iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52137

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Brain Stroke Predictive Analysis using Various Machine Learning Algorithms

Steffi Niveditha[1], Suresh[2], Yengunti Purna Chandra Rao4, Yarangalli Ganesh[3], Jangama Manjunath[5]
[1, 2, 3, 4, 5]*Ballari Institute of Technology and Managemenet, Ballari*

Abstract: ***Our project offers a reliable stroke detection technique. Data analysis and the creation of prediction applications are both done using the Pythonlanguage. The main goal of this project is to enable anyuser to perform a stroke self-test. The data on thepatients' characteristics is used to create the GUI interface and data analysis components. We collect the stroke training and test datasets and do exploratory data analysis. Through feature selection, the most effective and precise variables needed to predict strokein an individual are obtained, and according to thevariables obtained, the features that affect the prognosis of the disease are obtained. On this processed data, predictive modeling is done using different categorization models, including Logistic Regression, Support Vector Machine, Decision Treemodel and Random Forest algoritham. The bestaccurate model is used by the GUI interface to handle user inputs and forecast the incidence of stroke.***

Keyword: Logistic Regression, Support Vector Machine (SVM),Random forest algoritham and Decision Tree(Keyword)

I. INTRODUCTION

Neurological illnesses may caused due to the damage in CNS and (PNS) peripheral nervous system. While some of these disorders are curable and treatable, others are not treatable. Age is a important risk factor to getting this disease because of age people experience asfew neurological problems such as Alzheimer's disease and the Parkinson's disease primarily after. The causes might range from genetic abnormalities, infections, lifestyle factors, and other health issues that could have an impact on the brain. More than 600 different disorders of the neurological system exist, including epilepsy, brain tumors, and stroke. Every year, all over world 15 million peoples suffering from a stroke. The main focus of this research is the lack of effective methods for patients to determine whether they might have such disorders. For any user to test and become aware of their medical state, they urgently require a simple-to-use disease prediction application that operates on properly analyzed data. This may facilitate receiving the required medical care from a hospital. The application is helpful in educating users about the importance of leading healthy lives. In the 2016–17 academic year, 1,654,577 individuals were hospitalized with neurological illnesses. Among the neurological disorders are Autism, Dementia, Epilepsy, and others. Men are found to get a stroke at a younger age than women, while women die from stroke-related causes more frequently. The dataset utilized in this study includes records of people who had theirstroke risk assessed and includes information on their gender,age, lifestyle factors, BMI, demographic regions, and other factors.

II. LITERATURE REVIEW

Tasfia Ismail Shoily et al. has made a different comparisons on different models as like as Naive Bayes, J48, k-NN and Random Forest and found that Naive Bayes has as given better precision as compared to the other realated models. Various medical documents orreports are visulaized to obtain the dataset and those datasets was cross-verified by medical experts and used with WEKA (Waikato Environment for Knowledge Analysis).That the developed model will assist patients to be careful and check whether they can have stroke or not. 4 different trained models sre used such as Naive Bayes, J48, k-NN, and Random Forest. Precision and accuracy were very important aspects to validate the models.The dataset is applied to the machine learning models to get the desired output [1].

JoonNyung Heo et al., taken three machine learning classification models which were considered based on specific parameters which are related to deep neural network, random forest algoritham, and logistic regression algoritham to predict a person having a stroke or not. By studying above paper, we came to know that Deep neural network (DNN) is majorly and efficiently used for ischemic stroke or acute stroke patients which also has an drawback for long term prediction.With the DNN model we get has an accuracy of 88% with respect to the given inputs which was better as it is compared with other models. Automated and more precise calculations are done to improve the model and to get a higher accuracy, decreasing use of simpler models [2].

Jaehak Yu et al. prefered the C4.5 algorithm it is based on decision tree algorithm, which again takes the user's real- time attributes and classifies the stroke based on that dependent attributes, which is divided into four different classes. This collected data helps in identifying the timing of the stroke and the uncertanities that may occurs, which in turn helps in taking necessary precautionary measures and other medical checkups. Using naive bias, we getting an accuracy of 85.4% and by using random forest, we obtain an accuracy of 88.9% [3].

