Early Brain Stroke Prediction Using Machine Learning Technique

BY
Toufika Sharmin
151-15-5285

&

Md. Shibli Sadik
151-15-4704

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Md. Azizul Hakim
Lecturer
Department of CSE,
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

November 2018

APPROVAL

This Project/internship titled **"Early brain stroke prediction using machine learning technique"**, submitted by **Toufika Sharmin, ID No: 151-15-5285 & Md. Shibli Sadik**, **ID No: 151-15-4704** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 09-12-18.

BOARD OF EXAMINERS

Dr. Syed Akhter Hossain **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Narayan Ranjan Chakraborty **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Tarek Habib **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Mohammad Shorif Uddin **External Examiner**
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this research has been done by us under the supervision of **Md. Zahid Hasan** Assistant Professor and co-supervision of **Md. Azizul Hakim** Lecturer, Department of CSE, Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for the award of any degree.

SUPERVISED BY:

Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

CO-SUPERVISED BY:

Md. Azizul Hakim
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Toufika Sharmin
ID: -151-15-5285
Department of CSE
Daffodil International University

Md Shibli Sadik
ID: -151-15-4704
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

At first, we are thankful to Almighty Allah for his mercy and grace without which we wouldn’t be able to complete our project. We had to work hard to get the job done but we are grateful to some other people, without the help of whom this project couldn’t be as it is. We tried our best and finally this research-based project is completed.

At first, we’d like to thank our respected **supervisor**, **Md. Zahid Hasan, Assistant Professor and co-supervisor, Md. Azizul Hakim,** Department of Computer Science & Engineering, Daffodil International University. This whole time they have supported us, inspired us and showed us the right way. They made it easier for us to work continuously with all their patience and inspirations. Every time we had a difficulty, we contacted them and they helped us with a suggestion and that helped us a lot throughout the whole time.

They were so friendly with us this whole time and that was the main inspiration for us. We are so lucky to work under their supervision and of course, it has been an honor to work under their supervision. We also want to our express our deepest gratitude to honorable Professor and Head of CSE department, **Prof. Dr. Syed Akhter Hossain.**

At last, again we want to thank all the good wishers, friends, family, seniors for all the help and inspirations. This research is a result of hard work and all those inspirations and assistance.

ABSTRACT

Sometimes, a stroke called a "brain attack" occurs when blood supply to an area in the brain is cut off. Often due to doctor perception, deciding whether an observation diagnostic test is brain stroke or not which causes controversy. As a stroke can happen many problems in our brain which causes brain stroke. One of the most common reason of brain stroke is blood supply occurrences. But the examination of this kind of brain stroke, doctor observes the result of this test then decide a result of brain stroke. But human can’t show a result early and accurate result all the time. In this matter doctor make their decision on diagnostic test result. But human perception cannot be accurate all the time. Besides, it is not always possible to conclude the accurate judgement early. In that case, the benefit of early prediction, we can get accurate result. In this project, we define our classification task as the prediction of brain stroke. This system gives 97% accuracy of brain stroke. Our approach eliminates the time consuming of human perception.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

In restorative science mind stroke is one of the real infections everywhere throughout in the world. When blood supply is hindered to cerebrum then a cerebrum stroke occurs. The stroke influences the vast majority of the parts of cerebrum region. Chiefly stroke can occur in two different ways (a)Ischemic (80% everything being equal) (b) hemorrhagic. Transient ischemic assault (TIA) additionally causes mind stroke which is incidentally hindered in the cerebrum blood supply and causes a smaller than normal stroke [1]. At the point when the supply routes to mind wind up blocked which is causing intense diminished blood stream then ischemic strokes happen. The most ischemic strokes have thrombotic stroke and embolic stroke. When a vein in cerebrum releases then hemorrhagic stroke occurs. These incorporate uncontrolled hypertensions, overtreatment with anticoagulants, frail spots in your vein dividers [2].

In provincial Bangladesh the greater part of the general population are experiencing mind stroke. The World Health Organization (WHO) positions as number 84 mortality because of stroke in Bangladesh on the planet [3]. The stroke is the third driving reason for death. The dominant part of cases hypertension (63 percent) happen in people beyond forty 24 years old, by coronary illness (24 percent), and diabetes (21 percent) and hyperlipidemia (7percent). The entire level of hazard factors was greater than 100% because of the way that a portion of the patients had different hazard factors [4].

The vast majority of individuals dependably take high utilization of sugar, salt, cigarette, tobacco, liquor; physical idleness, stoutness, and hereditary components are in charge of hypertension, or, in other words strokes [5]. Numerous individuals even don't think about the stroke indications which learning can safe a real existence from stroke. An absence of data and poor control and furthermore administration of hazard factors have added to the developing rates of stroke.

In cerebrum stroke, it is trying to grow such innovation that can ready to choose any patient has mind stroke or not without testing his body. And furthermore, tedious to get a legitimate after effect of this patient conditions. In this circumstance we utilized a machine learning method in our framework which is KNN (K Nearest Neighbors). K implies calculation mostly use for a settled number of groups where k is the quantity of bunches. Our framework utilizes likewise a settled an incentive for preparing and test set to characterize that mind stroke exists or not.

1.2 Motivation

In case of brain stroke, when brain strokes occur, sometimes can't understand which stroke is occurred. Is this brain stroke or other diseases? For this reasons people have to go to hospital and do many testing's for this disease which is very time consuming and wasting money. In this situation patient cannot get proper treatment in early which is bad for this patient's condition. Many of the time doctor almost done getting result but sometimes result become wrong for doing early. That's why many patient's do not get right treatment and can't understand which will be the next treatment. So many people suffer more and more. Our system helps to give a good result for this situation and get early result that brain stroke have or not.

