RISK PREDICTION OF STROKE USING DATA MINING CLASSIFICATION TECHNIQUES

Masum Mohammad Jubayel[1][0000-1111-2222-3333], AsimaAkter Chowdhury[2][1111-2222-3333-4444]Mohammad Golam Sabbir[3] [2222-3333-4444-5555], Rahatara Ferdousi[4] [4444-5555-6666-7777] and M M Faniqul Islam[5][5555-6666-7777-8888]

[1]Metropolitan University, Sylhet 3100 Bangladesh
[2]Metropolitan University, Sylhet 3100 Bangladesh
[4]Metropolitan University, Sylhet 3100 Bangladesh
[4]Metropolitan University, Sylhet 3100 Bangladesh
[5]Queen Mary University of London, UK

masumjubayel@gmail.com
chowdhuryasima@gmail.com
sabbirsabbu@gmail.com
rahatara@metrouni.edu.bd
m.islam@smd17.qmul.ac.uk

Abstract. Stroke, a non-communicable fatal disease causing death to the mass of Bangladesh each year increasingly. If the risk factors of stroke are under control, the treatment would be beneficial. Thus, predicting the risk of developing stroke can aid the individuals to keep their risk factors under control through maintaining precautions, which would delay the progress of developing the risk of stroke. As Data Mining has multiple classification techniques, which are able to predict risk, analyzing the hidden patterns of patients data, contemporary researches have been carried out to predict disease risks from sign, symptoms or risk factors. In this work firstly, we have prepared a dataset of common risk factors of stroke of 635 patients from a direct questionnaire. Secondly, a detailed analysis of this dataset with Data Mining classification algorithms namely Naïve Bayes (NB), Random Forest (RF), Random Tree (RT) Support Vector Machine (SVM) have been carried out. In addition, the performance of these algorithms has been evaluated through 10 fold Cross-Validation and 80:20 Percentage Split to identify the most accurate one for this dataset. As we found Random Forest Decision Tree as the best algorithm for this dataset, finally we generated the decision tree using Jupyter a Python IDE and proposed a knowledge-based system prototype to predict the risk of developing stroke and to provide essential suggestions to the mass.

Keywords. Stroke, Data mining, risk factor, risk prediction, Naïve Bayes, SVM, Random Forest, Random Tree, Decision Tree.

1 Introduction

A stroke is a sudden interruption in the blood supply of the brain. The effects of a stroke depend on which part of the brain is injured, and how severely it is injured. Strokes may cause sudden weakness, loss of sensation, or difficulty with speaking, seeing, or walking [5]. If the prediction of risk factors is possible, it might be possible to lower risk factors and prevent or delay a stroke [6]. Almost 60% of patients are unaware of their mild stroke, 25% seek medical attention lately and another 15% suffer from recurrent stroke, before seeking attention [3]. Moreover, in a third world developing country like Bangladesh, the treatment procedures after stroke result in an economic burden.

In the healthcare industry, Data Mining plays an important role in predicting diseases or chance of developing diseases [1]. Data mining techniques have great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis [2].

Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Random Tree (RT) Support Vector Machine (SVM), etc. supervised classification DM algorithms have been applied to predict disease or its risk mostly. However, the performance varies due to the type of dataset as well as a pattern of information. As the focus of disease or risk of disease prediction systems is the mass of a country, an algorithm with better accuracy should be identified before developing and deploying a system to the mass [4]

The rest of the paper is organized as, in section 2 we have provided literature studies, in section 3 we have provided our proposed system architecture, in section 4 we have provided our dataset description, in section 5 we have demonstrated our experimental analysis with result, and finally in section 6 we have concluded this paper.

2 Literature review

In [6] the authors used two different data mining techniques named Naïve Bayes and J48 decision tree for the prediction of various diseases such as heart disease, diabetes, and breast cancer and compared their performance for evaluating the best classifier. They found the highest accuracy for diabetes using J48. However, they didn't provide the details of their dataset. They used WEKA tool but they did not show the results which were obtained by the percentage split and cross-validation separately.

In [7] this paper V. Kirubha and S. Manju Priya analyzed the application of the most popular data mining techniques in the medical domain and they used some of the algorithms for disease prediction. They showed that by using various tools and techniques on different disease diagnosis a variety of results can be gained. However, they did not show the result which was obtained by using different data mining techniques.