Jeena R.S. and Dr. Sukesh Kumar used on classification models such as SVM with preferred supported kernel functions used for stroke analysis and detection. To remove redundant and incompatible data, preprocessing should be performed. About 350 data inputs were used formaking a prediction. It was run on a platform like MATLAB, resulting in an accuracy such as 91% [4].

Chutima Jalayondeja did the prediction has been done with the help of demographic data and decision tree model, Naïve Bayes classifier and artificial Neural Network are the 3 algorithms considered and by using decision tree data model we get highest accuracy and having a low FP (false positive), by low FP rate means high accuracy in predicting wheather the patients had stroke but not stroke, while FN (false negative) says no stroke but the patients actually had stroke. The FN is the one of the dangerous as it leads to cause mortality since the patient has stroke but predicts the opposite and sends the output the person don't have a stroke. From the accuracy point of view of, the decision tree model was considered, but in terms of safety and security, the artificial neural network must be taken and considered. The neural network was chosen because it has a high FP value and a low FN value [5].

III. PROBLEM STATEMENT

To Design and implement a brain stroke predictive analysis system using various algorithms such (SVM) Support Vector Machine, Decision Tree, Logistic Regression and Random Forest.

.

IV. EXISTING SYSTEM

We present all the notable studies on stroke disease in this part based on the survey that was completed. The field of stroke prediction has seen significant development in both therapy and imaging. Electrical cell simulation, a technology that helps to boost the brain's diminished blood flow, is one that appears to have promise. Along with the previously mentioned technologies, significant scientific advancements have made it possible for patients to wearable smart devices that can track their heart rate, blood pressure, and other bodily indicators to determine their risk of stroke. Utilizing the expanding information technology fields of data science, machine learning, and artificial intelligence, systematic research and studies havebeen conducted. Those studies and activities have had a significant impact on the field of medical science.

V. PROPOSED SYSTEM

Nowadays, a lot of people pass away too soon due to the effect of strokes. In today's world people seen the development of numerous methods and systems that use medical data analytics to forecast stroke

Symptoms to identify the stroke. These systems analyses a person's medical history with the help of machine learning algorithms such as prediction algorithm's to predict a stroke and allow a user to do self test to save lives.

These systems begin by pre-processing the dataset to eliminate negative and missing values or to improve categorical data before supplying the data to machine learning algorithms. Gender, age, hypertension, heart disease, BMI, blood sugar level, and smoking habits are major characteristics of a dataset used to predict stroke. Some predictive ML algorithms like Logistic Regression algoritham, Decision Tree model, Random Forest Classifier and support vector machine (SVM) are used to create models and makepredictions accurately.

VI. OBJECTIVES

1) To enable any user to perform a stroke self- test.
2) To analyse user inputs and forecast the likelihood of a stroke using the most precise model.
3) To training and test the datasets and do exploratory data analysis.
4) To predict the brain stroke or not of a person.

.

VII. METHODOLOGY

A. Support Vector Machine (svm) Algorithm

Support Vector Machine Algorithm is on of the supervised ML algoritham. It is useful for solving both regression and classification problem statement.

Now let us consider the classification problem to just understand the Geometrical intuition since it is a classification problem here we can easily separable two classes points. We can separable these points with hyper plane. SVM make sure that when we are creating hyper path plane apart from that it also creates two margin lines. These two margin lines obtain some distance so that it will be linearly separable for both the classification points having some distances. SVM is also used for Regression problems to maintain the main features that the algorithm is characterize. Regression problem is same as the classification problem with only having minor changes.

