ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse289952020.pdf

https://doi.org/10.30534/ijatcse/2020/289952020

# Stroke Prediction Using Machine Learning based on Artificial Intelligence

**Youngkeun Choi** [1], **Jae Won Choi** [2]

[1] Associate Professor, Division of Business Administration, College of Business, Sangmyung University Seoul, Korea, penking1@smu.ac.kr

[2] Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA, jxc190057@utdallas.edu

## ABSTRACT

Machine learning technology is implemented to analyze different kinds of medical problems. Stroke is complicated, and every year it causes a lot of death. If the early symptoms of stroke are ignored, in a short period, the patient may end up having drastic consequences. For this, this study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in stroke modeling better. Secondly, the study seeks to evaluate the predictive performance of the decision trees. This study analyzed the dataset of 43,400 patients and 11 predicted attributes from a third-party website, international challenge on the popular internet platform Kaggle. Based on these results, first, All of predicative variables except bmi influence stroke significantly. Second, for the full model, the accuracy rate is 0.981, which implies that the error rate is 0.019. Among the patients who were predicted not to have stroke, the accuracy that would not have stroke was 98.17%, and the accuracy that had strike was 16.67% among the patients who were predicted to have stroke.

**Key words :** Stroke; Artificial intelligence, Machine learning; Decision tree.

## 1. INTRODUCTION

Machine learning and artificial intelligence (AI) are the first choices for data mining and big data [1]. It offers a number of applications, including technologies such as neural network modeling, simulation models, DNA computing, and quantum computing. AI in the field of biomedical science significantly reduces random problems when processing this type of data. Many technological advances have helped AI technologies evolve in a way that facilitates efficient and convenient processing of such data. This means that machine learning and AI models are also used in advanced data analysis and optimization approaches, such as drug design and analysis, medical imaging, biological inspiration learning and adaptation to analysis.

A brief description of the function of the machine learning algorithm is learned from previously diagnosed patient cases [2]. To save lives, you need to diagnose heart problems quickly, efficiently and accurately. Researchers have become interested in predicting the risk of disease and using a variety of machine learning techniques to create a variety of health care risk forecasting systems. Missing and outlier data in training sets often degrade model performance and inaccurate predictions. Therefore, it is important to process missing and specific values before making a prediction.

Stroke has become one of the most significant threats to public health worldwide [3]. Stroke disease ranks second in the post-cardiac life span. Stroke is a sudden onset of focal neurological defects that last more than 24 hours. And it's caused by cerebral artery occlusion or bamboo sclerosis. Signs of stroke appear suddenly, but often occur gradually. In the United States in 2016, the number of stroke patients increased sharply, causing significant load on the health care system. Stroke disorders cause physical, mental and financial burdens on patients, families and communities, but early detection is considered to improve healing and reduce disability. Early prediction of stroke disease is useful for prevention or early treatment intervention. Machine learning and data mining play an essential role in stroke forecasting, such as support vector machines, logistic regression, random forest classifiers and neural networks. Machine learning is a form of artificial intelligence aimed at building computers with human thinking abilities. The goal of machine learning is to enable the computer to perform certain tasks that rely on patterns and interference without using clear instructions [4].

The purpose of this study is to find and analyze the causes of stroke so that doctors and scientists can use this study to formulate possible solutions to problems. The pre-access and modeling methodology used in this white paper can be seen as a roadmap for readers to follow the steps taken in this study and apply procedures to identify the causes of many other medical problems.

## 2. RELATED STUDY

Classification in medicine is one of the most important, important and widely used decision-making tools. Many modern techniques have been introduced to accurately and accurately predict strokes. Some tasks related to this area are described briefly as follows:

