Prediction of Stroke Using Machine Learning

Conference Paper · June 2020

Srikanth .S
Visvesvaraya Technological University

Prediction of Stroke Using Machine Learning

KUNDER AKASH MAHESH
Dept. of Computer Science & Engineering
CMRIT, Bangalore
Karnataka, India
akma16cs@cmrit.ac.in

SHASHANK H N
Dept. of Computer Science & Engineering
CMRIT, Bangalore
Karnataka, India
shhn16cs@cmrit.ac.in

SRIKANTH S
Dept. of Computer Science & Engineering
CMRIT, Bangalore
Karnataka, India
sris16cs@cmrit.ac.in

THEJAS A M
Dept. of Computer Science & Engineering
CMRIT, Bangalore
Karnataka, India
them16cs@cmrit.ac.in

Abstract - Stroke is a blood clot or bleeds in the brain, which can make permanent damage that has an effect on mobility, cognition, sight or communication. Stroke is considered as medical urgent situation and can cause long-term neurological damage, complications and often death. The majority of strokes are classified as ischemic embolic and Hemorrhagic. An ischemic embolic stroke happens when a blood clot forms away from the patient brain usually in the patient heart and travels through the patient bloodstream to lodge in narrower brain arteries. Hemorrhagic stroke is considered another type of brain stroke as it happens when an artery in the brain leaks blood or ruptures. Stroke is the second leading cause of death worldwide and one of the most life- threatening diseases for persons above 65 years. It injures the brain like “heart attack” which injures the heart. Once a stroke disease occurs, it is not only cost huge medical care and permanent disability but can eventually lead to death. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage.

INTRODUCTION

Burden of Stroke in the World-Stroke is the second leading cause of death and leading cause of adult disability worldwide with 400-800 strokes per 100,000, 15 million new acute strokes every year, 28,500,000 disability adjusted life-years and 28-30-day case fatality ranging from 17% to 35%. The burden of stroke will likely worsen with stroke and heart disease related deaths projected to increase to five million in 2020, compared to three million in 1998. This will be a result of continuing health and demographic transition resulting in increase in vascular disease risk factors and population of the elderly. Developing countries account for 85% of the global deaths from stroke. The social and economic consequences of stroke are substantial. The cost of stroke for the year 2002 was estimated to be as high as $49.4 billion in the United States of America (USA), while costs after discharge were estimated to amount to 2.9 billion Euros in France.

Causes of mortality from stroke- Death from stroke is as a result of co-morbidities and/ or complications. Complications of stroke may arise at different time periods. The beginning of stroke symptoms and the first month following the stroke onset is the most critical period for survival with the highest number of fatalities in the first week. Complications of stroke include hyperglycemia, hypoglycemia, hypertension, hypotension, fever, infarct extension or rebreeding, cerebral edema, herniation, coning, aspiration, aspiration pneumonia, urinary tract infection, cardiac dysrhythmia, deep venous thrombosis and pulmonary embolism among others. During the first week from stroke onset, death is usually due to transtentorial herniation and hemorrhage, with death due to hemorrhage happening within the first three days and death due to cerebral infarction usually occurring between the third to sixth day. One week after the onset of stroke, death is usually due to complications resulting from relative immobility such as pneumonia, sepsis and pulmonary embolism.

Different studies have found varied factors associated with stroke mortality in their setting. For example, the most common predictors of death from stroke for those aged more than 65 years of age reported by Mackay included previous stroke, atrial fibrillation and hypertension. Nigeria 6 reported a 12.6% 30-day case fatality of all strokes. Among patients with hemorrhagic stroke: fixed dilated pupil(s), a Glasgow coma score of less than 10 on admission, swallowing difficulties at admission, fever, lung infection, and no aspirin treatment were independent risk factors for a lethal outcome. Yikona J et al also observed that stroke severity, neurological deterioration during hospitalization, non-use of antithrombolytics during hospital admission and lack of assessment by a stroke team were the most consistent predictors of case fatality at seven days, 30 days and one year after stroke. In Pretoria, South Africa, case fatality at 30 days was much higher, 22% for ischemic stroke, 58% for cerebral hemorrhagic stroke and hypertension was significantly associated with stroke. At Mulago hospital, 30 day case fatality of 43.8% was reported among 133 patients (mean age 65.8+ 15.8 years) with, fever > 37.50 (OR 2.81 (95%CI; 1.2-6.6) and impaired level of consciousness with a GCS <9 (OR0.13 95%CI; 0.005-0.35) significantly associated with increased mortality.