1.3 The rationale of the study

Presently a-days, brain stroke is considered as one of the most serious diseases. Numerous people groups are being kicked the bucket because of brain stroke. It considered as a most alluring sickness for getting to be identified with the heart. In the wake of seeking and breaking down we picked brain stroke as our exploration point. For turning into an expansive number of dead on the brain stroke, the exploration subject has been chosen. At long last, the paper has been chipping away at this to give a superior proposal that encourages us to diminish the dead number for our advanced age people groups.

1.4 Research questions

a) Does it show the accurate value to predict brain stroke in early prediction?
b) Does it classify brain stroke diseases by machine learning algorithm?

Already, several harmful diseases have been detected for a human being. Although each disease has a solution for prevention it's not possible for everyone due to only for unconsciousness. Everyone wants to lead a happy life in where a disease is the only obstacle. Any kind of disease prevention is possible if that in remain primary stage. For that reason, we built a prediction system that helps to identify the disease stage and provides us the result that he or she has brain stroke or not. All of the diseases, brain stroke disease is considered one of the leading diseases. Many peoples are died due to this disease. Brain stroke disease is the biggest killer of both men and women in the world. In our Bangladesh there are no well-known system for brain stroke. Finally, we selected it as our research topic in Bangladesh people for our pleasure and also try to make a good system for prediction brain stroke diseases.

1.5 Expected Output

In our brain stroke system is a system that helps to generate an expected result based on the given dataset. In this system, we used 70% of the training to get more accurate predictions. How accurate is it, it depends entirely on the training dataset? After completing all the needed procedure of the proposed system, our system has been ready for preparing out on the given dataset. We have applied various strategies to achieve our desired results. We got 97% accuracy from the K-Nearest Neighbors (KNN) among all that we have used.

1.6 Layout of the Report

- Chapter 1 have demonstrated an introduction to the project with its motivation, research questions and expected outcome.
- Chapter 2 will have "Background" demonstrates introduction, related works, research summary and challenges.
- Chapter 3 will have Research Methodology.
- Chapter 4 will have Experimental Results and Discussion.
- Chapter 5 will have Summary and Conclusion.

CHAPTER 2

Background Study

2.1 Introduction

In this section, we will discuss related works, research summary and challenges about this research. In related works section, we will discuss other research paper and their works, their methods and accuracy which are related to our work. In research summary section we will give the summary of our related works. In challenges section, we will discuss how we increased the accuracy level.

2.2 Related Work

Patrick Luckett, Elena Pavelescu, Todd McDonald, Lee Hively, Juan Ochoa proposed a technique complex nonlinear dynamical framework for foreseeing state transitions in cerebrum elements through phantom difference of stage space graph and get high sensitivity (90–100%) [6]. Jeroen de Bresser , Marileen P. Portegies , Alexander Leemans , Geert Jan Biessels ,L. Jaap Kappelle , Max A. Viergever proposed US and KNN method to get a good precision, accuracy and comparability for cerebrum volume estimations [7]. Harold P. Adams Jr., Birgitte H. Bendixen , L. Jaap Kappelle, Jose Biller, Betsy B. Love, David Lee Gordon, E. Eugene Marsh, and the TOAST Investigators discussed a TOAST classification to get future clinical preliminaries that enroll patients with intense ischemic stroke [8]. Mariana Bento, Yan Sym, Richard Frayne, Roberto Lotufo, and Let´ıcia Rittner proposed automated WML division technique requires no from the earlier data and plays out a surface based classification of pixels inside the mind white issue to show signs of improvement results[9]. R. Guerreroa, C. Qina, O. Oktaya, C. Bowlesa, L. Chena, R. Joulesb, R. Wolzb,a, M.C. Valdés-Hernándezc, D.A. Dickiec, J. Wardlawc, D. Rueckerta proposed segmentation of WMH to find with the expert-annotated volumes[10]. Md. Nazmul Islam, Mohammed Moniruzzaman, Md. Ibrahim Khalil1, Rehana Basri, Mohammad Khursheed Alam, Keat Wei Loo, and Siew Hua Gan discussed risk factor of brain stroke [11]. Rita V. Krishnamurthi, Suzanne Barker-Collo, Varsha Parag, Priya kumari Parmar, Emma Witt, Amy Jones, Susan Mahon, Craig S.

Anderson, P. Alan Barber, Valery L. Feigin, MD proposed TOAST method to discover the recurrence of all hazard factors expanded in IS [12].

2.3 Research Summary

In this table 2.1, we are shown some short description of some research paper which is related our topic.

Table 2.1: Research paper summary

SL No	Author	Methodology	Description	Outcome
1	Patrick Luckett, Elena Pavelescu, Todd McDonald, Lee Hively, Juan Ochoa [6].	complex nonlinear dynamical system	state transitions in cerebrum elements through phantom difference of stage space graph	Sensitivity (90–100%)
2	Jeroen de Bresser, Marileen P. Portegies, Alexander Leemans, Geert Jan Biessels ,L. Jaap Kappelle , Max A. Viergever[7].	US and KNN method	cerebrum volume estimations	A good precision, accuracy and comparability
3	Harold P. Adams Jr., Birgitte H. Bendixen, L. Jaap Kappelle, Jose Biller, Betsy B. Love, David Lee Gordon, E. Eugene Marsh III, and the TOAST Investigators [8].	TOAST classification	future clinical preliminaries that enroll patients with intense ischemic stroke	get future clinical trials