Some researchers used machine learning algorithms to predict stroke. This section describes the contributions of some research studies. Shanthi, et al. [5] used an artificial neural network (ANN) to predict thrombotic embolism stroke disease. Medical data set stroke data with eight important attributes of the patient was used. This study shows an ANN-based prediction of stroke disease by improving accuracy to 89% at a high consistent rate. ANN shows the appropriate performance level for predicting stroke conditions. The researchers found that DT was the best sorter among other methods used. Adam et al. [6] used the Support Vector Machine (SVM), Stochastic Gradient Boosting (SGB), and PAL (Penalized Logistic Regulation) to predict the stroke of a data set collected from TongutOzal Medical Center in Malatia, Turkey. Studies show that SVMs achieve a 98% peak accuracy. The results of this study showed that LR is best suited to the classification of acute ischemic stroke compared to ANN. The results showed that the support vector machine (SVM) achieved a higher area of the ROC curve compared to the Cox proportional risk model. Adam et al. [6] compared KNN with two algorithm decision trees for stroke classification on data sets at Tamil Nadu Kum Baco Nam Prison in India. And the researchers concluded that the classification of the decision tree was better than the KNN algorithm. Cheng et al. [7] also worked on predicting ischemic stroke using two ANN models for data sets from Indian Tamil Nadu Kum Baco Nam Prisoner General Hospital. And the researchers concluded that accuracy reached 79.2% and 95.1%.

## 3. METHODOLOGY

### 3.1 Dataset

The corresponding variable information is drawn from a third-party website, international challenge on the popular internet platform Kaggle (www.kaggle.com), which provides data in the title of ‘Healthcare Dtaset Stroke Data’ that was uploaded by Saumya Agarwal. It contains the data of 43,400 patients and 11 attributes, including the predicted attribute. The Kaggle asked participants to predict heart disease. To help with algorithmic development, the organizers provided the types of a data stream for a large set of individual factors. These variables are listed and defined in Table 1.

**Table 1:** The variables in each category

| Variables | Definition |
|---|---|
| Id | Patient ID |
| Gender | Gender of Patient |
| Age | Age of patient |
| Hypertension | 0 - no hypertension, 1 - suffering from hypertension |
| Heart_disease | 0 - no hypertension, 1 - suffering from heart disease |
| Ever_married | Yes/No |
| Work_type | Type of occupation |
| Residence_type | Area type of residence (Urban/Rural) |
| Avg_gluscose_level | Average of Glucose level (measured after meal) |
| Bmi | Body mass index |
| Smoking_status | Patient’s smoking status |
| Stroke | 0 - no stroke, 1 – suffered stroke |

### 3.2 Decision Tree

Among the various analytical techniques, decision tree (DT) is a powerful and widely used machine learning algorithm to predict and classify medical data to date [8]. Used for both classification and regression issues. Now you may have questions about why you want to use DT classifiers rather than other classifiers. I can give you two reasons to answer that question. One is that because decision trees often try to imitate the way the human brain thinks, understanding data and making good conclusions or interpretations is very simple. The second reason is that the decision tree allows you to see the logic that the data interpret, not the black box algorithms such as SVMs and NNs. It has simple and clear expertise and has become one of the favorites among the programmers of this generation. Now we've looked at why we can look at the decision tree in more detail at what the decision tree categorizer is. The decision tree start is a tree with multiple nodes, each node represents a function (properties), each link represents a decision called a rule, and each leaf in the tree represents a different known result. The idea of category type or duration is to create a tree for the entire data and get results from all riffs. Now we know a little bit more about the decision tree. We will continue to discuss how to create a decision tree separator. A decision tree can be built with two algorithms. One is Cart (classification and regression tree) and the other is ID3.

For ID3, first use the x and y values in the column. This value remains in the last position in the column and only has a "YES" or "NO" value. The above chart has x values (view, temperature, humidity, wind) and only two options 'YES' or 'NO' are at the end of the column or are y values. You must now map x and y. As you can see, this is a binary classification problem, so I'm going to use the ID3 algorithm to build the tree. To create a tree, you must first select the root node to be the root node. A general rule of thumb is to first select the function that most affects the y-value as the root node. Select the next most influential function as the next node. Here we're going to use the entropy concept. Entropy concepts measure the degree of uncertainty in a data set. Entropy must be calculated for all category type values of binary classification problems. In summary, the entropy of the data set should be calculated first. For all properties / functions, first calculate the entropy for all category type values, then import the average value information entropy for the current property to calculate the amount obtained for the current property. Then you must select the highest gain property and repeat it until you get the tree you want. This is the process of ID3.

As discussed above, decision tree classifiers were written by a different algorithm called Cart that represents classification and regression trees. This algorithm uses Gini Index as a cost function used to evaluate segmented anger in a data set. Here, the target variable is actually a binary variable, so you use two values (Yes and No). As we all know, there can be four combinations. Now you need to understand the Gini score to get a good idea of how to divide the data. If the Gini score is 0, the worst-case scenario is 50/50 split, but it is a complete separation. The problem is now how to calculate the Gini index value.