In [8] the authors presented a study about the model of logistic regression and obtained the result with "XLSTAT" software. They showed several steps of the logistic regression model in their study. They have found the sensitivity rate was 77.58%, the specificity rate was 83.03% and the error rate was 19.7%.

In [9] this paper the authors presented the prediction of risk of stroke within 10 years. And their dataset was within 1500000 men and 1200000 women had used data from the national health examination of the entire nation and they classified the total population in five ranges such as normal, slightly high, high, risky and very risky.

In [10] this paper author presented a study to find the possible risk of stroke by subjecting the risk factors to SVM. They used 100 patient's data with 8 attributes. They used SVM model parameters through its kernel function named polynomial kernel and Gaussian (RBF) kernel. They evaluated the result through the Confusion matrix and showed that the rate of the correctness of prediction by RBF was 98% whereby polynomial was 92%. So, the author told in this paper that the application of SVM models can be used for the processing of stroke-related risk factor data.

3 System Architecture

Our system architecture has been delineated in Fig. 1. Initially, an original dataset including the risk factors of 635 people has been used for selecting the best prediction algorithm. After cleaning data, risk factors of 606 peoples has been used for the experiment as most of the people don't have any idea of their Cholesterol level and all of them are in the same races. Then the processed dataset was feed to the database (which will be used as a trained dataset for the end-user tool) and to the classification algorithms for simulation. The performance accuracy has been evaluated using 10-Fold Cross-Validation and Percentage Split techniques. Finally, according to the best accuracy, the best algorithm has been chosen for enabling the risk prediction feature of the tool.

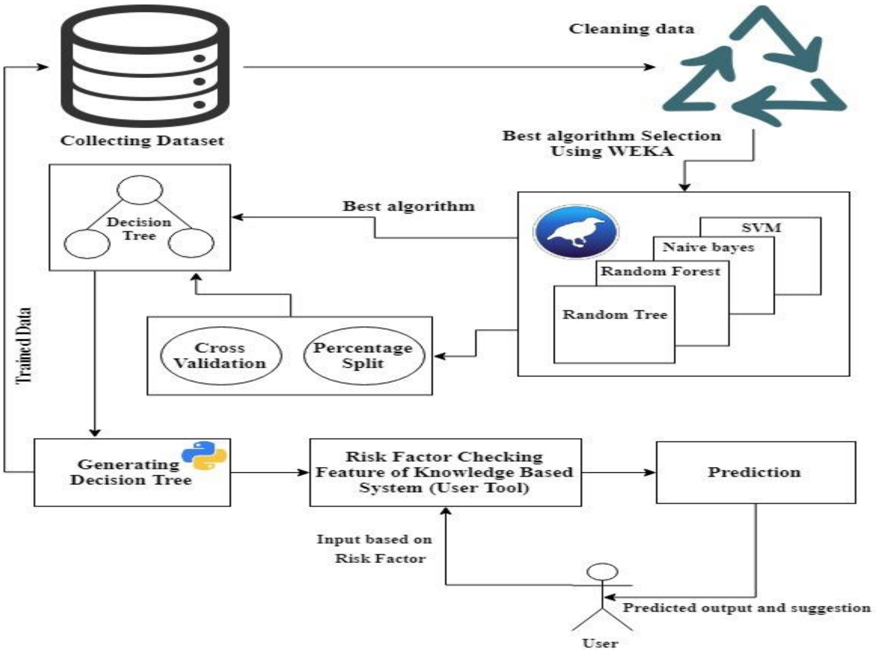


Fig. 1. Proposed System Architecture

4 Dataset Description

In this section, we represented the details of our dataset and attribute description. This dataset contains the information of 606 persons. This dataset has been created from a direct questionnaire to people who have recently developed stroke, or who are still not developed the stroke but having few or more risk factors of stroke. The data has been collected from the patients using direct questionnaires from the different hospitals in Sylhet, Bangladesh. We have collected the information from Sylhet Woman Medical College & Hospital and Jalalabad Ragib Rabeya Medical College & hospital. The description of the dataset is given below.