4	Mariana Bento, Yan Sym, Richard Frayne,Roberto Lotufo, and Let´ıcia Rittner [9].	automated WML segmentation	technique requires no from the earlier data and plays out a surface based classification of pixels inside the mind white issue	demonstrates better results
5	R. Guerreroa, C. Qina, O. Oktaya, C. Bowlesa, L. Chena, R. Joulesb, R. Wolzb,a, M.C. Valdés-Hernándezc, D.A. Dickiec, J. Wardlawc, D. Rueckerta[10].	segmentation of WMH	segmentation of WMH to find with the expert-annotated volumes	find with the expert-annotated volumes
6	Md. Nazmul Islam, Mohammed Moniruzzaman, Md. Ibrahim Khalil, Rehana Basri, Mohammad Khursheed Alam, Keat Wei Loo, and Siew Hua Gan [11].	Analysis	risk factor of brain stroke	Highly details of stroke in Bangladesh
7	Rita V. Krishnamurthi, Suzanne Barker-Collo, Varsha Parag, MSc; Priyakumari Parmar, Emma Witt, MSc; Amy Jones, Susan Mahon,Craig S. Anderson, P. Alan Barber, Valery L. Feigin.[12].	TOAST	discover the recurrence of all hazard factors expanded in IS	Better result

Brain stroke is a disease that attacks the brain. Without a doubt, the brain is a very important part of every human being. Therefore, if we want to lead a healthy life, we have to be cautious. If we find this disease early in the initial stage, we can easily overcome it. Otherwise we must suffer for this disease for our future life. After making the decision based on the current situation, we wanted to establish a system that provides better performance due to disease and understand the situation of affected patients. Finally, we touch our expected goal for God's blessing, which we have thought to implement.

2.4 Challenges

Data collection is one of the big challenges for getting predicting accuracy. Without data, the prediction is not possible and it can't predict. After that, another challenge is preprocessing. After doing preprocessing our data set has no null value and helps us to get a good prediction. Next, Feature scaling helps to take all feature values into the same scale with respect to value. Therefore, different algorithm has been applied to the proposed architecture. Finally, the implementation process has been established to get accurate predicted value. There were several challenges rising according to the working procedure. We are tried to increase and get a better result for this model by using machine learning algorithm of K nearest neighbors.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In our system, we try to collect data from different medical hospital in our country (Bangladesh). We also try to unique and make accurate prediction in our research. We use 12 features and 200 data from Bangladesh medical hospital's patient dataset. After that, we got some missing value which we resolved. To get the proper prediction, we've already completed the feature scaling process. Datasets are used for training and testing purposes and here are some of them included algorithm Logistic Regression (LR), Classification and Regression Trees (CART), k-Nearest Neighbor (KNN), support vector machines (SVM). The appropriate algorithm scenario has been given based on the working procedure.

3.2 Data Collection Procedure

However, a few hospital-based stroke studies have been carried out involving five hospitals: Dhaka Medical College and Hospital(DMCH),MMCH(Mymensingh Medical College Hospital),Green Life Medical College Hospital (GLMCH),Shaheed Taj Uddin Ahmad Medical College-Gazipur and Khulna Medical College Hospital(KMCH).The study was conducted about 100 brain stroke patients and 100 normal patient data (have no stroke) in Mymensingh Medical College Hospital(MMCH),Green Life Medical College Hospital (GLMCH),Dhaka Medical College Hospital(DMCH), Khulna Medical College Hospital(KMCH) between 2017 to 2018.We also try our work can be unique research. We also search many research papers and get many research papers in brain stroke. But our work is for classification of brain stroke in early prediction which is unique. In Bangladesh, nobody can't research for classification of brain stroke. That's why we collect brain stroke dataset and try to make a good system for our people. Here are some details of dataset collection table 3.1.

Table 3.1: Dataset details

SL No	Name Of the medical	No. of data
1	Dhaka Medical College and Hospital	28
2	Mymensingh Medical College Hospital	90
3	Green Life Medical College Hospital	55
4	Shaheed Tajuddin Ahmad Medical College-Gazipur	15
5	Khulna Medical College Hospital	12

In this investigation of doctor's facility present that the quantity of stroke patients is 80% are male and rest of female, with 70% experiencing ischemic stroke (IS) and 20% experiencing hemorrhagic stroke (HS). A little level of the stroke patients had past history of stroke (7%) or transient ischemic assault (TIA) (3%). As per the statistic information, 54% and 46% lived in urban and country regions, individually, while 47% and 53% were from low-and center salary gatherings, separately. Here are the brain stroke dataset type in the figure 3.1 .

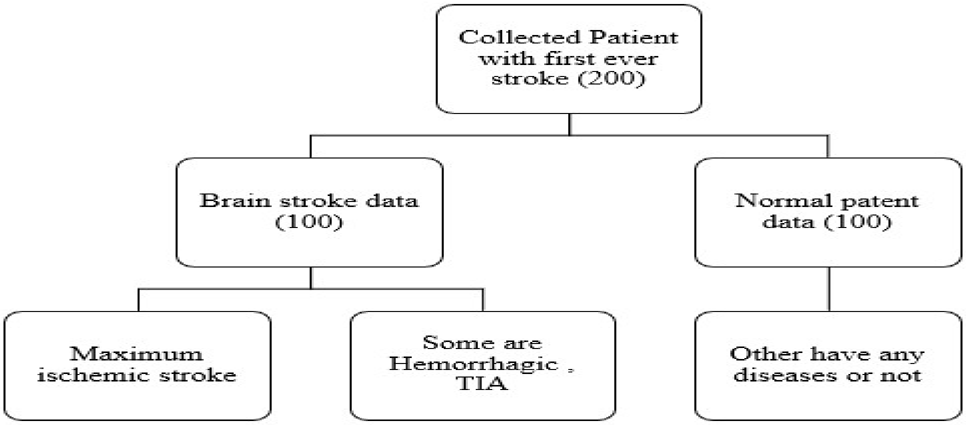
Collected Patient with first ever stroke (200)
Brain stroke data (100)
Normal patent data (100)
Maximum ischemic stroke
Some are Hemorrhagic , TIA
Other have any diseases or not

Figure 3.1: Brain Stroke dataset type

In this figure 3.1, we discuss that our dataset type are brain stroke and non-brain stroke patient data. In brain stroke data most of the patient has ischemic stroke and some are hemorrhagic and TIA ((Transient Ischemic Attack).