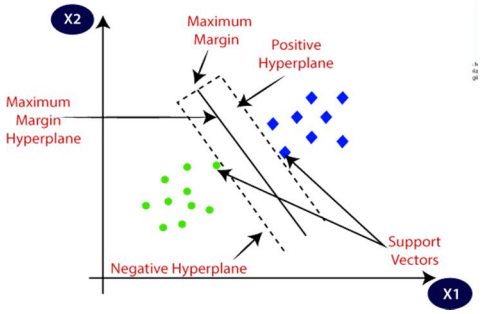


Fig.1. Support Vector Machine Algorithm

B. Logistic Regression Algorithm

Logistic Regression is one of the simple and commonly used for Machine Learning algorithms for two-classes classification. It is one of the statistical approach to predicts the probability of a binary event utilizing a logic function. Binary event means it will identify person is having disease or not. Logistic Regression is regression model or statistical approach to predict probability on dependent on a certain attribute for given data entry belongs the category number one just like linear regression using a certain mathematical functions. Using sigmoid function logistic regression models the data and it provide constant output. It is used to predict the probability of dependent variable. These dependent variable has only two necessary classes. In dependent variable data is coded as either 1 or 0.

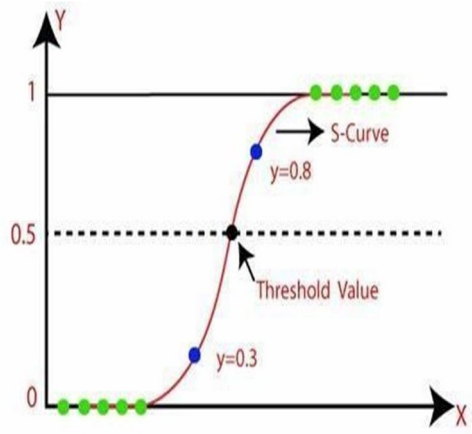


Fig.2. Logistic Regression Algorithm

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

C. Random Forest Algorithm

Random Forest Algorithm is a collection of multiple random decision tree and it is much less sensitive to the training data. In the creation Random Forest, the first step is to build new dataset from the original data. The process of creating new data is called boot strapping. Then we willrandomly select features for each tree which are used for training. After the construction of all the decision trees, a new data point is passed to all the trees. The next step is combining all the trees. As it a classification problem, we will take the majority voting. This process of combining results of all the decision tree is taken which is called aggregation. In Random Forest, we first perform bootstrapping and then aggregation.

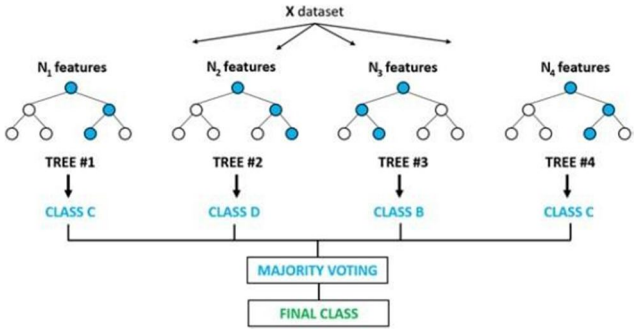


Fig. 3. Random Forest Algorithm

D. Decision Tree Algorithm

This represents the given data into a tree like structure in ahierarchal structure. That structure having a N number of nodes which is also called as attributes or independent variables , there are some directed links or edges between an attribute so that it establishes a relation between all the entities available in that dataset. A data set divided into large number of rows and columns where last column is referred as a class and the rows or tuples are referred as instance and other columns are called as attributes or features. Decision tree has a primary node or a root node which has no incoming edges and it has two or more number of out-going edges. At the end of the tree there areleaf nodes or terminal nodes which has incoming edges but no out-going edges. And there are some nodes in between root and leaf nodes which are called as internal nodes it has exactly one incoming node and several out- going nodes. Edges consists of conditions known as attribute test conditions. This algorithm is represented in the form of a linked list in memory map. This tree is traversed in the left to right format.

