Stroke Prediction System Using Machine Learning Algorithms

[1]Mr. T.N.Sudhahar, [2]Dr. S.Meera, [3]K.Saraswathi, [4]S.Sheryl Catherine, [5]B.Sathya Priya

[1]Assistant Professor, [2]Head of the Department/Professor, [3]Student, [4]Student, [5]Student
Computer Science and Engineering,
Agni College of Technology, Chennai, India

***Abstract*: Stroke could be a medical condition that may result in the death of someone. It's a severe condition and if treated on time we are able to save one's life and treat them well. The target of this study is to construct a prediction model for predicting stroke and to assess the accuracy of the model. Therefore, the project mainly aims at predicting the possibilities of occurrence of stroke using the emerging Machine Learning techniques. The most of strokes are classified as ischemic embolic and Hemorrhagic. An ischemic embolic stroke happens when a blood forms off from the patient brain usually within the patient heart and travels through the patient bloodstream to live narrower brain arteries. Hemorrhagic stroke is taken into account as another sort of brain stroke because it happens when an artery within the brain leaks blood or ruptures.**

***Index Terms*: Machine Learning, Stroke, Classification, Pre- diction**

I. Introduction

A stroke could be a medical condition within which poor blood flow to the brain causes cell death. There are two main kinds of stroke: ischemic, because of lack of blood flow, and hemorrhagic, thanks to bleeding. Both cause parts of the brain to prevent functioning properly. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to at least one side. Signs and symptoms often appear soon after the stroke has occurred. Two principal types of stroke are CVA and hemorrhage. Ischemic stroke happens due to absence of blood stream and hemorrhagic stroke happens due to bleeding. Hemorrhage is ordered in to two kinds' subarachnoid hemorrhage and intracerebral hemorrhage. Transient ischemic assault is otherwise called "ministroke". Stroke denies individual's mind of oxygen and supplements, which ends up within the death of dead cells when stroke occurs. It's not only very expensive for the medical treatments and a permanent disability but can ultimately prompt demise. A machine learning model would take the patients information and propose a bunch of suitable Expectation. The framework can remove concealed information from a chronicled clinical data set and may anticipate patients with infection and utilize the clinical profiles like Age, force per unit area, Glucose, then forth it can foresee the probability of patients getting an illness. Grouping calculations are utilized with the amount of properties for expectation of illness. Prevention of stroke - over 70% of strokes are first events, thus making primary stroke prevention a very important aspect. Interventions should be targeted at behavior therapy, which however requires information about the baseline perceptions, knowledge and prevalence of risk factors in defined populations. The goal of our project is to use principles of machine learning over large existing data sets to effectively predict the stroke supported potentially modifiable risk factors. Then it intended to develop the appliance to produce a personalized warning on the premise of every user's level of stroke risk.

II. Related work

Researchers have worked on early diagnosis of neurological disorders, which may be hard to attain. Their contributions and related work are presented here to grasp the premise of the project. Ane Alberdi et al.(2018) had collected home behavior data of patients so as to detect the symptoms for Alzheimer's disease. The symptoms that they were ready to obtain as a final result were associated with the mood, cognition and mobility of patients. The researchers worked on a wise home solution by which sensors can monitor the patients and help in detecting the multiple symptoms. The information assessed was from 29 older adults, who lived in smart homes and monitored for a duration of but 1 month to 60 months. Regression models and classification models were also used on the information. This research helped in obtaining necessary behavioral features for detecting AD in patients and also the changes in patients that might indicate the AD symptoms. Stroke is one more up- set whose risk detection could be a challenge. Yonglai Zhang et al.(2018) conducted a pursuit on stroke patients to detect the danger of stroke. The dataset worked upon comprised of medical tests and other archived data on 792 patients at a Beijing.Predicting Stroke Risk with an Interpretable Classifier, The work of Sergio Peñafiel was supported by the Conicyt (Chile) Master Scholarship under Grant 22180506. The development of a prediction method which supplies information about the most probable causes of a high stroke risk and may cope with incomplete data records. It's based on the Dempster-Shafer theory of plausibility. It determines whether or not a patient will have a stroke within the next year automatically using her/his historical medical information. This problem is seen as a binary classification problem, where our input is that the historical information defined within the previous section and also the outcome are two classes, namely "The patient will have stroke within the subsequent year", and "The patient won't have stroke within the following year". The range of 1 year was chosen because it's an inexpensive period of your time for taking preventive action within the case there's high stroke risk.