Traditional risk factors associated with stroke- Stroke can occur in anyone regardless of race, gender or age however the chances of having a stroke increase if an individual has certain risk factors that can cause a stroke. The best way to protect oneself and others is to understand personal risk and how to manage it.

Studies have shown that 80% of strokes can be prevented in this way. Stroke risk factors are divided into modifiable and non-modifiable. The modifiable risk factors are further subdivided into lifestyle risk factors or medical risk factors. Lifestyle risk factors which include smoking, alcohol use, physical inactivity and obesity can often be changed while medical risk factors such as high blood pressure, atrial fibrillation, diabetes mellitus and high cholesterol can usually be treated. A large multicenter (INTERSTROKE) case control study showed that there are ten factors that are associated with 90% of stroke risk and half of these are modifiable. Non-modifiable risk factors on the other hand though they cannot be controlled, they help to identify individuals at risk for stroke.

Prevention of stroke - More than 70% of strokes are first events, thus making primary stroke prevention a particularly important aspect. Interventions should be targeted at behavior modification, which however requires information about the baseline perceptions, knowledge and prevalence of risk factors in defined populations.

PROBLEM STATEMENT

Stroke is the second leading cause of death worldwide and remains an important health burden both for the individuals and for the national healthcare systems. Potentially modifiable risk factors for stroke include hypertension, cardiac disease, diabetes, and dysregulation of glucose metabolism, atrial fibrillation, and lifestyle factors. Therefore, the goal of our project is to apply principles of machine learning over large existing data sets to effectively predict the stroke based on potentially modifiable risk factors. Then it intended to develop the application to provide a personalized warning on the basis of each user’s level of stroke risk and a lifestyle correction message about the stroke risk factors.

LITERATURE SURVEY

In order to get required knowledge about various concepts related to the present analysis existing literature were studied. Some of the important conclusions were made through those are listed below.

“**Computer Methods and Programs in Biomedicine**” - Jae–woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho – The Purpose of this paper was Calculation of 10-year stroke prediction probability and classifying the user's individual probability of stroke into five categories.
“**Probability of Stroke: A Risk Profile from the Framingham Study**” - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD - In this paper, A health risk appraisal function has been developed for the prediction of stroke using the Framingham Study cohort.

“**Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study**” - Min SN, Park SJ, Kim DJ, Subramaniyam M, Lee KS – In this research, this paper aimed to derive a model equation for developing a stroke pre- diagnosis algorithm with the potentially modifiable risk factors.

"**Stroke prediction using artificial intelligence**"- M. Sheetal Singh, Prakash Choudhary - In this paper, Here, decision tree algorithm is used for feature selection process, principle component analysis algorithm is used for reducing the dimension and adopted back propagation neural network classification algorithm, to construct a classification model.

"**Medical software user interfaces, stroke MD application design (IEEE)**" Elena Zamsa-The article presents the design of an application interface for associated medical data visualization and management for neurologists in a stroke clustering and prediction system called Stroke MD.

"**Focus on stroke: Predicting and preventing stroke**" Michael Regnier- This paper focuses on cutting-edge prevention of stroke.

"**Effective Analysis and Predictive Model of Stroke Disease using Classification Methods**"-A.Sudha, P.Gayathri, N.Jaisankar- This paper, principle component analysis algorithm is used for reducing the dimensions and it determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not.

"**Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study**" - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal - In this paper Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. This article aimed to develop and validate a set of deep learning algorithms for automated detection.

PROPOSED SYSTEM

Algorithms Involved-

Few methodologies used in our projects are:

1. Decision Tree
2. Naïve Bayes
3. Artificial Neural Network

Decision Tree- A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision tree is one of the important methods for handling high dimensional data. Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike the linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand. Fig 1 represents part of the decision tree model for prediction of stroke diseases.