In self-reported dataset conduct that RBS, LDL, HDL, Triglyceride are main factor of brain stroke. In this dataset RBS are less than 7.8mmol/L, HDL (High-density lipoprotein) levels and LDL (Low-density lipoprotein) levels, Triglyceride which recommend that patient have brain stroke. In the study ischemic stroke is main stroke which is increased day by day. Dataset of features reference value are shown in table 3.2.

Table 3.2: Dataset of features reference value

Feature	Reference value	Feature	Reference value
RBS	<7.8mmol/L	FBS	3.6-5.8mmol/dl
S.Creatinine	<1.2mg/d	HbA1C	4-5.6%
S.Cholesterol	150-220m	Hb	13-18
LDL	<180	RBC	M:3.8-5.8, f:4.5-6.5
HDL	>=35	AGE	>40
Triglyceride	<150mg/dl	GENDER	Maximum male

In this dataset table 3.2, we try to show 12 dataset features reference value. In table 3.2, we want to show our dataset information where features are Random blood Sugar test (RBS), Serum Creatinine(S.Creatinine), Serum Cholesterol(S.Cholesterol), Low-density lipoprotein (LDL), High-density lipoprotein (HDL), Triglyceride, Fasting blood sugar (FBS), hemoglobin A1c (HbA1C), Hemoglobin(Hb), red blood cell count (RBC), age, gender (sex).

Some ischemic strokes are caused by a narrowing in the carotid artery, which is an artery in the neck, which takes blood to the brain. The narrowing, known as carotid stenosis, is caused by a build-up of fatty plaques. Bleeding can develop inside areas of ischemia, a condition known as "hemorrhagic transformation". It is unknown how many hemorrhagic strokes actually start as ischemic stroke [13].

3.3 Statistical Analysis

In our system dataset, we have 100 datasets for brain stroke patient and 100 datasets non-brain stroke patient. Here, we selected 70% dataset to train and 30% dataset to test. In this system we try to find out accuracy from the dataset and use KNN method to predict brain stroke diseases. In this figure 3.2, we are shown that dataset flowchart and how we use the dataset for proposed model.

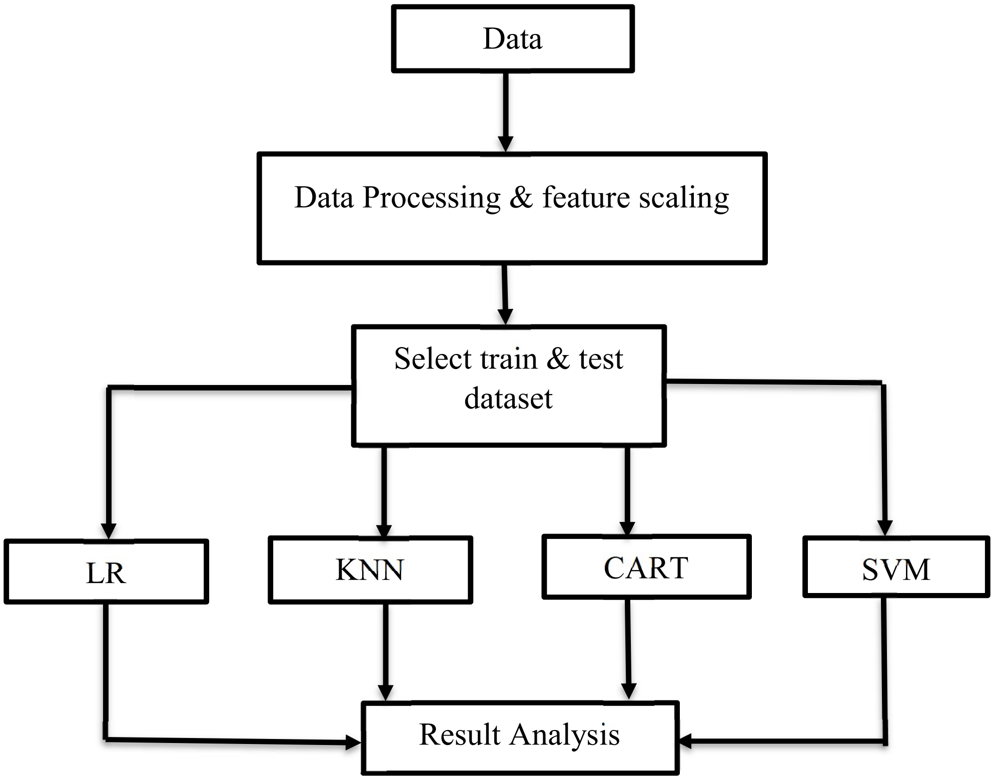
Data
Data Processing & feature scaling
Select train & test dataset
LR
KNN
CART
SVM
Result Analysis

Figure 3.2: Architecture of proposed model.

In this figure, we have shown that how we do our research in shortly details. In this figure 3.2 we can know how to go ahead for our target step by step.

3.4 Research Subject and Instrumentation

As of late, the fame of machine learning calculations is rising exponentially. Machine Learning Algorithms give PCs the capacity to gain from information with the assistance of measurable methodologies. A machine can discover the interior information example and deliver a choice or prescient learning as a result without the assistance of express coding is considered as the most premium part. In this way, a similar calculation can be connected to datasets of various areas without having a change of its interior structures. There are distinctive kinds of machine learning calculations, however we have utilized some of them to our framework. The algorithm details are given below:

3.4.1 Logistic Regression

Logistic Regression, falls under Supervised Machine Learning. It takes care of the issues of Classification (to settle on forecasts or take choices dependent on past information). It is utilized to anticipate two fold results for a given arrangement of free factors. The reliant variable's result is discrete. Logistic relapse is another procedure acquired by machine gaining from the field of statistics. It is the go-to strategy for paired characterization problems. Logistic relapse utilizes a condition as the portrayal, particularly like straight relapse [14].

Info esteems (x) are joined straightly utilizing weights or coefficient esteems (alluded to as the Greek capital letter Beta) to foresee a yield esteem (y). A key contrast from direct relapse is that the yield esteem being demonstrated is a paired quality (0 or 1) instead of a numeric esteem.

The following is a model strategic relapse condition:

$y = e^{(b0 + b1*x)}/(1+e^{(b0 + b1*x)})$ ……………. (1)

From equation 1, here y is the anticipated yield, b0 is the inclination or capture term and b1 is the coefficient for the single information esteem (x). Every segment in your information has a related b coefficient (a consistent genuine esteem) that must be gained from your preparation information [14].

3.4.2 Decision tree (CART)

Decision trees are a vital kind of calculation for prescient demonstrating machine learning. The traditional decision tree calculations have been around for quite a long time and present-day varieties like arbitrary timberland are among the most ground-breaking procedures available. the humble decision tree calculation known by its more current name CART which represents Classification and Regression Trees. Decision tree technique is utilized as the most integral asset for taking in the machine since it gets compelling outcomes as quickly as time permits. Choice tree has diverse kinds of calculations: Cart, ID3, C 4.5, CHH and H48. Among them J48 is utilized and it is exceptionally mainstream algorithm. J48 utilizes pruning technique for building a tree. This calculation keeps on being a recursive procedure until the point that the normal outcomes are found. It gives great exactness and adaptability. This formula is made available from the following equations [15].

$E = \sum_{i=1}^{K} P_i \, log_2 \, P_i$………… (2)

From equation 2,
K defines the number of classes of target attributes,
Pi defines the number of occurrences of class,
i is divided by the total number of instances.
Traditionally, this calculation is alluded to as " decision trees", yet on a few stages like R they are alluded to by the more present-day term CART [15].

3.4.3 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a discriminative classifier formally characterized by an isolating hyperplane. At the end of the day, given marked preparing information (managed taking in), the calculation yields an ideal hyperplane which arranges new precedents. In two-dimensional space this hyperplane is a line partitioning a plane in two sections where in each class lay in either side the numeric info factors (x) in your information (the sections) shape a n-dimensional space. For instance, on the off chance that you had two information factors, this would frame a two-dimensional space [16].

A hyperplane is a line that parts the information variable space. In SVM, a hyperplane is chosen to best separate the focuses in the info variable space by their class, either class 0 or class 1. In two-measurements you can envision this as a line and we should expect that the majority of our info focuses can be totally isolated by this line. For instance:

$B0+(B1*X1)+(B2*X2)= 0$……………. (3)

From equation 3, here the coefficients (B1 and B2) that decide the incline of the line and the capture (B0) are found by the learning calculation, and X1 and X2 are the two info factors [16].

3.4.4 K-nearest neighbor (KNN)

KNN falls in the managed adapting group of calculations. Casually, this implies we are given a marked dataset consisting of preparing perceptions (x, y) and might want to catch the connection among x and y. All the more formally, we will likely take in a capacity h:X→Y with the goal that given a concealed perception x, h(x) can unquestionably foresee the relating yield y. The KNN classifier is likewise a non-parametric and occasion-based learning calculation. Non-parametric means it makes no unequivocal presumptions about the utilitarian type of h, staying away from the risks of mismodeling the basic dispersion of the information. For instance, assume our information is exceptionally non-Gaussian however the learning model we pick accept a Gaussian shape. All things considered, our calculation would make to a great degree poor forecast. Case based learning implies that our calculation doesn't expressly take in a model. Rather, it retains the preparation occurrences which are thusly utilized as "learning" for the forecast stage. Solidly, this implies just when an inquiry to our database is made (i.e. when we request that it foresee a name given an information), will the calculation utilize the preparation occasions to release an answer [17].

In the arrangement setting, the K-closest neighbor calculation basically comes down to shaping a larger part vote between the K most comparative occurrences to a given "concealed" perception. Similitude is characterized by a separation metric between two information focuses. A prominent decision is the Euclidean separation given by

$d(x,x')=\sqrt{(x_1-x'_1)^2+(x_2-x'_2)^2+\ldots+(x_n-x'_n)^2}$………………(4)

From equation 4, here given a positive whole number K, an inconspicuous perception x and a likeness metric d, KNN classifier plays out the accompanying two stages:

It goes through the entire dataset registering d among x and each preparation perception. We'll call the K focuses in the preparation information that are nearest to x the set. Note that K is normally odd to counteract tie circumstances [17].

It essentially ascertains the separation of another information point to all other preparing information focuses. The separation can be of any kind e.g. Euclidean or Manhattan and so on. It at that point chooses the K-closest information focuses, where K can be any number.

Now, there are likely thinking about how to pick the variable K and what its impacts are on this classifier. All things considered, as most machine learning calculations, the K in KNN is a hyperparameter that you, as a creator, must pick with the end goal to get the most ideal fit for the informational index. Naturally, it can consider K controlling the state of the choice limit we discussed before.

At the point when K is little, we are limiting the area of a given forecast and driving our classifier to be "more visually impaired" to the general conveyance. A little incentive for K gives the most adaptable fit, which will have low inclination yet high fluctuation. Graphically, our choice limit will be more jagged. On the other hand, a higher K midpoint more voters in every forecast and consequently is stronger to exceptions. Bigger estimations of K will have smoother choice limits which implies bring down change yet expanded predisposition [18].

The quantity of group, K, must be resolved before hand. Its drawback is that it doesn't yield a similar outcome with each run, since the subsequent bunches rely upon the underlying irregular assignments.

We never know the genuine bunch, utilizing similar information, supposing that it is inputted in an alternate request it might deliver diverse group if the quantity of information is few. As far we know that datasets are very much arranged for the KNN display building. Since KNN is a non-parametric calculation, we won't acquire parameters for the model. The KNN () work restores a vector containing element of characterizations of test set.

