International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-8, Issue-4, July 2021
https://doi.org/10.21276/ijirem.2021.8.4.2
www.ijirem.org

Stroke Prediction Using Machine Learning Algorithms

***Harshitha K V**
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore, Karnataka, India

Harshitha P
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore Karnataka, India

Gunjan Gupta
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore Karnataka, India

Vaishak P
Student, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore Karnataka, India

Prajna K B
Assistant Professor, Department of Electronics & Communication Engg, Nitte Meenakshi Institute of Technology, Bangalore Karnataka, India

ABSTRACT
Stroke is a destructive illness that typically influences individuals over the age of 65 years age. Prediction of stroke is a time consuming and tedious for doctors. Therefore, the project mainly aims at predicting the chances of occurrence of stroke using the emerging Machine Learning techniques. Five different algorithms are used and a comparison is made for better accuracy. Aim is to create an application with a user-friendly interface which is easy to navigate and enter inputs.

Keywords
AI, Data Pre-processing, Classification, Machine Learning.

1. INTRODUCTION
Stroke is a potentially lethal consequence of atrial fibrillation which can lead to death. Prediction of stroke is a time consuming and tedious for doctors. Stroke is a destructive illness that typically influences individuals over the age of 65 years age. It harms the cerebrum like "coronary episode" which harms the heart and has been positioned third driving reason for death in the US and agricultural nations. A stroke happens when the blood supply to an individual's cerebrum is hindered or decreased. Two principal sorts of stroke are ischemic stroke and hemorrhagic stroke. Ischemic stroke happens because of absence of blood stream and hemorrhagic stroke happens because of bleeding. Hemorrhagic stroke 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 results in the death of dead cells when stroke occurs. It's not only very expensive for the medical treatments and a permanent disability but can at last prompt demise. By and large, Data Mining assumes an imperative part in the forecast of illnesses in medical care industry. A significant subject of AI in medication is utilized in this project. 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 can anticipate patients with infection and utilize the clinical profiles like Age, blood pressure, Glucose, and so forth it can foresee the probability of patients getting an illness. Grouping calculations are utilized with the quantity of properties for expectation of illness. The clinical record additionally comprises his clinical history of illnesses and strokes also assume he has had a stroke before too and we take all that data and train the machine dependent on various models, for example, Decision tree, SVM, Logistic regression and so on.

2. LITERATURE SURVEY
Tasfia Ismail Shoily et al., has compared the different models between Naive Bayes, J48, k-NN, and Random Forest, we observe Naive Bayes has better precision. Different medical reports are observed to obtain the dataset which was cross referenced by medical experts and used with WEKA (Waikato Environment for Knowledge Analysis). The model which is developed will help patients to be cautious whether they may get a stroke or not. Trained 4 different models such as Naive Bayes, J48, k-NN and Random Forest. Precision and accuracy was observed to validate the models. The dataset is applied to the machine learning models [1].

JoonNyung Heo et al., have considered three machine learning models which were used based on specific parameters which are deep neural network, random forest, and logistic regression to predict stroke. Studying this paper, we understood that Deep neural network (DNN) is widely used for ischemic or acute stroke patients which also has an impact for long term prediction. The DNN model has an accuracy of 88% with respect to the inputs which was better than the other models. Automated and more precise calculations are done to improve the model, decreasing use of simpler models [2].

Jaehak Yu et al., preferred C4.5 decision tree algorithm which uses NIHSS score which in turn takes real time values and classifies stroke based on severity which is of four different classes, also gives us information on possible disabilities due to stroke. This information helps in determining the time of stroke and disability which may occur thereby helping in taking necessary precautions and other medications. Naive Bias has accuracy of 85.4% and random forest has high accuracy of 88.9% [3].

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. It was run on MATLAB which led to 91% accuracy [4].

Chutima Jalayondeja has conferred that in the prediction using demographic data and Decision Tree, Naïve Bayes, and Neural Network are the 3 models which were considered and Decision Tree was observed with highest accuracy and low FP (false positive), by low FP rate means high accuracy in predicting weather the patients had stroke but was not stroke, whereas FN (false negative) predicts no-stroke but patients had actual stroke. FN is the dangerous as it leads to mortality since the patients has stroke but predicts the opposite. In the view of accuracy, Decision Tree was the considered, but with respect to safety, Neural Network was taken because it had high FP value and the low in FN value [5].

Benjamin Letham et al., prophesized a predictive analysis using Bayesian model known as Bayesian Rule Lists (BRL), creates a distribution over permutations which starts from a large, processed set of data. The pre-processed data, reduces the model space for various sets of fragments and due to this the algorithm scales with least of the data set than have many features. The BRL method helps to attain high accuracy, precision, and tractability [6].