Table 1. Description of the training dataset

	Number of Attributes	**Number of Instances**
Report based dataset	15	606

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
Systolic BP	1.120>, 2.120-139, 3. 140-160, 4.160<
Diastolic BP	1.80>, 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
Alcoholism	1. No, 2. Yes
Less Physically Active	1. No, 2. Yes
Smoking	1. No, 2. Yes
Stress and depression	1. No, 2. Yes
Saturated Fat↑ ()	1. No, 2. Yes
Fibre↓ ()	1. No, 2. Yes
Chronic Kidney Disease (CKD)	1. No, 2. Yes
Class Attribute	1. Stroke, 2. Non-stroke

The data pre-processing has been conducted by handling the missing values following the technique of ignoring the tuples with incomplete values. After pre-processing, 606 instances have remained in total. Among them, 451 are positive values and 155 are negative values. The detail description of the attributes is shown in Fig 2. Two class variables are used to find whether the patient is having a risk of developing stroke (positive) or not (negative).

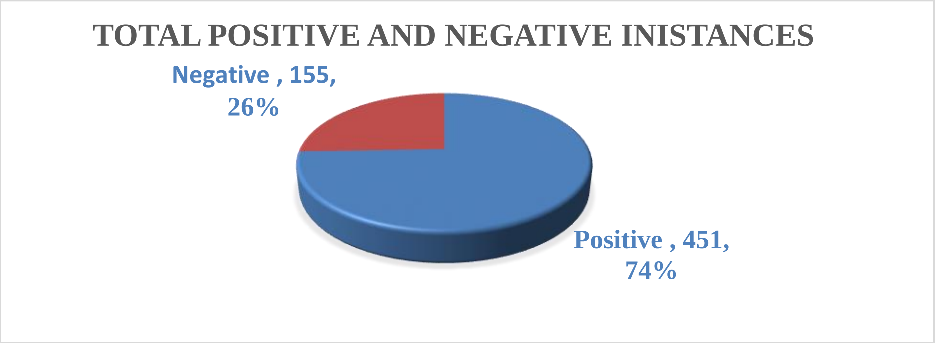


Fig2. Class Attributes Distribution

5 Result Analysis

Performance of different Data Mining techniques on our dataset with detailed accuracy information is represented in the following tables. Although Support Vector Machine is one of the most popular algorithms for data prediction. In the case of our dataset, the accuracy was the lowest for both of the Cross-validation methods and also for the Percentage Split. However, the best result was achieved by using the Random Forest decision tree. Where using 10-fold cross-validation, 84.16% instances were classified correctly and using percentage split technique, it could classify 80.99% of the instances correctly. In table 3 to table 18 we have depicted the detailed analysis of the result. We have found the correctly classified instances and incorrectly classified instances for each algorithm.

Table 3. Performance Results from RandomForest decision tree using (Cross Validation)

	Number of Instances	Percentage
Correctly classified Instances	510	84.16%
Incorrectly Classified Instances	96	15.84%

Table 4. Detailed Accuracy by class from RandomForest decision tree using 10-fold Cross Validation Technique

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.665	0.098	0.701	0.665	0.682
	0.902	0.335	0.887	0.902	0.895
Weighted Average	0.842	0.275	0.839	0.842	0.840

Table 5. Performance Results from RandomForest decision tree using Percentage Split

	Number of Instances	Percentage
Correctly classified Instances	98	80.99%
Incorrectly Classified Instances	23	19.0083%

Table 6. Detailed Accuracy by class from RandomForest decision tree using - Percentage Split (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.585	0.075	0.800	0.585	0.676
	0.925	0.415	0.813	0.925	0.865
Weighted Average	0.810	0.300	0.809	0.810	0.801

Table 7. Performance Results from Naïve Bayes Algorithm using (Cross Validation)

	Number of Instances	Percentage
Correctly classified Instances	449	74.09 %
Incorrectly Classified Instances	157	25.91%

Table 8 Detailed Accuracy from class Naïve Bayes with 10-fold Cross Validation technique

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.303	0.109	0.490	0.303	0.375
	0.891	0.697	0.788	0.891	0.837
Weighted Average	0.741	0.546	0.712	0.741	0.718

Table 9. Performance Results from Naïve Bayes - Percentage Split

	Number of Instances	Percentage
Correctly classified Instances	82	67.77%
Incorrectly Classified Instances	39	32.23%