3.5 Selected Algorithm

We use different algorithm to get highest accuracy from our dataset. In this figure, we are show which algorithm are given best accuracy among another algorithm.

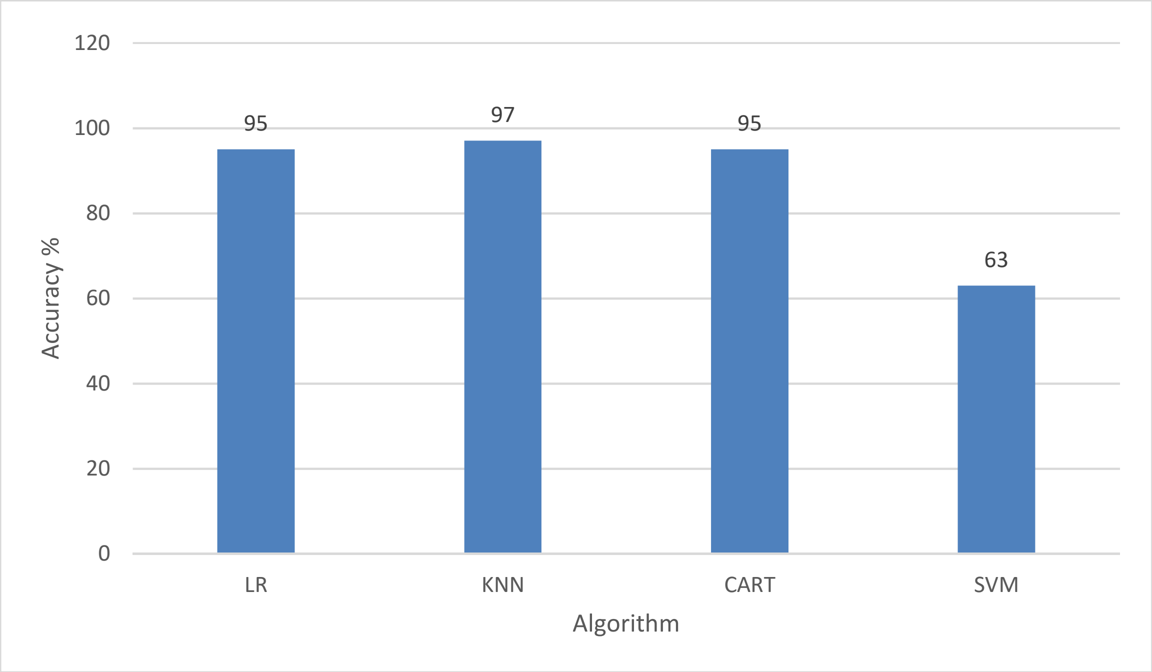

Figure 3.3: Applied Algorithms

In figure 3.3, we have 4 models Logistic Regression, k-nearest classifier, Decision Tree Classifier, SVM and exactness estimations for each. We have to contrast the models with one another and select the most exact KNN.

The little estimation of K will prompt an expansive fluctuation in expectations. On the other hand, setting K to an extensive esteem may prompt a substantial model inclination. Consequently, K ought to be set to an esteem sufficiently huge to limit the likelihood of misclassification and little enough (regarding the quantity of cases in the model example) so the K closest indicates are nearer the inquiry point. KNN is a decent decision when effortlessness and precision are the dominating issues. KNN can be prevalent, if the life expectancy of an information stream is less or where new sets of information arrives quickly and the preparation set is continue evolving. Despite the fact that KNN gives good outcomes, it is excessively costly as far as time and memory.

3.6 Proposed Algorithm

In our proposed method we use K-nearest neighbors' algorithm. In our algorithm we try to build a model which predict brain stroke disease. Our proposed technique means to improve the execution of KNN classifier for illness forecast. Implementation is considered as a fundamental sector for making any system. Anaconda is an environment that consist of python and all deep learning packages. Python version 3.6.3 has been used and considered as a latest version of python. Various types of library Function has been used for implementation. In our method -

Step1.Firstly we select our datasets which contain brain stroke and normal dataset.

Step2.Classification of dataset into patient with brain stroke and normal.

Step3: Input the dataset.

Step4: Apply machine learning algorithm in python.

Step5: Find out highest accuracy from dataset from different machine learning algorithm.

Step6: Get highest accuracy using KNN.

Step7: Measure the performance of the model.

KNN takes the brain stroke dataset and classify whether a person is having brain stroke or not. The above algorithm is applied on pre-processed dataset and performance is measured.

We have to realize that the model we made is any great. Afterward, we will utilize factual strategies to evaluate the exactness of the models that we make on concealed information. We additionally need a more solid gauge of the exactness of the best model on inconspicuous information by assessing it on genuine concealed information. That is, we will keep down a few information that the calculations won't get the opportunity to see and we will utilize this information to get a second and free thought of how precise the best model may really be. We will part the stacked dataset into two, 70% of which we will use to prepare our models and 30% that we will keep down as an approval dataset. We will use 10-fold cross validation to estimate accuracy. This will split our dataset into 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits.

Cross-approval is a resampling method used to assess machine learning models on a restricted information test.

The method has a solitary parameter considered k that alludes to the quantity of gatherings that a given information test is to be part into. In that capacity, the technique is frequently called k-crease cross-approval. At the point when a particular incentive for k is picked, it might be utilized instead of k in the reference to the model, for example, k=10 getting to be 10-overlay cross-approval.

Cross-approval is essentially utilized in connected machine figuring out how to evaluate the aptitude of a machine learning model on concealed information. That is, to utilize a restricted example with the end goal to gauge how the model is relied upon to perform as a rule when used to make forecasts on information not utilized amid the preparation of the model [19].