The Gini index is similar even if the target variable is a category type variable at a different level. Therefore, the step in this method is the first calculation of the Gini index for a data set. Then you must calculate the Gini index for all category type values for all functions, and then import the average information entropy for the current property and eventually calculate the Gini gain. When you complete this task, you can select the best Gini gain properties and repeat them until you get the tree you want. This is how the decision tree algorithm works.

### 3.3 Data mining models

In order to survive in a competitive market, many companies use data mining technology for decision-making predictive analysis. Building an effective and accurate decision-predictive model is more important than managing your customers effectively. Statistical and data mining techniques were used to construct decision forecasting models. Data mining technology can be used to predict or classify behavior by discovering interesting patterns or relationships in data and fitting models based on available data. If the learning data set and the test data set are separated for machine learning, the test data set must meet the following requirements: You first need to create a learning data set and a test data set in the same format. Second, the test data set should not be included in the training data set. Third, the learning data set and the test data set must be consistent with the data. However, it is very difficult to create a test data set that meets these requirements. Various verification frame works have been developed in data mining using one data set to address this problem. This study supports using the Split Validation operator provided by RapidMiner. The operator divides the input data set into training and test data sets to support performance evaluation. This study selects relative segmentation among the segmentation method parameters of this operator and uses 70 percent of the input data as learning data.

### 3.4 Performance evaluation

Performance assessments determine how well a model created using learning data works. Performance measurements can be divided into technical performance measurements and heuristic measurements. The technical performance measurements used in this study produce a model from training data, process the test data as a model, and compare the class labels of the original verification case to the predicted class labels to show performance results. Technical performance measurements can be divided into learning and learning and non-study. The learning used in this study is also classified and reversed. All data used for this learning and testing will have the original class values. Obtain performance by comparing and analyzing the original class values with the predicted results.

Classification issues are the most common data analysis issues. Various metrics have been developed to measure the performance of classification models. Classification problems of category types are often characterized by accuracy, precision, recall, and f measurements. RapidMiner includes performance (classification) that measures performance metrics for common classification problems and performance (differential classification) that provides performance metrics for binary classification problems. Table 2 shows how these metrics are calculated.

**Table 2:** Key performance indicators of binomial classification

| | | **Actual class (as determined by Gold Standard)** | |
|---|---|---|---|
| | | True | False |
| **Predicted class** | Positive | True Positive | False Positive(Type □ error) |
| | Negative | False Negative(Type □ error) | True Negative |

Precision = TP/(TP+FP), Recall = TP/(TP+FN), True negative rate = TN/(TN+FP), Accuracy = (TP+TN)/(TP+TN+FP+FN), F-measure = 2·((precision·recall)/(precision + recall))

## 4. RESULTS

### 4.1 Decision tree

Figure 1 shows the classification tree for the full model after pruning the tree using cross-validation to avoid overfitting [9].

**--------------------------------------------**

Insert Figure 1 here

**--------------------------------------------**

The key variables in the full model analysis consist of 11 ones, as shown below, based on the criterion established with each of these variables. In other words, the classifier has identified four potential questions along each of these variables and specific criteria as defined below to aid in the classification of unknown patients. All of predicative variables except bmi influence stroke significantly.

Tables 3 illustrate each of the confusion matrix measures. For the full model, the accuracy rate is 0.981, which implies that the error rate is 0.019. Among the patients who were predicted not to have stroke, the accuracy that would not have stroke was 98.17%, and the accuracy that had strike was 16.67% among the patients who were predicted to have stroke.

**Table 3:** Performance evaluation

| | **True 0** | **True 1** | **Class precision** |
|---|---|---|---|
| **Pred. 0** | 12,776 | 238 | 98.17% |
| **Pred. 1** | 5 | 1 | 16.67% |
| **Class recall** | 99.96% | 0.42%% | |

## 5. CONCLUSION

Stroke is complex and causes many deaths each year. Ignoring the initial symptoms of stroke can lead to serious patient consequences in a short period of time. Given this situation, the stroke was predicted using a Kaggle data set using machine learning techniques. Medical information is enormous, and this study shows how to effectively use vast amounts of data to predict stroke using machine learning techniques.

In summary, this study had two main goals in essence. First of all, this paper seeks to better understand the role of variables in stroke modeling. Second, this study tries to evaluate the predictive performance of the decision tree. Based on the above results, a series of meanings are derived. Regarding the first goal, the study suggests that assessing the role of a variable is complex and its impact depends on the classification method used. The decision tree method emphasizes the most important explanatory power in the analysis. Therefore, it is not possible to draw a unanimous conclusion on which explanatory variables are most important for stroke in all the methods used as a whole.

However, the results of this study provide additional information about the patient's profile. Doctors should predict strokes for the classification methods used. For example, all predictors except bmi have a significant effect on stroke. Second, for the entire model, the accuracy is 0.981 and the error rate is 0.019. The non-stroke accuracy rate of patients expected to have no stroke was 98.17% and the expected stroke accuracy was 16.67%.

This study provides some research contributions and practical implications. First, the study expands existing literature by experimenting with the combined effects of variables on stroke modeling. Stroke has a great effect on the patient. A lot of research has been done on stroke, but no one can say that we can create a universal human tool to predict stroke. Stroke is so complex and associated with so many factors that researchers tend to use fewer elements and ignore the effects of others. Patient demographics are often changed and monitored continuously, which can cause hospital problems and damage to personal information. Some studies examined age, gender, and geographic location. But researchers still cannot express cultural and behavioral factors that can affect stroke. This study contributes to the literature on strokes by providing a global model that summarizes the stroke determinants of individual factors in patients. Second, the methodology used in this paper can be seen as a roadmap to follow the steps taken by the reader in this case study and apply a day-long procedure to identify the causes of many other problems. This paper proposes the best performance model for predicting stroke based on a limited set of functions, including patient factors to achieve the best results in terms of accuracy using machine learning techniques and functional importance analysis, including decision trees and neural networks. In this way, the study identified stroke patterns that could predict a patient's stroke.

In fact, this application helps doctors manage patient health records and speed up treatment if you already have patient reports. Quick treatment saves lives. This application helps patients track their health records. Therefore, it is helpful to take care of your health regularly. Analyst reports help doctors easily and easily predict strokes. The proposed system has a database that stores patient records, and as the number of patients increases, more data is generated and storage is a problem. Therefore, in future releases, we will provide the cloud capability to store all records in the cloud. So if you have the right to protect and access your data, you can search everywhere. Smart devices are synchronized with applications in future releases. This monitors the patient's real-time health and alerts you in case of an emergency. This reduces risk.

In the future, the machine learning model will use a larger training data set that uses more than one million different data points maintained in electronic health recording systems. While calculations and software sophistication can be a big leap forward, a system that works with artificial intelligence allows doctors to provide the best care to the patient as soon as possible. Software APIs can be developed for free access to health websites and apps to patients. Probability predictions are made regardless of whether processing is delayed or not.

## REFERENCES

1. S. R. Guruvayur, and R. Suchithra. **A detailed study on machine learning techniques for data mining**. In *Proc. 2017 International Conference on Trends in Electronics and Informatics (ICEI)*
2. N. G. Molty, and S. Das. **Machine learning for improved diagnosis and prognosis in healthcare**. In *Proc. 2017 IEEE Aerospace Conference.*
3. M. Katan, and A. Luft. ***Global burden of stroke. in Seminars in neurology***. Thieme Medical Publishers, 2018.
4. T. T. Jos, L. G. Cindy, and M., Z. Elviawaty. **A review on applying machine learning in game industry**, *International Journal of Advanced Science and Technology*, Vol. 28, No. 2, pp. 258-264, 2019.
5. D. Shanthi, G. Sahoo, and N. Saravanan. **Designing an artificial neural network model for the prediction of thrombo-embolic stroke,** *International Journals of Biometric and Bioinformatics*, Vol. 3, No. 1, pp. 10-18, 2009.
6. S. Y. Adam, A. Yousif, and M. B. Bashir. **Classification of ischemic stroke using machine learning algorithms**. *International Journal of Computer Application*, Vol. 149 No. 10, pp. 26-31, 2016.
7. C. A. Cheng, Y. C. Lin, and H. W. Chiu. (2014). **Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks**. In *Proc. 2014* ICIMTH.
8. J. M. Gonzalez-Cava, J. A. Reboso, J. L. Casteleiro-Roca, J. L. Calvo-Rolle, and J. A. M. Pérez. **A Novel Fuzzy Algorithm to Introduce New Variables in the Drug Supply Decision-Making Process in Medicine.** *Complexity*, https://doi.org/10.1155/2018/9012720, 2018.
9. M. Kuhn, and K. Johnson. ***Applied predictive modeling***. New York: Springer, 2013.

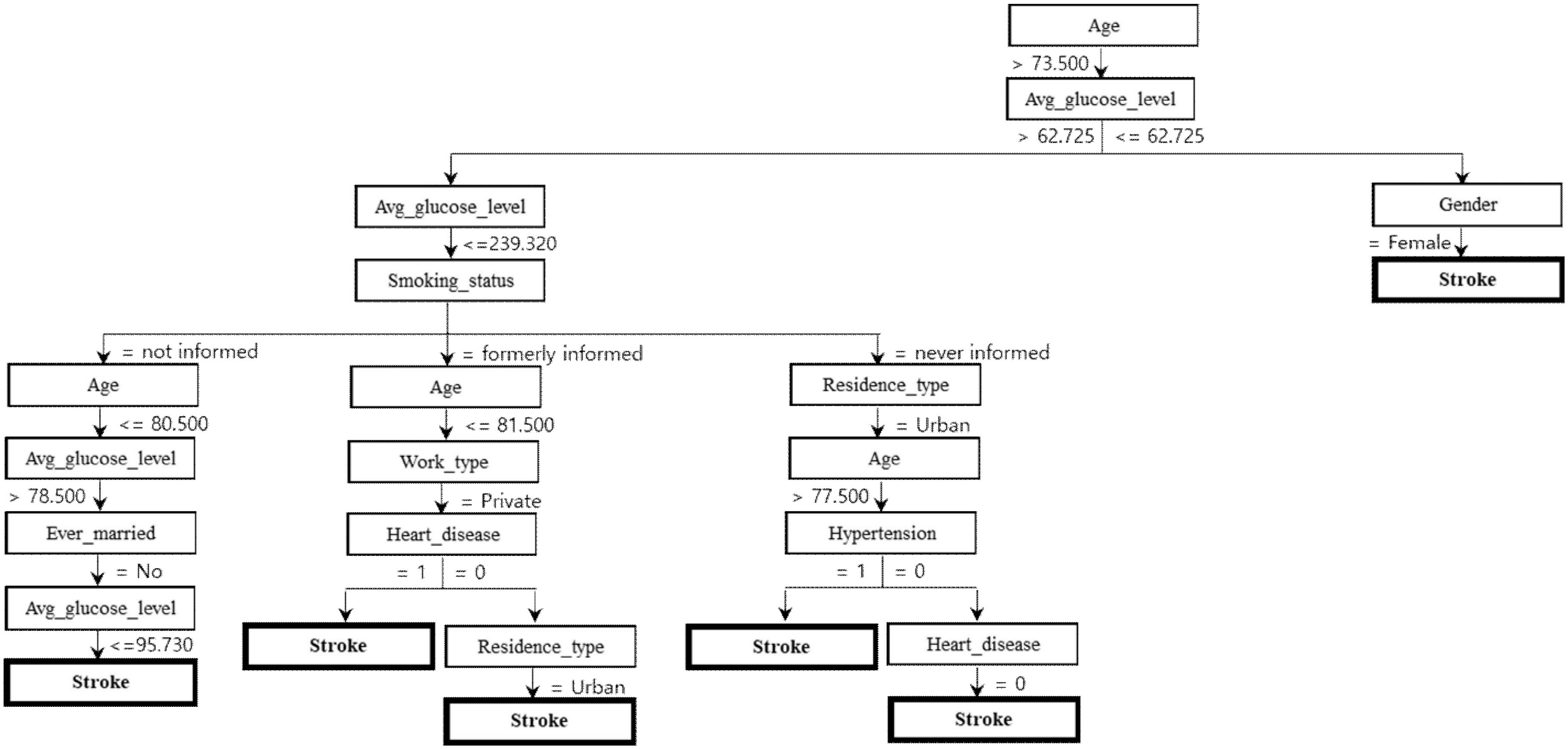

**Figure 1:** Classification Tree for the Full Model