Stroke Prediction Using Machine Learning Methodology

Swapnil Sanjay Aundhakar[1*], A. G. Patil[2]

[1]*M.Tech. Student, Department of Electronics and Telecommunication Engineering, Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli, India*

[2]*Professor, Department of Electronics and Telecommunication Engineering, Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli, India*

***Abstract*: Stroke is a major disease leading to death in adults and elderly people, as well as disability. Rapid detection of stroke is very difficult because the cause and cause of the onset are different for each individual. In this proposed model, we develop model stroke prediction using machine learning approach and implement a system on WHO provide dataset. The most efficient and accurate variables required to predict stroke in an individual is obtained through Feature Selection and as per the variables gained, the features which influence the disease prognosis is obtained. Predictive modelling is performed on this processed data with various classification models such as Random Forest, Decision tree, Logistic Regression and Support Vector Machines. The web application is made to process user inputs and predict the occurrence of stroke using the most accurate model.**

***Keywords*: Feature selection, disease prognosis, Random Forest, Logistic Regression, Support Vector Machine.**

1. Introduction

In particular, mortality from cardiovascular diseases and cerebrovascular diseases of the circulatory system is continuously increasing in elderly people over 60 years of age. Stroke is a fatal disease that causes dysfunction in adults and the elderly, and difficulties in social or economic activities, depending on the stroke severity [1]. Stroke, a disease with severe morbidity, disability and mortality, has become one of the major threats to public health worldwide.

Stroke can vary widely, depending on the patient’s symptoms pattern or the accompanying disease. Individuals with stroke should accurately assess the level of disability, and induce medical centre visits. Therefore, there is a desperate need for technology to support visits to medical institutions and the diagnosis and treatment of medical doctors within a short time by continuously monitoring patients with stroke.

Recent studies describe many ongoing attempts to apply National Institutes of Health Stroke Scale (NIHSS) techniques to a new solution to prevent the recurrence of stroke patients and adaptively evaluate the initial disease disorders, in order to cope with major risk factors for stroke.

In another methodology, we found major risk factors for stroke through previous studies and clinical trials. These risk factors include smoking, hypertension, diabetes, and obesity. It has been reported that the risk of stroke is caused by interaction of various risk factors, rather than by one factor. Therefore, a new methodology is emerging for assessing the risk factors of each individual stroke, and for prediction or early detection of the disease. Recently, studies on the prediction of stroke disease using various.

2. Literature Review

E. C. Jauch, J. L. Saver, H. P. Adams et al [1] The authors present an overview of the current evidence and management recommendations for evaluation and treatment of adults with acute ischemic stroke. The intended audiences are pre-hospital care providers, physicians, allied health professionals, and hospital administrators responsible for the care of acute ischemic stroke patients within the first 48 hours from stroke onset. These guidelines supersede the prior 2007 guidelines and 2009 updates.

Singh M. S. & Choudhary P. (2017) [6]. Stroke prediction using artificial intelligence, the aim is to take a Medical decision which is a highly specialized and challenging job due to various factors, especially in the case of diseases that show similar symptoms, or regarding rare diseases. It is a major topic of Artificial Intelligence (AI) in medicine. An AI system would take the patients data and propose a set of appropriate Prediction. The system can extract hidden knowledge from a historical clinical database and can predict patients with disease and use the medical profiles such as Age, Blood Pressure, Blood Sugar, etc. it can predict the likelihood of patients getting a disease.

Method by T. P. Hong, C. W. Wu [5] presented the work on most of the existing researches about stroke prediction are concerned with the complete and class balance dataset, but few medical datasets can strictly meet such requirements. For the incomplete data, a missing value imputation method based on iterative mechanism has shown acceptable prediction accuracy.

[*] Corresponding author: swapnil.9684@gmail.com

3. Proposed Research Work

To initiate with the work, we have started collecting data in each and every aspect towards the goal of the system. In the first place, the research is in the direction of the main causes or the factors which have strong influence on the heart health. Some factors are unmodifiable like age, sex and family background but there are some parameters like blood pressure, heart rate etc. which can be kept in control by following certain measures.

Many doctors suggest healthy diet and regular exercise to keep the heart healthy. Following are the parameters which are considered for the study in designing the system which have major risk percentage

1. Age
2. Sex
3. Blood Pressure
4. Heart Rate
5. Diabetes
6. Hyper cholesterol

A. Dataset Description

The data collection and analysis based on patient characteristics can be useful for our system user since any normal user can easily provide his/her general data [4].

Complicated clinical data information such as Magnetic Resource Image scans and images are not supported by this prediction system and is created for the user to test oneself on the go. The dataset as given in figure-1 offers 12 variables consisting of patient id, gender, hypertension, age, heart disease rate, marriage status, occupation, type of residence, glucose level, patient BMI, smoking status and stroke yes/no prognosis [4].

B. Attribute Information

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking status means that the information is unavailable for this patient

4. Proposed Methodology

Following are steps involved in building the model.

1. Analyse the business Problem.
2. Data collection: Defining the problem and assembling a dataset.
3. Data preparation: Preparing your data.
4. Splitting of data
5. Choose model
6. Train model: Developing a model that does better than a baseline.
7. Evaluate model: Choosing a measure of success, Deciding on an evaluation protocol.
8. Parameter tuning - Scaling up: developing a model that overfits, regularizing your model and tuning your parameters.
9. Prediction.

The data analysis is carried out on the dataset in different phases. Various functions and libraries in python are used for this purpose. The respective phases and methods followed for this research are listed and explained in detail.