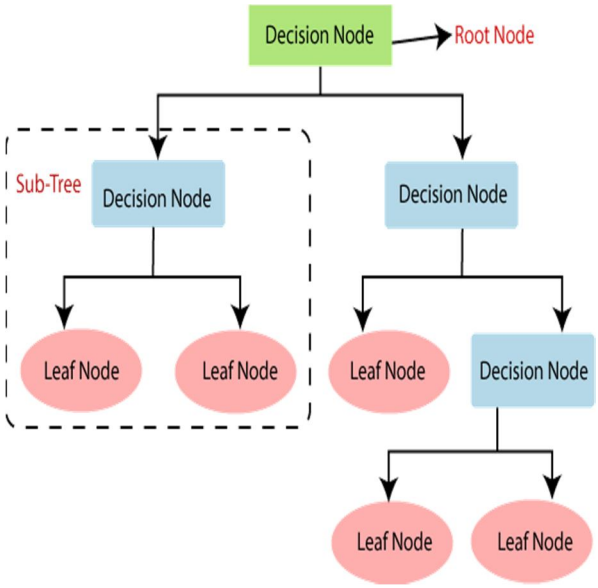


Fig. 4. Decision Tree Algorithm

VIII. EXPECTED OUTPUT

The expected output of given project which will determine the stroke of a person self-test by the user it self based on some parameters which are independent variable such as marital status, hypertension, age, gender, heart disease, work type, smoking status by taking these Boolean values from the user we predicted the the person having a stroke or not those are dependent variables . if the person havinga stroke it is indicated by Boolean 1 and not having a stroke indicated by the Boolean 0.

IX. CONCLUSION

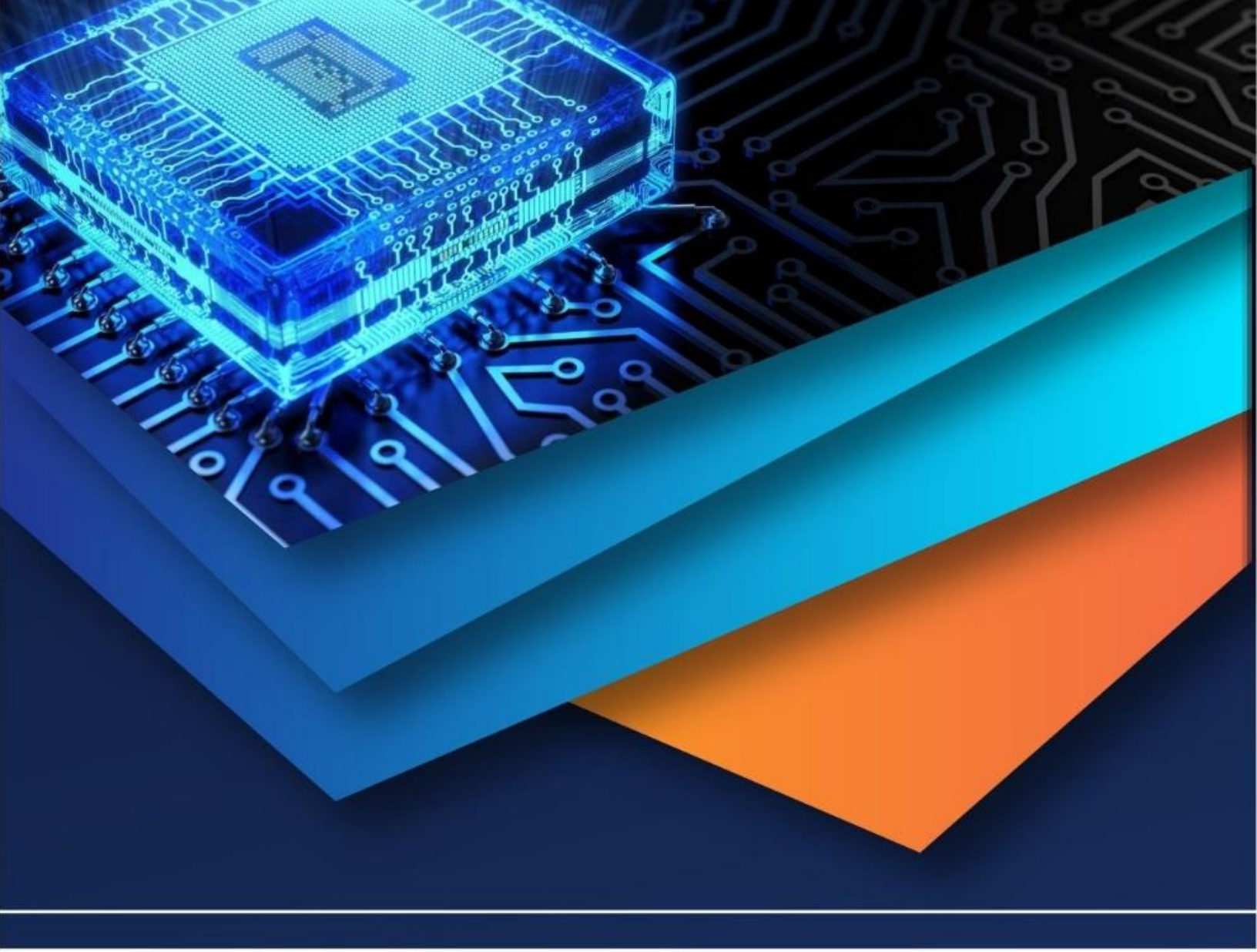
After the literature survey, we came to know various advantages and disadvantages of different research papers and thus, proposed a system that helps to predict brain stroke self test in a cost effective manner and efficient wayby taking few inputs from the user side and predicting accurate results with the use of trained Machine Learning algorithms.