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Results

To measure the performance of our proposed system we use dataset for 12 features to test the accuracy. Our dataset contains 200 data which is brain stroke and normal and also get the accuracy 97%. We use real dataset that's why we find the high accuracy from our dataset. We also applied Cross Validation Technique and portioned our final dataset into 10 equal subsamples to get the higher accuracy. We also use confusion matrix to calculate precision, recall, F-measure, Support, True Positive Rate, True Negative Rate and accuracy of the model.

The confusion matrix is a table to describe the performance of a classification model on a set of test data. Confusion matrix can define four terms:

True Positive (TP): We predicted result as no brain stroke which are actually no-brain stroke.

True Negative (TN): We predicted result as brain stroke which are actually brain stroke.

False Positive (FP): We predicted No-brain stroke, but these are not actually no brain stroke.

False Negative (FN): We predicted brain stroke, but these are actually no-brain stroke.

Precision: Precision is the piece of related instances among the retrieved instances. high precision means that an algorithm returned substantially more relevant results than irrelevant ones.

$$precision = \frac{tp}{tp + fp}$$

Recall: Recall is the piece of relevant instances that have been retrieved over the total amount of relevant instances. High recall means that an algorithm returned most of the relevant result.

$$Recall = \frac{tp}{tp + fn}$$

F-measure: F-score is a measure of test's accuracy by considering both precision and recall. it is a harmonic average of precision and recall.

$$F - score = \frac{2 * precision * recall}{precision + recall}$$

Accuracy: Accuracy refers to the familiarity of the measured value to a known value.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

True Positive Rate: False positive rate are refers that our proposed method predict the brain stroke is no brain stroke when it's actually brain stroke. Calculate the false positive rate by the given equation:

$$Truepositiverate = \frac{TP}{TN+Fp}$$

Specificity: Specificity refers that our proposed method predicts the brain stroke is a brain stroke when it's actually brain stroke [20]. Calculate the specificity of the given equation:

$$specificity = \frac{TN}{TN+Fp}$$

We know the confusion matrix and this can help in ascertaining further developed arrangement measurements, for example, precision, recall, specificity and sensitivity of our classifier. We also know the confusion matrix as follow as table 4.1 are shown that-

Table 4.1: Confusion matrix

	No-event	event
No-event	true negative 31	false positive 1
event	false negative 1	true positive 27

Here this table 4.1 we see that true negative and positive and false negative and positive in confusion matrix.

Here the details of confusion matrix are shown in table 4.2 and confusion matrix is showed by y-validation and prediction value.

Table 4.2: Measure accuracy

	Precision	Recall or sensitivity	F1-score	Support
no	.97	.97	.97	32
yes	.96	.96	.96	28
avg/total	.97	.97	.97	60

In this table 4.2 we can see that confusion matrix are described such as accuracy, precision, recall, fi-score, support for our dataset.

We also get high Accuracy 97% for our research using KNN algorithm. In this figure 4.1 we get accuracy curve for our dataset and we get 97% accuracy for our dataset.

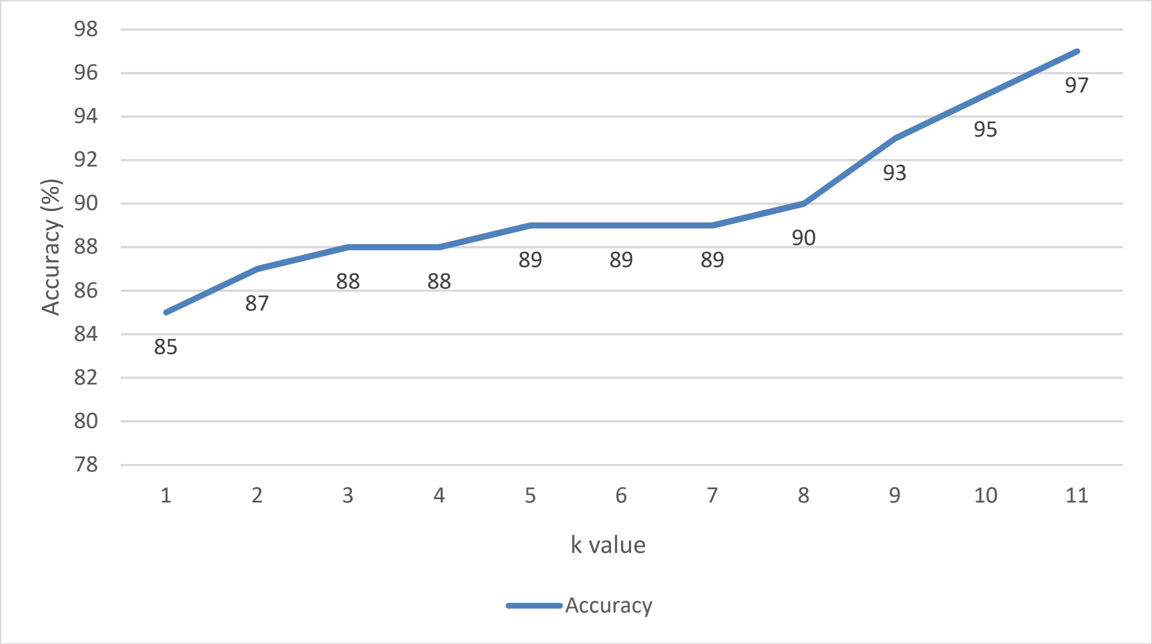

Figure 4.1: Classification of Accuracy

Here accuracy of our dataset is shown for k=11 and we also try to our best to make a good accuracy using machine learning algorithm.

In most straightforward terms, given an arrangement of information focuses from rehashed estimations of a similar amount, the set can be said to be exact if the qualities are near one another, while the set can be said to be precise if their normal is near the genuine estimation of the amount being estimated. In the primary, more typical definition over, the two ideas are free of one another, so a specific arrangement of information can be said to be either exact, or exact, or both, or not one or the other.

CHAPTER 5

Conclusion and Future works

5.1 Conclusion

Brain stroke is the main reason of blood supply. The current investigation has gone for utilizing existing information and to evaluate the weight of stroke. The estimations depend on a few suppositions where the estimation of every part is questionable [21].

In this paper, we have described KNN algorithms to anticipate 1 year 2-month stroke results from records of physiological parameters amid after stroke. We have evaluated enhancements through the incorporation of physiological pattern designs as highlights in our algorithms. Episode stroke can be precisely anticipated utilizing self-detailed data concentrated on well-being practices. Hazard evaluation can be performed with populace well-being overviews to help populace well-being arranging or outside of clinical settings to help quiet engaged avoidance. We trust that these patterns assume a vital job on early clinical medications of stroke patients. The productivity and precision of our algorithm have likewise been shown through analyses on a genuine informational index of stroke patients [22].

5.2 Future Work

All of the underlying methodological and computational complexities aside, our long-term goal is to design an easy to use online system, allowing for relative prediction of the clinical outcome based on the demographics and clinical findings. Such a system has the potential for fine adjustment from the continuous training provided via handling large-scale national or international multi-institutional users, with the advantage of easily incorporating newly available data to improve prediction performance. Another side to focusses our work perinatal Stroke, childhood Stroke.

REFERENCES

[1] Banglanews24.com available at https://www.banglanews24.com/health/article/9540/Brain-Stroke-Must-know(accessed on Monday October 08,2018).

[2]_Mayo_Clinic._Available_at_https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113.(accessed on May 16, 2018).

[3]_World_Health_Rankings.Available_at_https://www.researchgate.net/publication/230847528_Burden_of_stroke_in_Bangladesh (accessed on 3 February 2012).

[4] Hossain AM, Ahmed NU, Rahman M, Islam MR, Sadhya G, Fatema K. Analysis of sociodemographic and clinical factors associated with hospitalized stroke patients of Bangladesh. Faridpur Med Coll J 2011; 6:19–23.

[5] Rate of strokes very high in Bangladesh. Available at https://www.thedailystar.net/backpage/rate-strokes-very-high-bangladesh1577461(accessed on May 20, 2018).

[6] Patrick Lucket, Elena Pavelescu, Todd McDonald, Lee Hively, Juan Ochoa et al, Predicting state transitions in brain dynamics through spectral difference of phase-spacegraphs.12 October,2018. s10827-018-0700-1.

[7] Jeroen de Bresser, Marileen P. Portegies, Alexander Leemans, Geert Jan Biessels, L. Jaap Kappelle, Max A. Viergever et al, A comparison of MR based segmentation methods for measuring brain atrophy progression. 54 (2011) 760–768.

[8] Harold P. Adams Jr., MD; Birgitte H. Bendixen, PhD, MD; L. Jaap Kappelle, MD; Jose Biller, MD; Betsy B. Love, MD; David Lee Gordon, MD; E. Eugene Marsh III, MD; and the TOAST Investigators et al. Classification of Subtype of Acute Ischemic Stroke. doi: 10.1161/01.STR.24.1.35.

[9] Mariana Bento, Yan Sym, Richard Frayne, Roberto Lotufo1, and Let´ıcia Rittner a et al, Probabilistic Segmentation of Brain White Matter Lesions Using Texture-Based Classification. DOI: 10.1007/978-3-319-59876-5 9.

[10] R. Guerreroa, C. Qina, O. Oktaya, C. Bowlesa, L. Chena, R. Joulesb, R. Wolzb,a, M.C. Valdés-Hernándezc, D.A Dickiec, J.Wardlawc, D.Rueckerta a et al, White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. Neurology. 17 (2018) 918–934.

[11] Md. Nazmul Islam1,2, Mohammed Moniruzzaman1,2, Md. Ibrahim Khalil1,2, Rehana Basri3, Mohammad Khursheed Alam3, Keat Wei Loo4, and Siew Hua Gan4 a et al, Burden of stroke in Bangladesh. Doi:1747-4949.2012.

[12] Rita V. Krishnamurthi, Suzanne Barker-Collo, Varsha Parag, Priyakumari Parmar, Emma Witt, Amy Jones, Susan Mahon, Craig S. Anderson, P. Alan Barber, Valery L. Feigin, MD, Ph an et al, Stroke Incidence by Major Pathological Type and Ischemic Subtypes in the Auckland Regional Community Stroke Studies Changes Between 2002 and 2011 Neurology. 2018; 49:3-10.

[13] Donnan GA, Fisher M, Macleod M, Davis SM (May 2008). "Stroke". Lancet. 371 (9624): 1612–23.

[14]Logistic regression for machine learning , https://machinelearningmastery.com/logistic-regression-for-machine-learning/(accessed on April 1, 2016).

[15] Classification and regression trees, https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/ (accessed on April 8, 2016).

[16] Machine learning 101, https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72(accessed on April 10, 2016).

[17] Kevin Zakka's Blog.Available at https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/ (accessed on 13, 2016).

[18]_PMC_Available_at_https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/?fbclid=IwAR2Zc0VMkyLeFKzYw_xHg-eA8iIjJ1Cfo_x5FFd0I34hbg7zGmzVNGWfmVM (accessed on 2016 Jun 4).

[19] A Gentle Introduction to k-fold Cross Validation available at https://machinelearningmastery.com/k-fold-cross-validation/(accessed on May 23, 2018).

[20] Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures .Available at https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/ (accessed on September 9, 2016).

[21] Kidwell CS, Warach S (December 2003). "Acute ischemic cerebrovascular syndrome: diagnostic criteria". Stroke. 34 (12): 2995–8.

[22] Fairhead JF, Mehta Z, Rothwell PM (2005). "Population-based study of delays in carotid imaging and surgery and the risk of recurrent stroke". Neurology. 65 (3): 371–5.