Table 10. Detailed by class Accuracy from Naïve Bayes– Percentage Split (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.195	0.075	0.571	0.195	0.291
	0.925	0.805	0.692	0.925	0.791
Weighted Average	0.678	0.558	0.651	0.678	0.622

Table 11. Performance Results from RandomTree, a Decision Tree Algorithm using (Cross Validation)

	Number of Instances	Percentage
Correctly classified Instances	489	80.69%
Incorrectly Classified Instances	117	19.31%

Table 12. Detailed Accuracy from RandomTree using 10-fold Cross Validation technique

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.677	0.149	0.610	0.677	0.642
	0.851	0.323	0.885	0.851	0.868
Weighted Average	0.807	0.278	0.815	0.807	0.810

Table 13. Performance Results from RandomTree using - Percentage Split

	Number of Instances	Percentage
Correctly classified Instances	95	78.51%
Incorrectly Classified Instances	26	21.49%

Table 14. Detailed Accuracy from RandomTree using - Percentage Split (80:20)

	TP Rate	FP Rate	Precision	Recall	F-measure
	0.659	0.150	0.692	0.659	0.675
	0.850	0.341	0.829	0.850	0.840
Weighted Average	0.785	0.277	0.783	0.785	0.784

Table 15. Performance Results from Support Vector Machine Algorithm using (Cross Validation)

	Number of Instances	Percentage
Correctly classified Instances	451	74.42%
Incorrectly Classified Instances	155	25.56%

Table 16. Detailed Accuracy from Support Vector Machine using - Cross Validation

	TP Rate	FP Rate	Precision	Recall	F-measure
	0	0	0.582	0	0.623
	1	1	0.786	1	0.853
Weighted Average	0.744	0.744	0.029	0.744	0.021

Table 17: Performance Results from Support Vector Machine Algorithm – Percentage Split

	Number of Instances	**Percentage**
Correctly classified Instances	80	66.12 %
Incorrectly Classified Instances	41	33.88%

Table 18. Detailed Accuracy from Support Vector Machine using 10 Percentage Split (80:20)

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-measure**
	0	0	0.366	0	0.223
	1	1	0.661	1	0.796
Weighted Average	0.661	0.661	0.026	0.661	0.013

For the more semantic view of the performance of used algorithms using both evaluation techniques are depicted in graphs. In Fig. 3, the performance of the algorithms using Cross-validation evaluation is depicted and in Fig. 4, the results from percentage split have been shown to represent the comparative accuracy of the used algorithms.