X. FUTURE ENHANCEMENT

Stroke depends on both lifestyle and medical history of that self tested patient. In this paper, we have taken some important lifestyle characteristics or some independent attributes and some related medical issues or conditions. In the future, more medical independent attributes can be considered for better performance and getting higher accuracy of the model, such as systolic bp, diastolic bp , pulse pressure, mean or average blood pressure, minimum, maximum, and mean pulse. Also mRS score and NIHSS score and CHADS2 score can be added to get more accurate and precise result on any model.

REFERENCES

[1] Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat and Sharmin Akter Tanna "Detection of stroke using machine learningalgorithms",10thInternationalConferenceonCompu ting,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 strokel , 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] Ane Alberdi, Alyssa Weakley et al., ―Smart home- based prediction of multi-domain symptoms related to Alzheimer's Disease,‖ IEEE Journal of Biomedical and Health Informatics, 2018.
[7] Yonglai Zhang, Wenai Song et al., ―Risk Detection of Stroke Using a Feature Selection and Classification Method,‖ IEEE Access, vol. 6, pp. 31899-31907, 2018.
[8] Farrikh Alzami, Juan tang et al.,‖ Adaptive hybrid feature selection-based classifier ensemble for epileptic seizure classification,‖ IEEE Access, vol. 6, pp. 29132- 29145, 2018
[9] Wanchat Threeanaew, James MacDonald et al., ―Collection and Analysis of Multimodal Data for SUDEP Biomarker Discovery,‖IEEE Sensors Letters, vol.3, no.1, 2019.
[10] Pholpat Durongbhan, Yifan Zhao et al., ―A Dementia Classification Framework using Frequency and Timefrequency Features based on EEG signals,‖ IEEE Transactions on Neural Systems and Rehabilitation Engineering, pp. 1-10, 2018.
[11] Mohamed Mahyoub, Martin Randles et al.,―EffectiveUse of Data Science Toward Early Prediction of Alzheimer's Disease,‖ IEEE 20th International Conference on High Performance Computing Communications,IEEE 16th International Conference on Smart City, IEEE 4th Intl. Conference on Data Science and Systems, pp. 1455- 1461, 2018.
[12] 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.
[13] https://www.kaggle.com/fedesoriano/stroke- prediction-dataset
[14] Chiu, I.-M., Zeng, W.-H., & Lin, C.-H. R. (2020). Using multiclass machine earning model to improve outcome prediction of acute ischemic stroke patients after reperfusion therapy. 2020 International Computer Symposium (ICS). doi:10.1109/ics51289.2020.00053
[15] Fang, G., Xu, P., & Liu, W. (2020). Automated Ischemic Stroke Subtyping Based on Machine Learning Approach. IEEE Access, 8,118426– 118432.doi:10.1109/access.2020.3004977



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY