**RISK PREDICTION OF STROKE USING DATA MINING CLASSIFICATION TECHNIQUES**

1st Rahatara Ferdousia, 2nd Masum Mohammad Jubayelb, 3rd Asima Akter Chowdhury

*a Metropolitan University, Sylhet 3100 Bangladesh.*

[*rahatara@metrouni.edu.bd*](mailto:rahatara@metrouni.edu.bd)

*b Metropolitan University, Sylhet 3100 Bangladesh.*

[*masumjubayel@gmail.com*](mailto:masumjubayel@gmail.com)

*c Metropolitan University, Sylhet 3100 Bangladesh.*

[*chowdhuryasima@gmail.com*](mailto:chowdhuryasima@gmail.com)

1. **Abstract**

Stroke, a fatal non-communicable disease of any age, kills more people than AIDS, Tuberculosis and Malaria put together in each year. WHO estimated around 6.2 million deaths because of stroke in 2008. As the incidence, prevalence, mortality and disability rates are increasing, overall stroke burden has increased globally. Almost 70% of patients are unaware of their mild stroke, 30% seek medical attention lately and another 30% suffer from recurrent stroke, before seeking attention. Data mining, with its several techniques for classification and regression, plays a leading role in developing an effective model of risk prediction in the context of healthcare. Even though stroke prevention is a complex medical issue, primary prevention could be feasible by using data mining classification techniques that will assess risk factors to predict the likelihood of the disease among mass people. This work is aimed at providing an analysis of different data mining classification algorithms like Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), etc. on a newly created dataset of 435 patient’s risk factors to find the algorithm with the best accuracy to propose a tool for the end users to check risk prediction.

*Keywords***:** Stroke risk factor, Data mining, Naïve Bayes, Random forest, Risk Prediction.

1. **Introduction**

A stroke happens when blood flow is occluded in a part of the brain. The lack of oxygen damages the brain cells that can have potentially disabling effects on the patient. (Services, 2018) When it happens the brain cells divested of oxygen and brain cells begin to die. Then the abilities controlled by that area of the brain such as memory and muscle control are lost. Some strokes affect the muscles used to urinate. There are two types of stroke. They are

An Ischemic stroke is the most common type. It occurs when blood supply is cut off to part of the brain. It accounts for the majority of all strokes. Approximately 85% of strokes are ischemic caused by vascular occlusion.

An ischemic can occur because of lesions caused by atherosclerosis. These lesions may form in the small arteries of the brain and they can block blood flow to the brain. (Services, 2018)

A haemorrhagic stroke is a different kind of stroke caused by bleeding in the brain. It happens when a blood vessel breaks and bleeds into the brain. Blood spills into or around the brain and creates swelling and pressure, damages cells and tissue in the brain.

Haemorrhagic stroke is caused by a rupture in a weakened blood vessel in the brain. Haemorrhagic stroke account for about 20% of all strokes. (Services, 2018)

In this thesis paper, we will predict the risk of stroke. A man who has not yet stroked will talk about the possibility of having a stroke in the future. It is a more challenging task in healthcare sectors to predict the diseases from the voluminous medical databases. At present, data mining techniques will help us a lot to predict risk. Data mining techniques which include classifications, clustering, association rule mining for finding risk prediction. In this research work, Naïve Bayes Support Vector Machine (SVM) classifier algorithm is used for stroke risk prediction.

The brain is one of the largest and most complex organs in the human body. It is made up of more than 100 billion nerves that communicate in trillions of connections called synapses. [1] It is the central organ of the human nervous system. It controls most of the activities of the body, processing, integrating and coordinating the information it receives from the sense organs and making decisions as to the instructions sent to the rest of the body. Because of an ischemic stroke, brain cells start to die or damage. So, symptoms occur in the body parts (face, eyes, arms, legs etc.) that these brain cells control. There is the number of factors which increase the risk of having an ischemic stroke. Some of them are listed below:

* Age and gender
* High blood pressure
* Diabetes
* Ischemic heart diseases
* Smoking
* The family history of stroke
* Stress and depression
* Overweight and obesity
* Abnormal cholesterol levels

1. **Structure:**

The figure1 showed the system architecture of stroke risk prediction. Here the train data set are collected from the reports of some patients from different four hospitals in Sylhet, Bangladesh. The training dataset was pre-processed in the pre-processing stage. Then we tested the data set with some algorithms like SVM, ANN, NB, RF, J48, DT, LG etc. and we evaluated the performance accuracy of all those algorithms within 10-Fold Cross Validation and Percentage Split techniques. According to the best accuracy, the best algorithm will be chosen for developing the tool. A questionnaire form will be filled up by the user as system input to predict the risk level and to give some tips and suggestions to the end user.

Report based train dataset

Database

Select the best algorithm

J48

RF

SVM

ANN

NB

Preprocessing

Use of algorithms for risk prediction

 J

Performance evaluation of algorithms

Risk level, tips and suggestions

Checking feature end user tool

End User

Input data

Fig 1: System Architecture for stroke risk prediction

1. **Methodology**
   1. **Naïve Bayes**

Naïve Bayes is a type of classification which follows Bayes theorem and also known as probabilistic classifier method [5]. Naïve Bayes algorithm uses to predict the probability of the result or the output for an unseen or unlabeled test input. There are 4 equations are given below those show how the classifier works and we have two classes namely positive (pos) and negative (neg) corresponds to with stroke and without stroke. (Ray, 2017)

P (pos| A) = p (pos|a1) \* p (pos|a2) \*………p (pos |an) \* p(pos ) (1)

P (neg| A) = p (neg|a1) \* p (neg|a2) \*………p (neg |an) \* p(neg ) (2)

P (Ai |pos)

Here, I start at 1 and increase until it reaches the last number of our train attributes.

* 1. **Random forest**

Random forest consists of many decision trees is an algorithm which can work for both classification and regression problem. It follows bagged methods. We showed an equation for a random forest in equation (3). If there are X= x1,,,,,xn and Y= y1,,,,,,,,yn are two variables then for making decision random forest takes B time replacement and takes the final decision by pooling or voting all the results of those replacements. It predicts the result for an unlabeled or unseen data x

(3)

1. **Experimental Analysis**

In this section, we represented the details of our dataset and the result analysis.

**4.1 Details of the training dataset**

Our train dataset contains information about 435 persons. We have collected our dataset from four different hospitals in Sylhet, Bangladesh. At first, we preprocessed our data in the preprocessing step. In this step, we handled our missing data. For collecting data, we made questionnaire form and then collected the data by filling up those questionnaire forms from the test report of the people who already have the stroke and we also collected data from those people who have not stroked but they also have similar problems**.**

Table 1: Description of the training dataset

|  |  |  |
| --- | --- | --- |
|  | Number of Attributes | Number of Instances |
| Report based dataset | 15 | 435 |

In table 1, we gave the description of the dataset and we gave the description of our attributes in table 2. We have collected 435 data and we have 15 attributes. Among 435 data, there are 342 are positive data and 93 are negative data. Class attribute has two variables those are stroke and non-stroke.

Table 2: Description of attribute

|  |  |
| --- | --- |
| Attributes | Values |
| Age | 1.25-34, 2.35-44, 3.45-54,4.55-65,5.65< |
| Gender | 1. Male 2. Female |
| SBP | 1.120>, 2.120-139, 3. 140-160, 4.160< |
| DBP | 1.180>, 2.80-95, 3.95< |
| Diabetes | 1. No, 2. Yes |
| Ischemic Heart Disease | 1. No, 2. Yes |
| Family History of stroke | 1. No, 2. Yes |
| Alcoholic | 1. No, 2. Yes |
| Less Physically Active | 1. No, 2. Yes |
| Smoking | 1. No, 2. Yes |
| Stress and depression | 1. No, 2. Yes |
| Situated Fat↑ | 1. No, 2. Yes |
| Fibre↓ | 1. No, 2. Yes |
| CKD-Kidney | 1. No, 2. Yes |
| Class Attribute | 1. Stroke, 2. Non-stroke |

**4.2 Details of the result analysis**

In our work, we wanted to provide an analysis of different data mining classification algorithms like SVM (Support Vector Machine), NB (Naïve Bayes), RF (Random Forest), J48, ANN( Artificial Neural Network ), LR (Logistic Regression), RT (Random Tree) etc. We applied those classification algorithms on our dataset and we evaluated the performance accuracy of each classification algorithms through both 10-Fold Cross Validation and percentage split evaluation techniques. That’s also shown in table 3.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Evaluation Metrics | 10- Fold Cross Validation | | | | | | | Percentage Split (80:20) | | | | | | |
| NB | SVM | LG | J48 | RT | RF | ANN | NB | SVM | LG | J48 | RT | RF | ANN |
| Total number of the instance | 435 | 435 | 435 | 435 | 435 | 435 | 435 | 87 | 87 | 87 | 87 | 87 | 87 | 87 |
| Correctly Classified  Instance | 345 | 342 | 352 | 365 | 332 | 358 | 353 | 69 | 66 | 67 | 70 | 65 | 68 | 69 |
| 79.31% | 78.62% | 80.92% | 83.91% | 76.32% | 82.29% | 81.15% | 79.31% | 75.86% | 77.01% | 80.46% | 74.71% | 78.16% | 79.31% |
| Incorrectly classified instance | 90 | 93 | 83 | 70 | 103 | 77 | 82 | 18 | 21 | 20 | 17 | 22 | 19 | 18 |
| 20.69% | 21.38% | 19.08% | 16.1% | 23.68% | 17.7% | 18.85% | 20.69% | 24.14% | 22.99% | 19.54% | 25.29% | 21.84% | 20.69% |