A. Pre-Processing

Pre-processing is the initial step in conducting the analysis. This module provides elimination of unwanted values as well as cleaning the dataset. A dataset can be loaded with Not Applicable (NA) values or even empty values. These values need to be removed by removing the whole row which contains such a value. Zero values also can be neglected and the dataset needs to be rid of these values. This is where pre-processing comes in. This step is important mainly because the rest of the modules need to work upon cleaned data for proper analysis [7].

B. Exploratory Data Analysis

When cleaned data is obtained, data exploration must be done in order to get insight from it. Analysis of the variables is done in this phase in order to get valuable information about the explanatory variables and their relation with the response variable. Yes/no prognosis of stroke and depression act as the response variable in both datasets for this process. The aim of this phase is to display and get information on how the other variables are influencing and relating with our response variable through plots and visualisation. The different variables of stroke dataset such as age, BMI, glucose levels, gender and so on are checked for their relation with the stroke prognosis variable [7].

C. Feature Selection

Feature selection is the process in which we select the most important features out of all the independent variables. The features are selected for our study based on two feature selection methods.

D. Predictive Modelling

This phase follows the feature selection methods by which the best nine features - gender, hypertension, age, heart disease rate, and marriage status, type of work, glucose level, patient BMI, and smoking status are selected. The training and the test dataset consists of nine features each. The test dataset does not contain the outcome variable i.e. stroke variable and the main aim of this phase is to predict this variable by fitting classification models into the training dataset. The classification models used for this study include Random Forest, Decision Tree, Logistic Regression and Support Vector

Machines. The accuracy of the models are checked individually and the most accurate model is chosen for constructing the stroke prediction web application [7].

E. Algorithm Flowchart

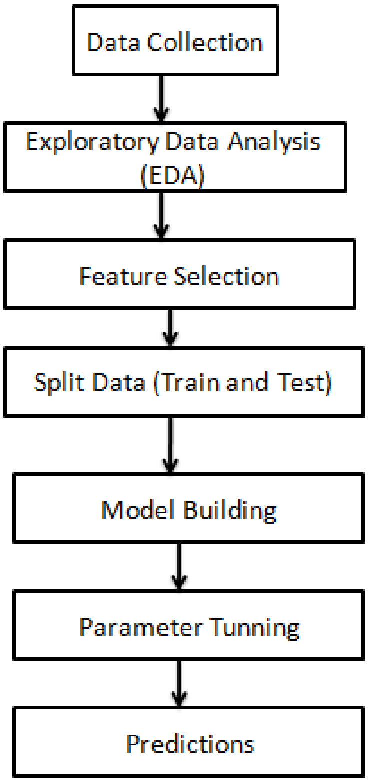

Fig. 1. Algorithm flowchart

5. Conclusion

This paper presented an overview on stroke prediction using machine learning methodology.

References

[1] C. Jauch, J. L. Saver, H. P. Adams, A. Bruno, B. Connors, B. M. Demaerschalk, P. Khatri, P. W. McMullan, A. I. Qureshi, K. Rosenfield, P. A. Scoot, D. R. Summers, D. Z. Wang, M. Wintermark, and H. Yonas, “Guidelines for the Early Management of Patients with Acute Ischemic Stroke,” American Heart Association, Vol. 44, No. 3, pp. 870-947, 2013.

[2] S. Wang, J. Zhang, and W. Lu, “Sample size calculation for the proportional hazards model with a time-dependent covariate,” Computational Statistics & Data Analysis, Vol. 74, pp. 217-227, 2014.

[3] Benjamin Letham. Cynthia Rudin. Tyler H. McCormick. David Madigan. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model." Ann. Appl. Stat. 9 (3) 1350 - 1371, September 2015.

[4] WHO DATASET, According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

[5] T.-P. Hong, C. W. Wu, Mining rules from an incomplete dataset with a high missing rate, Expert Systems with Applications 38 (4) (2011).

[6] Singh, M. S., & Choudhary, P. (2017). Stroke prediction using artificial intelligence. 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON).

[7] Soodamani Ashokan, Suriya G. S Narayanan, Mandresh S, Vidhyasagar BS, Paavai Anand G, An Effective Stroke Prediction System using Predictive Models, IRJET, 2020.

[8] H. J. Lee, J. S. Lee, J. C. Choi, Y. J. Cho, B. J. Kim, H. J. Bae, D. E. Kim, W. S. Ryu, J. K. Cha, D. H. Kim, H. W. Nah, K. H. Choi, J. T. Kim, M. S. Park, J. H. Hong, S. I. Sohn, K. S. Kang, J. M. Park, W. J. Kim, J. Lee, D. I. Shin, M. J. Yeo, K. B. Lee, J. G. Kim, S. J. Lee, B. C. Lee, M. S. Oh, K. H. Yu, T. H. Park, J. Y. Lee, and K. S. Hong, “Simple Estimates of Symptomatic Intracranial Hemorrhage Risk and Outcome after Intravenous Thrombolysis Using Age and Stroke Severity,” Journal of Stroke, Vol. 19, No. 2, pp. 229-231, 2017.

[9] V. Vijayaganth, P. Purusothaman, M. Krishnamoorthi, ‘A Comprehensive Survey on Security Challenges and Techniques in Big Data’, International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 06, 2020.

[10] Aishwarya Roy, Anwesh Kumar, Navin Kumar Singh, Shashank D, ‘Stroke prediction using decision trees in artificial intelligence’ International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4, Issue 2, 2018.