Pei-Wen Huang1 et al., considered Multi-modal analysis method and predicted stroke taking physiological data. This data included Electrocardiography (EKG), Arterial blood pressure (AKG) and Photoplethysmography (PPG). They have studied each of these signals for its accuracy. Also, they clubbed all the three signals stating that multi model analysis has higher accuracy for the prediction of stroke [7].

3. OBJECTIVE AND SCOPE

The prime objective of this project is to construct a prediction model for predicting stroke using machine learning algorithms. The dataset was obtained from Kaggle website "Healthcare dataset stroke data"[8]. Categorical features, numerical features and multicollinearity analysis will be carried on for better understanding of the data. Five different models - SVM, Decision tree, Random Forest, K-nearest neighbor, Logistic regression are considered. Finally, better performing algorithm will be chose to predict stroke and a simple Graphical User Interface is created using tkinter.

4. METHODOLOGY

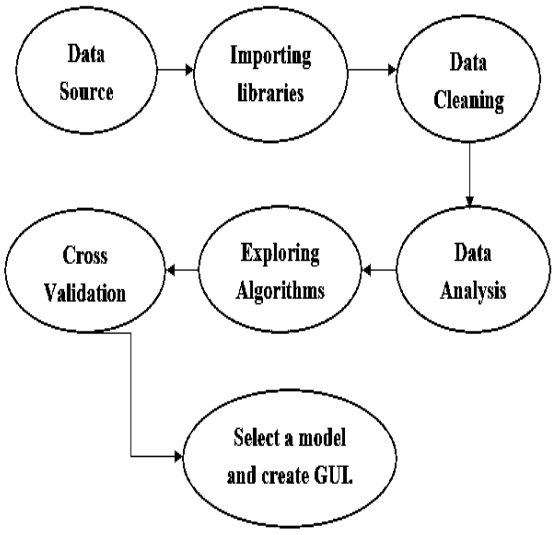

Figure 1: Proposed Method of Stroke Prediction Using Machine Learning Algorithms.

4.1 Data source

4.1.1 Primary data

Collected reference materials and common features/attributes by consulting neurologists and cardiologists.

4.1.2 Secondary data

Obtained from Kaggle website. “Healthcare-dataset-stroke data” which is publicly available. Totally consists of 5110 entries, out of which 2995 were female and 2115 were male. Consists of 12 features.

4.2 Libraries imported

4.2.1 NumPy

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

4.2.2 Pandas

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

4.2.3 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.

4.2.4 Seaborn

It is a data visualization based on matplotlib. It provides high-level interface for drawing informative and attractive graphs.

4.2.5 Tkinter

It is a simple and easy to use in-built python model used for designing user interface. Here this module is used to create 1920x1080 window with the necessary widgets for taking inputs.

4.3 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.

4.4 Data analysis

There are three types of data analysis which is performed i.e.,Categorical feature analysis, Numerical feature analysis and Multicollinearity analysis. Data analysis is done to show us the hidden relationships and attributes present in the dataset which help the machine learning model to perform better.

4.5 Implementing algorithms

Five different algorithms are selected after literature survey. Comparative study is made between these five algorithms - Decision Tree, Logistic Regression, Random Forest, Support Vector Machine and K Nearest Neighbor.

4.6 Cross validation

Effectiveness of all the models is verified to solve overfitting problems. Overall assessment on how the model will perform for an independent test dataset.

Finally, the best performing model will be used to predict stroke using the input data given by the user.

4.7 Creating GUI

The graphical user interface is a user interface which permits users to interact with digital gadgets through graphics and the audio indicator, as opposed to textual content, mostly based on user interfaces, written command labels or textual content. In response to an impressive curve of command-line interfaces that requires directions to be typed on a pc-keyboard, GUIs were provided.

4.8 Software specifications

The programming language which is used in this project is Python. Python is an extensively helpful translated, intuitive, object-orchestrated, and irrefutable level programming language. Python is used for web improvement, machine learning, AI, working systems, compact application headway, and computer games. Like Perl, Python source code is in like manner available under the GNU General Public License (GPL). This educational exercise gives adequate understanding on Python programming language. Python is expected to be extraordinarily comprehensible. It uses English expressions routinely where as various tongues use highlight, and it has less semantic improvements than various lingos.

3.8.1 Jupyter notebook

In the past it was referred as IPython Notebooks. It is an online intelligent computational platform for making Jupyter journal records. The "notebook" term can conversationally make reference to various elements, primarily the Jupyter web application, Jupyter Python web server, or Jupyter report design contingent upon setting

5. RESULTS & DISCUSSION

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.

Table 1: Accuracy for all five models.

Algorithms	Accuracy
Decision Tree	0.9113
Logistic Regression	0.955
Random Forest	0.955
Support Vector Machine	0.9243
KNN	0.9524

i. From table 1, the Logistic Regression, Random Forest, and KNN models all have a high accuracy score of 0.95. However, the error type and recall value of each model must also be considered.

ii. Models usually with an accuracy score of 0.95 have a high false negative rate. A high number of false negatives indicates a type 2 error. Identification of the subjects that have stroke and consider them free of stroke cannot be accepted in order to avoid a type 2 failure of our stroke prediction.

iii. Thus, according to the classification report, the Random Forest Model has a low false negative value and a high accuracy, which fits our goal. We have chosen Random Forest model with an accuracy of 0.955 to fit as our model.

iv. Installing tkinter is same as importing other modules in Python. Added various widgets for taking inputs.

Figure 2: Prediction of stroke where chances of stroke is high

v. Figure 2 represents a user interface window of a certain new patient whose data is given as input where it predicts that the patient has a high chance of getting stroke.

Figure 3: Prediction of stroke wherechances of stroke is less

vi. Figure 3 represents a user interface window of a certain new patient whose data is given as input where it predicts that the patient has a low chance of getting stroke.

6. CONCLUSION

- In this project, we have constructed a model for predicting stroke using machine learning algorithms. After, thoroughly reviewing various IEEE papers we selected five different models such as decision tree, random forest, support vector machine, logistic regression and K nearest neighbor. Key attributes/features were selected under the guidance of medical practitioners.
- Visualizing health data allows professionals to present key/common trends and information via graphs, charts and visuals that helps even a data analysts understand the dataset. Hence, data visualization was our main objective. Used libraries like pandas, matplotlib, seaborn and Pywaffle for informative and attractive representation of data.
- Predictive analytics is a popular business intelligence trend. They help doctors make data driven decisions in no time which can even predict and prevent deadly diseases. In this project, we have carried on categorical feature analysis, numerical feature analysis and multicollinearity successfully.
- Applied different model on the dataset. A comparative study amongst the five different models showed that random forest, logistic regression and K nearest neighbor has an accuracy of 95.5%, whereas decision tree was 91.13% accurate and support vector machine exhibited accuracy of 92.43%.
- Finally, Random Forest was chosen as the best model with high accuracy and less false negative. To facilitate seamless use of the application, a Graphical User Interface (GUI) was created using tkinter.

7. FUTURE WORK

Stroke is dependent on a lifestyle attributes as well as past medical history. Here in this paper, we have considered seven lifestyle attributes and three medical conditions. In the future, for better performance of the model more medical attributes can be considered such has Systolic blood pressure, diastolic blood pressure, pulse pressure, mean blood pressure, The min, max and mean value of a pulse. Also, mRS score, NIHSS score, CHADS2 score can be added to get a more accurate and precise output.

REFERENCES

[1] Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat and Sharmin Akter Tanna "Detection of stroke using machine learning algorithms", 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, July 2019.

[2] JoonNyung Heo , Jihoon G. Yoon , Hyungjong Park , Young Dae Kim , Hyo Suk Nam and Ji Hoe Heo. "Stroke prediction in acute stroke", Stroke. 2019;50:1263-1265, AHA Journal, 20 Mar 2019.

[3] Jaehak Yu,Damee Kim,Hongkyu Park,Seung-chul Chon,Kang Hee Cho,Sun-Jin Kim,Sungkyu Yu,Sejin Park and Seunghee "Semantic analysis of NIH stroke" , 2019 International Conference on Platform Technology and Service (PlatCon), IEEE, 30 Jan 2019.

[4] Jeena R.S and Dr.Sukesh Kumar "Stroke prediction using SVM", International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2016.

[5] Chutima Jalayondeja "Stroke risk prediction model based on demographic data" , The 2015 Biomedical Engineering International Conference (BMEiCON-2015), IEEE, 2015.

[6] Benjamin Letham,Cynthia Rudin Tyler, H. McCormick and David Madigan "An interpretable model for stroke prediction using bayesian analysis", The Annals of Applied Statistics 2015, Vol. 9, No. 3, 1350–1371, Institute of Mathematical Statistics, IEEE, 5 November 2015.

[7] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[8] https://colab.research.google.com