Table 3: Comparison of evaluation metrics using 10fold cross-validation and percentage split (80:20)

We know Random Tree algorithm is a very famous and very well-known algorithm for classification but in this work, it gives the lowest performance accuracy. It has been correctly classified only 76.32% in cross-validation evaluation techniques where Random forest algorithm has been correctly classified 82.29% and J48 has been correctly classified 83.91%. In percentage split evaluation, J48 has been correctly classified 80.46% where Random tree has been correctly classified 74.71%.

1. **Proposed Tool to the End User**

As we wanted to provide a tool for our end user to help them to know the probability of having a stroke. In this modern era, every educated person knows the use of websites and web technology has become a great tool for everyone. So, we chose web technology to give instant help to people all over the world for predicting the probability of having a stroke. With the help of this website, people can predict his/her risk level of stroke from anywhere in the world. This website also gives some helpful suggestions and healthy tips to the end user. Those helpful suggestions and tips can help people who have not stroked yet to remove or decrease the probability of being a stroke. This is a very easy, understandable and simple website. So, the user can use it very easily and quickly. In figure 3, we showed a demo homepage of our proposed tool.

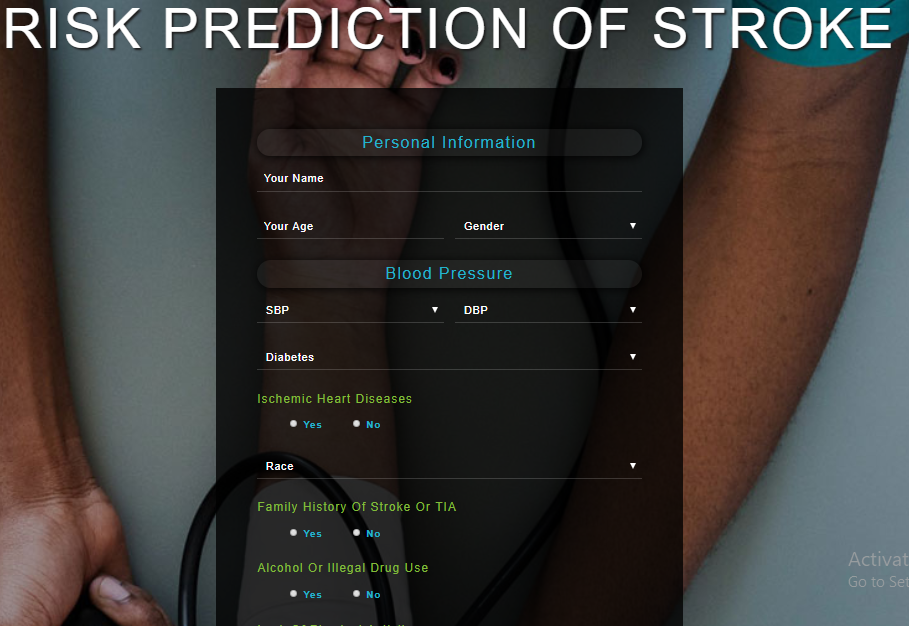


Fig. 2. Proposed tool for the end user

1. **Conclusion**

The number of people who have stroke are increasing day by day. The people are not being conscious as they are going to face this. So, if they can know the probability of having stroke of them then they will be alert and they will try to decrease this probability. This research paper presented a system for risk prediction of stroke by using the different data mining techniques. A new dataset has been created for this research work and the dataset has been collected from different hospitals of Sylhet. As data mining techniques have some algorithms for predicting many diseases so we used some algorithms of data mining techniques namely SVM, NB, RF, J48, MLP etc.(Pragati Agrawal, 2015) We observed in this work that the RF algorithm gave the best accuracy. We also provide a tool for the end users so that they can know the risk level of their having stroke. With the help of this tool we can increase awareness about stroke. However, we have collected only 435 data as our train dataset, it can be updated by increasing the number of instances and can be implemented in others data mining techniques for prediction purpose.

# **8.0 References**

1 Pragati Agrawal, A. K. D., 2015. A Brief Survey on the Techniques using for the Diagnosis of Diabetes-mellitus. *International Research Journal of Engineering and Technology (IRJET),* 02(03), p. 1039.

2 Ray, S., 2017. *Analytics Vidhya.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/  
[Accessed 13 March 2019].

3 Services, U. D. o. H. &. H., 2018. *National heard, Lung, and Blood Institut.* [Online]   
Available at: https://www.nhlbi.nih.gov/health-topics/stroke  
[Accessed 13 March 2019].