III. Algorithms

1. DECISION TREE:

A decision tree is a decision support tool that uses a tree- like model of choices and their possible consequences, including occurrence outcomes, resource costs, and utility. It's a method to display an algorithm that only contains conditional control statements.

2. LOGISTIC REGRESSION:

Logistic regression estimates the probability of an event occurring, like voted or didn't vote, based on a given dataset of independent variables. Since the result is a probability, the variable quantity is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure.

3. NAIVE BAYES:

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, rep- resented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

4. RANDOM FOREST CLASSIFIER:

Random forest is a classifier that contains variety of decision trees on various subsets of the given dataset and takes the typical to enhance the predictive accuracy of that dataset. Rather than looking forward to one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the ultimate output.

IV. Libraries Imported

1. NumPy

Python library which deals with arrays, basically used for scientific computations. Used for performing linear algebra, matrix multiplication, Fourier transform.

2. Pandas

Used analyze data. Works on various file formats such as SQL, JSON, and Microsoft Excel. Data manipulation operations such as merging, selecting, reshaping and data cleaning in general.

3. Matplotlib Pyplot

Has collection of functions that makes matplotlib works like MATLAB. Basically, used for data visualization, also includes functions such as creating a figure, plotting area etc.

Data Cleaning

Data was cleaned for missing data and null values. Missing data was dealt by removing the rows with null values or redundant values.

Data Analysis

There are three varieties of data analysis which is per- formed i.e., Categorical feature analysis, Numerical feature analysis and Multicollinearity analysis. Data analysis is completed to point out us the hidden relationships and attributes present within the dataset which help the machine learning model to perform better.

Implementing algorithms

Four different algorithms are selected after literature survey. Comparative study is formed between these four algorithms - Decision Tree, Logistic Regression, Random Forest and Naïve Bayes.

Cross Validation

Effectiveness of all the models is verified to unravel over- fitting problems. Overall assessment on how the model will perform for an independent test dataset. Finally, the simplest performing model are used to predict stroke using the file given by the user. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality.

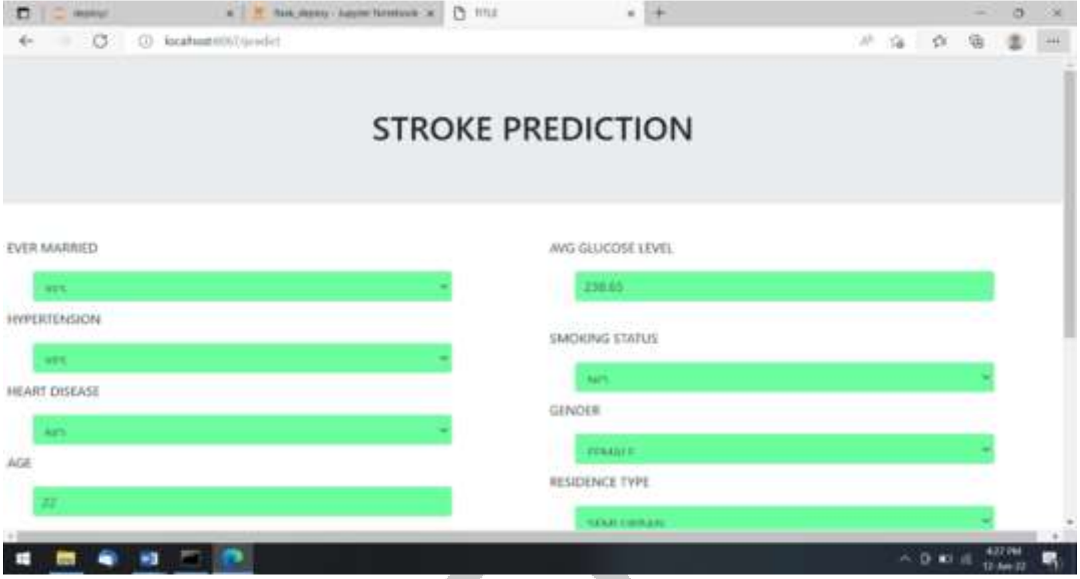
Implementing algorithms

In the past it had been referred as IPython Notebooks. It's an internet intelligent computational platform for creating Jupyter journal records. The "notebook" term can conversationally make reference to various elements, primarily the Lupyter web application, Jupyter Python web server, or Jupyter report design contingent upon setting.