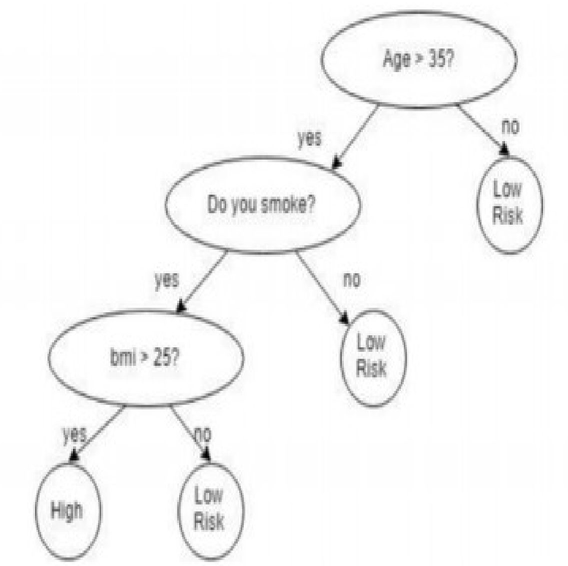


Fig 1: - Decision tree

Naive Bayes- A Naïve Bayes classifier is a probabilistic machine-learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of A happening, given that B has occurred. Hence, B is the evidence and A is the hypothesis. The assumption made here is that the predictors/features are independent. That is the presence of one particular feature does not affect the other. Hence it is called naïve.

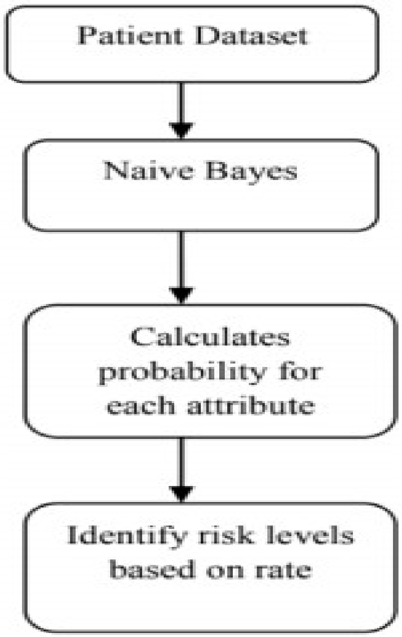

Patient Dataset
Naive Bayes
Calculates probability for each attribute
Identify risk levels based on rate

Fig 2: - Bayesian classifier

Naive Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real-life cases, the predictors are dependent; this hinders the performance of the classifier.

Artificial Neural Network- Neural networks are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

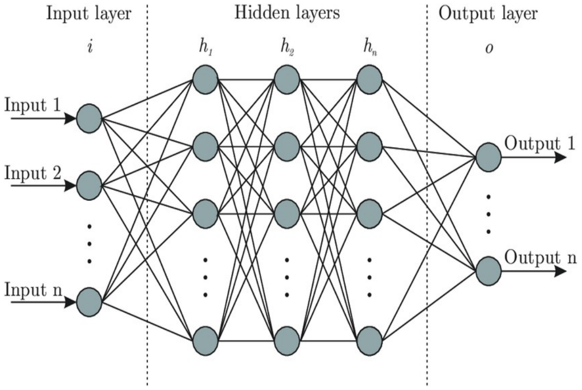
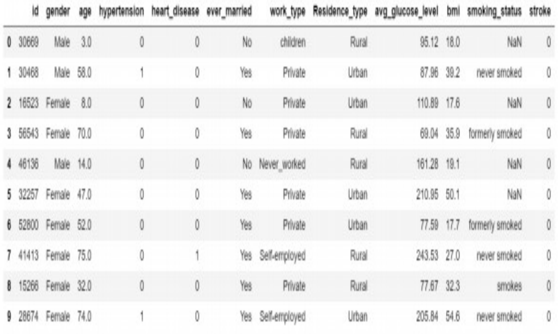


Fig 3: Artificial Neural Network

Neural networks help us cluster and classify. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on.

DATASET USED-



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0
5	32257	Female	47.0	0	0	Yes	Private	Urban	210.95	50.1	NaN	0
6	52800	Female	52.0	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
7	41413	Female	75.0	0	1	Yes	Self-employed	Rural	243.53	27.0	never smoked	0
8	15266	Female	32.0	0	0	Yes	Private	Rural	77.67	32.3	smokes	0
9	28674	Female	74.0	1	0	Yes	Self-employed	Urban	205.84	54.6	never smoked	0

Fig 4: -Dataset

Architectural Design

System architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. The overall logical structure of the project is divided into processing modules and a conceptual data structure is defined as Architectural Design.