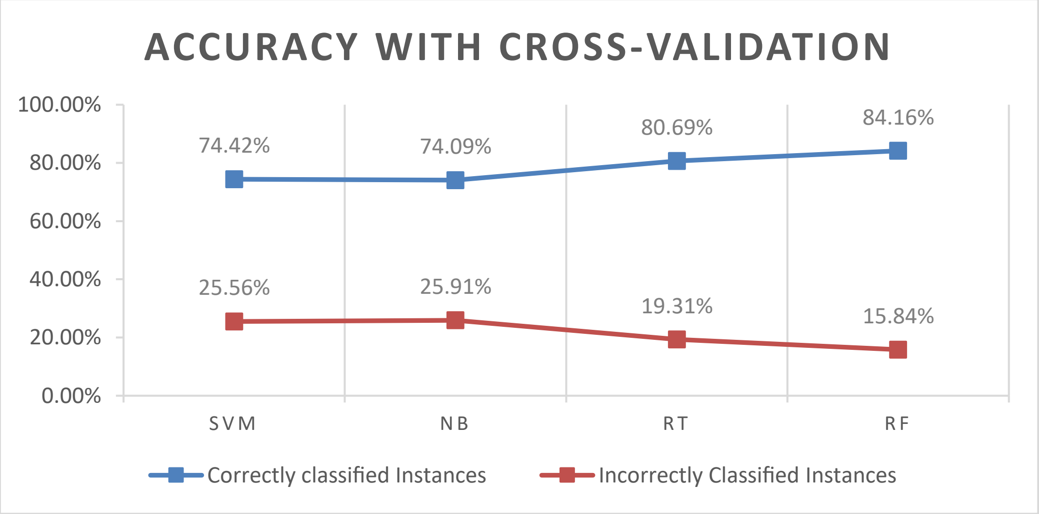


Fig 3. Performance of Classification Algorithms Using Cross-Validation Technique

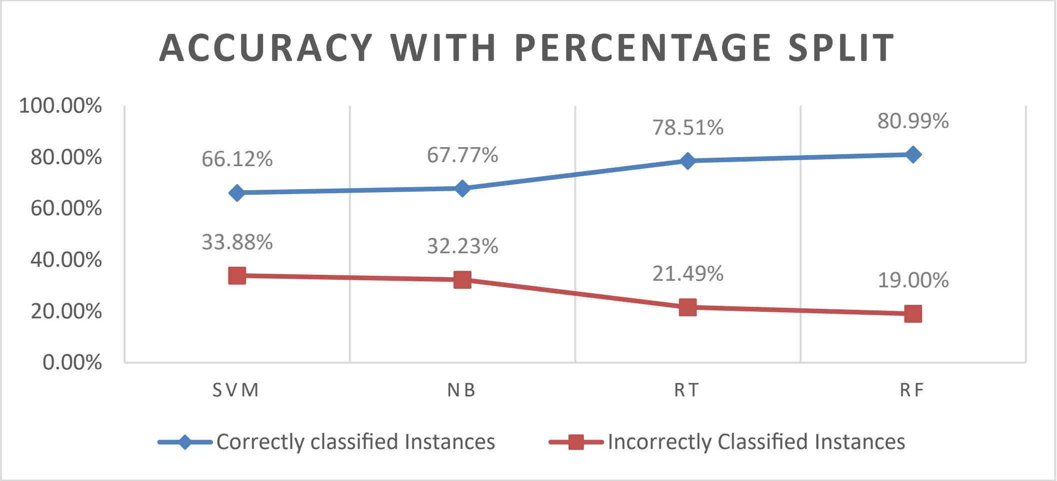
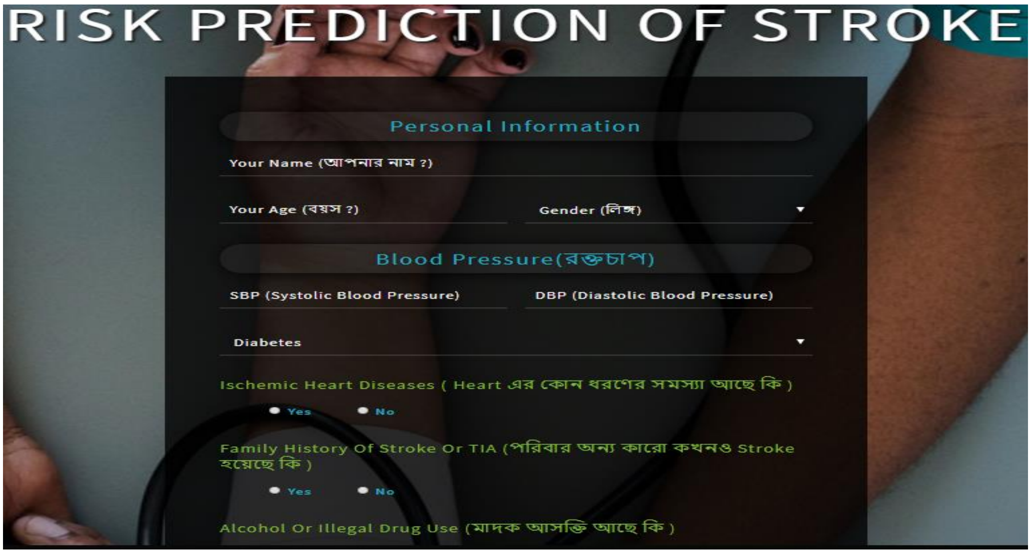


Fig. 4. Performance of Classification Algorithms Using Percentage Split Technique

From table 3 to 18 we can see that the decision tree gives the best accuracy performance according to our dataset. After seeing this we have analyzed our dataset by implementing the code of the decision tree in python. We implemented the code in python in two ways. First, we implemented the code with 10-fold cross-validation. For this, we split the dataset into 10-fold and calculate the percentage of the performance accuracy. Then we calculate the Gini Index of the attributes and attributes values for the dataset and create a terminal node value. Then we generate a decision tree by following the CART (Classification and Regression Tree) algorithm and find 80.33% the mean accuracy of the algorithm..

6 Proposed Tool for the End Users

We propose a web application with a user-friendly UI, where a knowledge-based web-site asking questions in both Bengali and English could be beneficial to check the risk of developing stroke by using user's common risk factors as input. However, any other region adopting this idea can change the language according to theirs. This concept was made due to reach mass people of every stage with the contribution of this research work. This website will also provide some useful suggestions and tips to the end-users to avoid developing stroke and seek medical attention timely. In figure 5 and figure 6 a demo input page is given below. In figure 7 and figure 8 a demo output page is depicted, which shows the generated prediction of having a stroke resulting in whether you are at risk or not.



RISK PREDICTION OF STROKE

Personal Information

Your Name (আপনার নাম ?)

Your Age (বয়স ?)

Gender (লিঙ্গ)

Blood Pressure(রক্তচাপ)

SBP (Systolic Blood Pressure)

DBP (Diastolic Blood Pressure)

Diabetes

Ischemic Heart Diseases (Heart এর কোন ধরণের সমস্যা আছে কি)

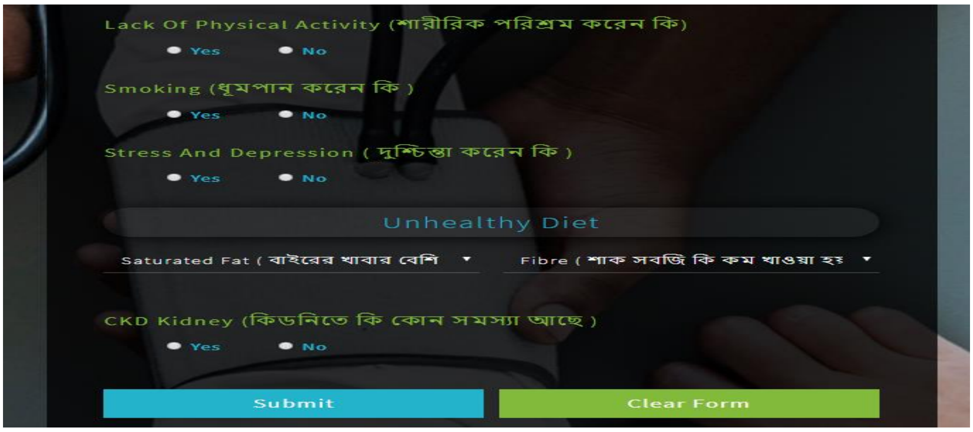
● Yes ● No

Family History Of Stroke Or TIA (পরিবার অন্য কারো কখনও Stroke হয়েছে কি)

● Yes ● No

Alcohol Or Illegal Drug Use (মাদক আসক্তি আছে কি)

Fig 5. Homepage of the Proposed Website/System



Lack Of Physical Activity (শারীরিক পরিশ্রম করেন কি)

● Yes ● No

Smoking (ধূমপান করেন কি)

● Yes ● No

Stress And Depression (দুশ্চিন্তা করেন কি)

● Yes ● No

Unhealthy Diet

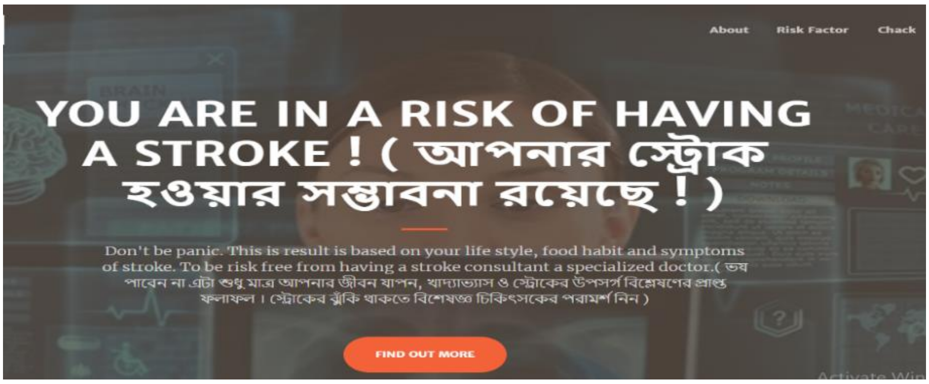
Saturated Fat (বাইরের খাবার বেশি

Fibre (শাক সবজি কি কম খাওয়া হয়

CKD Kidney (কিডনিতে কি কোন সমস্যা আছে)

