosengren, A., Wilhelmsen, L.: Long-term risk factors for stroke: twenty-eight years of follow-up of 7457 middle-aged men in Goteborg, Sweden. Stroke **37**(7), 1663–1667 (2006)
8. Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from electronic health records. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, pp. 5704–5707. IEEE (2019)
9. Pathan, M.S., Jianbiao, Z., John, D., Nag, A., Dev, S.: Identifying stroke indicators using rough sets. IEEE Access **8**, 210318–210327 (2020)
10. Kim, J., Hong, D., Park, S.: A case-control study of risk factors for cerebrovascular disease. J. Prev. Med. **28**, 473–486 (1995)
11. Park, J.K., Kang, M.G., Kim, C.-B., et al.: A meta-analysis on the risk factors of cerebrovascular disorders in Koreans. J. Prev. Med. Public Health **31**, 27–48 (1998)
12. Shi, Y., et al.: Risk factors for ischemic stroke: differences between cerebral small vessel and large artery atherosclerosis aetiologies. Folia Neuropathol. **59**(4), 378–385 (2021)
13. Hanifa, S.M., Raja-S, K.: Stroke risk prediction through nonlinear support vector classification models. Int. J. Adv. Res. Comput. Sci. **1**, 4753 (2010)
14. Clissold, B.B., Sundararajan, V., Cameron, P., et al.: Stroke incidence in Victoria, Australia—emerging improvements. Front. Neurol. **8**, 180 (2017)
15. Rana, S., et al.: Application of machine learning techniques to identify data reliability and factors affecting outcome after stroke using electronic administrative records. Front. Neurol. **12**, 670379 (2021)
16. Khosla, A., Cao, Y., Lin, C.C.Y., Chiu, H.K., Hu, J., Lee, H.: An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 183–192 (2010)
17. Hung, C.Y., Lin, C.H., Lan, T.H., Peng, G.S., Lee, C.C.: Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. PLoS ONE **14**, e0213007 (2019)
18. Teoh, D.: Towards stroke prediction using electronic health records. BMC Med. Inform. Decis. Making **18**(1), 1–11 (2018)
19. Hung, C.Y., Chen, W.C., Lai, P.T., Lin, C.H., Lee, C.C.: Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3110–3113. IEEE (2017)
20. Fed Soriano, Stroke Prediction Dataset. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-datase. Accessed 23 June 2023