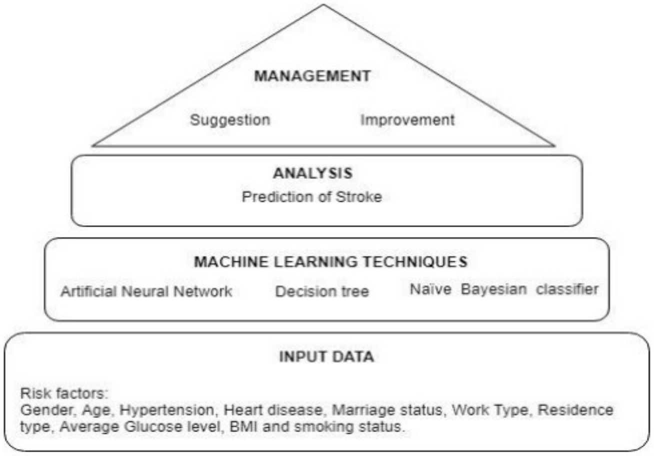

MANAGEMENT
Suggestion
Improvement
ANALYSIS
Prediction of Stroke
MACHINE LEARNING TECHNIQUES
Artificial Neural Network
Decision tree
Naïve Bayesian classifier
INPUT DATA
Risk factors:
Gender, Age, Hypertension, Heart disease, Marriage status, Work Type, Residence type, Average Glucose level, BMI and smoking status.

Figure 5: System Architecture

Figure 5 shows the overall logical structure of the project with following modules:

1. Input data: Risk factors like age, gender, hypertension, heart disease, BMI, Smoking status, Glucose level.
2. Machine Learning Techniques: Artificial Neural Networks, Decision Tree, Naïve Bayes classifier.
3. Analysis: Prediction and analysis of stroke whose performance is based on machine learning techniques.
4. Management: Suggestion and improvement of stroke victims.

WORK FLOW

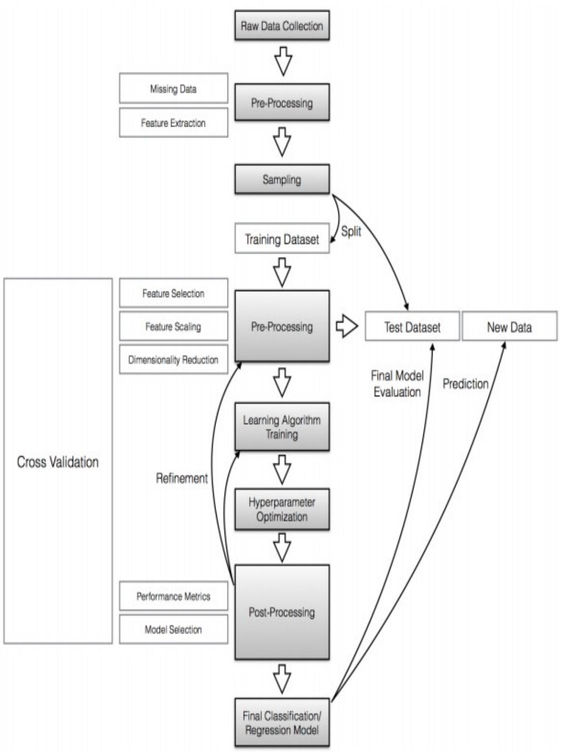


Fig 6: - Work Flow

IMPLEMENTATION STEPS-

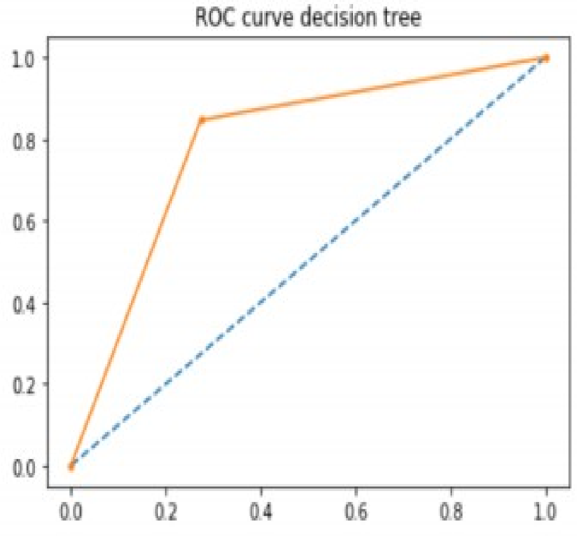
1. Clean the missing values both training and testing data
2. Applying Label Encoder to convert object into integer
3. Balancing Dataset
4. Split the data into training and testing
5. Building Decision Tree Model
6. Building Naïve Bayes Model
7. Building Artificial Neural Networks Model
8. Create a GUI and extract models into GUI module
9. Enter the new data for which stroke has to be predicted
10. Result: -Predicted data with respect to each model