● Yes ● No

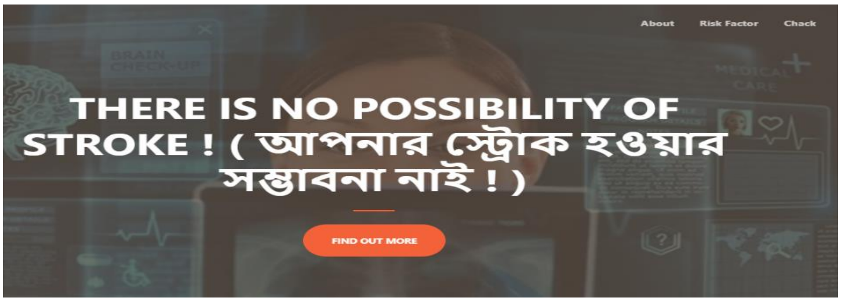
Fig 6. Homepage of the Proposed Website/System



YOU ARE IN A RISK OF HAVING A STROKE ! (আপনার স্ট্রোক হওয়ার সম্ভাবনা রয়েছে !)

Don't be panic. This is result is based on your life style, food habit and symptoms of stroke. To be risk free from having a stroke consultant a specialized doctor.(ভয় পাবেন না এটা শুধু মাত্র আপনার জীবন যাপন, খাদ্যাভ্যাস ও স্ট্রোকের উপসর্গ বিশ্লেষণের প্রাপ্ত ফলাফল । স্ট্রোকের ঝুঁকি থাকতে বিশেষজ্ঞ চিকিৎসকের পরামর্শ নিন)

Fig 7. Risk Checking Page of Stroke



THERE IS NO POSSIBILITY OF STROKE ! (আপনার স্ট্রোক হওয়ার সম্ভাবনা নাই !)

Fig 8. Risk Checking Page of Stroke

7 Conclusion

Most of the statistics are showing that the global prevalence of stroke is rising. Where people are still unaware of the risk factors of developing a stroke. Knowing the risk factors by any means could make them alert to reduce the incidence of stroke and its aftermaths effectively. In this work, we have conducted a detailed analysis using multiple Data Mining Techniques and found Random forest Decision Tree as the best one. We observed that the Random Forest Decision Tree algorithm provided the best accuracy at 84.16%. We also proposed a prototype for the end-user application. However, analysis with more data and deployed mobile apps could be an integrated research scope of this work.

References

1. *AICN*. (2014, April 30). Retrieved from AICN: http://www.ieeeottawa.ca/aicn/data-mining-and-knowledge-discovery-in-healthcare-and-medicine/Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
2. Kamaraj, K. (2016). Multi Disease Prediction using Data Mining Techniques. *Research Gatee*, 1-2.
3. Islam MN1, M. M. (2013, 4 8). *Burden of stroke in Bangladesh*. Retrieved from National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/pubmed/22974096
4. Jiawei Han, M. K. (2001). *Data Mining Concepts and Techniques.* Academic Press, Morgan Kaufmann Publishers.
5. Siri Krishan Wasan, V. B. (2006). The Impact of Data Mining Techniques On Medical Diagnostics. *Data Science Journal*, 119-126.
6. D. D. S. P. K.GomathiKamaraj, "ResearchGate," 17 September 2017. [Online]. Available: https://www.researchgate.net/publication/319851535_Multi_Disease_Prediction_using_Data_Mining_Techniques. [Accessed 03 July 2019].
7. S. M. P. V. Kirubha, "Survey on Data Mining Algorithms in Disease Prediction," International Journal of Computer Trends and Technology (IJCTT) , vol. 38, pp. 124-128, 2016.

8. N. A. Balar Khalid, "A Model for Predicting Ishcemic Stroke Using Data Mining Algorithms," IJISET- International Journal of Innovative Science, Engineering & Technology, vol. 2, no. 11, pp. 18-23, 2015.
9. H.-s. L. Jae-woo-Lee, "The development and implementation of stroke risk prediciton model in National Health insurance service's personal health record.".
10. K. R. S. Sabibullah Mohamed Hanifa, "Stroke Risk Prediction Through Non-linear Support Vector Classification Models," Internation Journal of Advanced Research in Computer Science, vol. 1, pp. 47-53, 2010.