Result and Conclusion

In this paper, five learning techniques were explored to predict Stroke. We made the following observations after significant analysis. All the five models are compared and best performing model is considered for prediction. Conclusion Several assessments and prediction models, Decision Tree, Logistic Regression, Naïve Bayes and Random Forest, 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.

Fig.1 Output



STROKE PREDICTION

EVER MARRIED

yes

HYPERTENSION

yes

HEART DISEASE

no

AGE

22

AVG GLUCOSE LEVEL

238.65

SMOKING STATUS

no

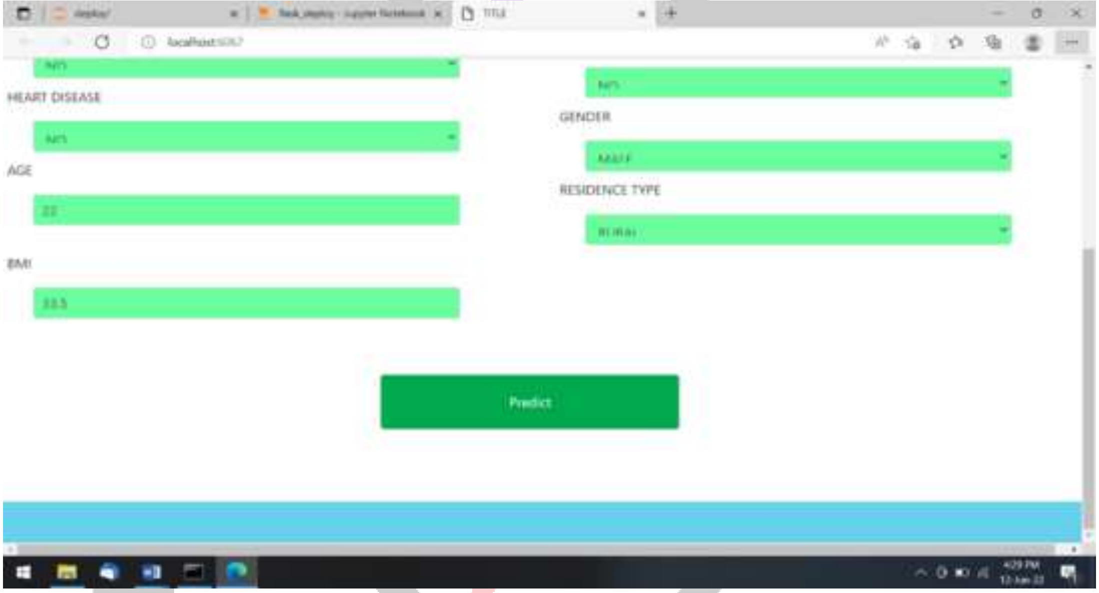
GENDER

FEMALE

RESIDENCE TYPE

SEMI URBAN

Fig.2 Output



no

HEART DISEASE

no

AGE

22

BMI

33.5

no

GENDER

MALE

RESIDENCE TYPE

RURAL

References

[1] Jeena R.S and Dr.Sukesh Kumar used SVM with appropriate kernel functions which had been investigated for analysis of stroke. Pre-processing was done to remove redundant and incompatible data, 350 inputs were taken for the prediction.

[2] Prediction of Stroke Using Machine Learning Kunder Akash Mahesh, Shashank H N, Srikanth S and Thejas A M from Dept. of Computer Science & Engineering CMRIT, Bangalore Karnataka, India, used principles of machine learning over large existing data sets to effectively predict the stroke based on potentially modifiable risk factors.

[3] "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.

[4] Stroke Prediction Using Machine Learning Algorithms by Harshitha K V, Harshitha P, Gunjan Gupta, Vaishak P and Prajna K B from Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India. The project mainly aims at predicting the chances of occurrence of stroke using the emerging Machine Learning techniques.

[5] Predicting Stroke Risk with an Interpretable Classifier, The work of Sergio Peñafiel was supported by the Conicyt (Chile) Master Scholarship under Grant 22180506. The development of a prediction method which gives information about the most probable causes of a high stroke risk and can deal with incomplete data records. It is based on the Dempster-Shafer theory of plausibility.