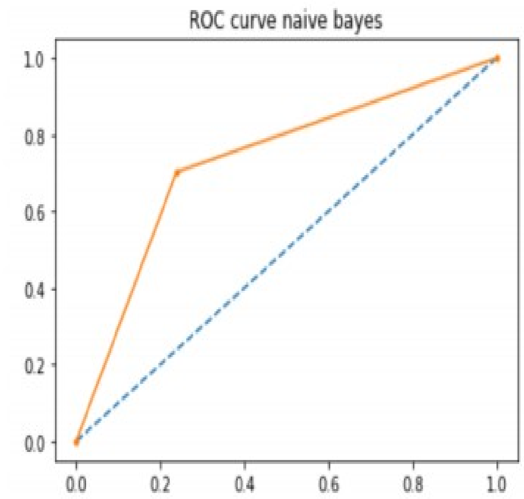
RESULTS AND PERFORMANCE EVALUATION-

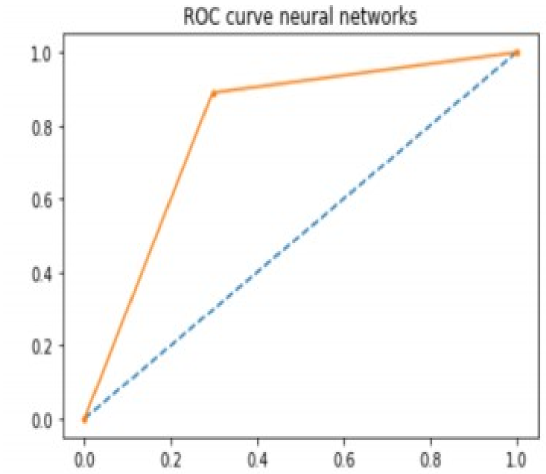
Performance Analysis

In this section snapshot showing the performance of three algorithms proposed in this project i.e. Decision Tree, Naïve Bayes, Artificial Neural Network are compared. AUC – ROC (Area Under The Curve - Receiver Operating Characteristics) curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

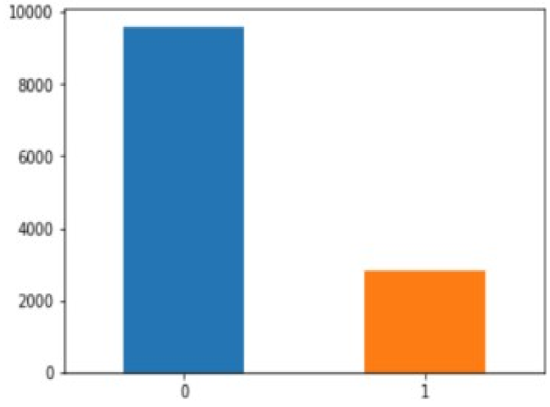
The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.



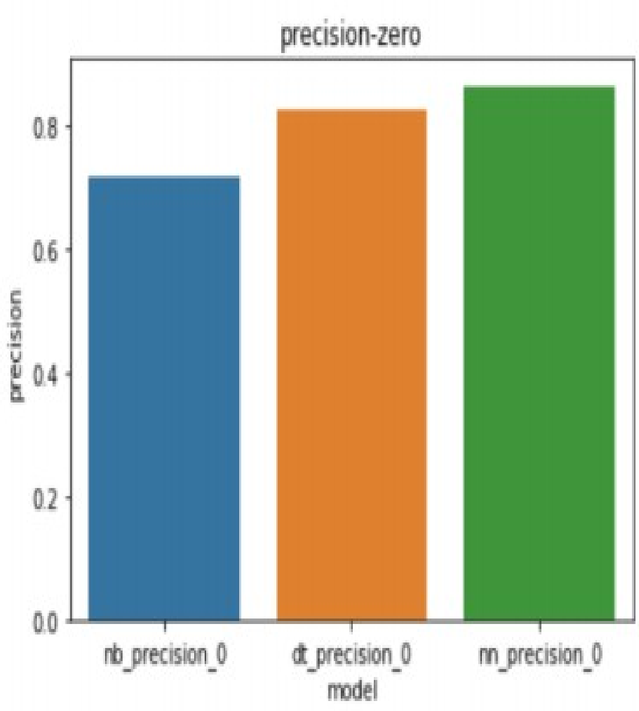


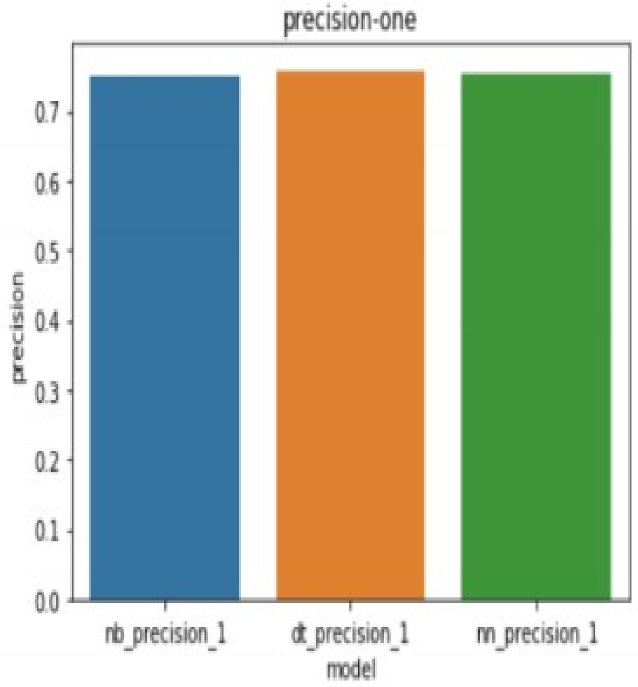


Graphs and Analysis

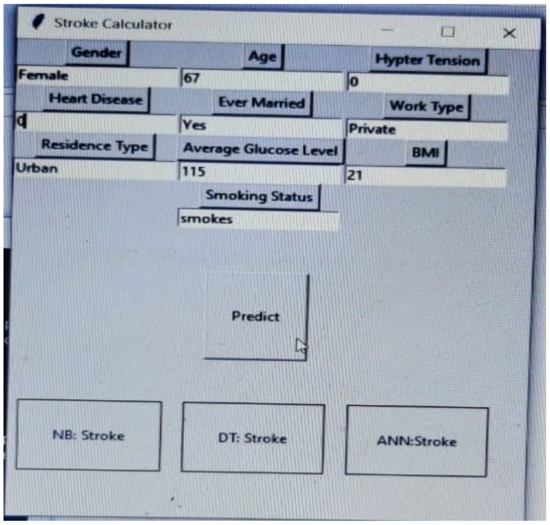


Frequency of Stroke before Balanced Dataset





GUI



Stroke Calculator

Gender: Female
Age: 67
Hypter Tension: 0
Heart Disease: 0
Ever Married: Yes
Work Type: Private
Residence Type: Urban
Average Glucose Level: 115
BMI: 21
Smoking Status: smokes

Conclusion

Several assessments and prediction models, Decision Tree, Naive Bayes and Neural Network, showed acceptable accuracy in identifying stroke-prone patients. This project hence helps to predict the stroke risk using prediction model and provide personalized warning and the lifestyle correction message through a web application. By doing so, it urges medical users to strengthen the motivation of health management and induce changes in their health behaviors.

Future Scope

This project helps to predict the stroke risk using prediction model in older people and for people who are addicted to the risk factors as mentioned in the project. In future, the same project can be extended to give the stroke percentage using the output of current project. This project can also be used to find the stroke probabilities in young people and underage people by collecting respective risk factor information's and doctors consulting.

REFERENCES

[1]. “Computer Methods and Programs in the Biomedicine” - Jae–woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho

[2]. “Probability of Stroke: A Risk Profile from the Framingham Study” - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD

[3]. “Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study” - Min SN, Park SJ, Kim DJ, Subramaniyam M, Lee KS

[4]. “Stroke prediction using artificial intelligence”- M. Sheetal Singh, Prakash Choudhary

[5]. “Medical software user interfaces, stroke MD application design (IEEE)” - Elena Zamsa

[6]. “Focus on stroke: Predicting and preventing stroke” - Michael Regnier

[7]. “Effective Analysis and Predictive Model of Stroke Disease using Classification Methods” - A.Sudha, P.Gayathri, N.Jaisankar

[8]. “Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study” - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal