International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2283121

Stroke Risk Prediction Using Machine Learning Algorithms

Rishabh Gurjar[1], Sahana H K[1], Neelambika C[1], Sparsha B Sathish[1], Ramys S[2]

[1]Department of Computer Science and Engineering. The National Institute of Engineering, Mysore, Karnataka, India

[2]Assistant Professor, Department of Computer Science and Engineering. The National Institute ofEngineering, Mysore, Karnataka, India

Article Info
Publication Issue :
Volume 8, Issue 4
July-August-2022

Page Number : 20-25

Article History
Accepted: 20 June 2022
Published: 05 July 2022

ABSTRACT

The majority of strokes are brought on by unforeseen obstruction of pathways by the heart and brain. Distinct classifiers have been developed for early detection of different stroke warning symptoms, including Logistics Regression, Decision Tree, KNN, Random Forest, and Naïve Bayes. Furthermore, the proposed research has obtained an accuracy of around 95.4%, with the Random Forest outperforming the other classifiers. This model has the highest stroke prediction accuracy. Therefore, Random Forest is almost the perfect classifier for foretelling stroke, which doctors and patients can utilise to prescribe and identify likely strokes early. Here in our research we have created a website to which model is dumped/loaded, such that the interface will be friendly to the end-users.

Keywords: Stroke, Machine Learning, Data Analysis, Normalization, Scalarization, ML Algorithms, Accuracy, Results.

I. INTRODUCTION

The cells in regions of the brain are deprived of nutrients and oxygen and start to die when blood flow to those regions is interrupted or diminished. A stroke is a medical emergency that requires prompt medical care. Early detection and appropriate treatment are required to prevent further harm to the damaged area of the brain and associated consequences in other body areas. According to the WHO (World Health Organization), 15 million people worldwide suffer from stroke every year, with one person dying every 200-300 seconds.

Strokes come in two different varieties: ischemic and hemorrhagic. In the event of an isocheimal stroke, clots prevent drainage, whereas in the event of a hemorrhagic stroke, a weak blood vessel bursts and causes bleeding into the brain. A healthy/balanced lifestyle, like quitting drinking and smoking, regulating BMI (body mass index) and average level of glucose, and keeping kidney and good heart function, can help prevent stroke. Predicting strokes is crucial, because they must be treated to prevent permanent harm or death. The criteria utilised in this investigation to predict stroke included hypertension, BMI, heart disease, and average blood glucose levels. Furthermore, machine learning can play an important part in the suggested prediction system's decision-making processes [1]– [3].

Only a few researches that have been published [4]–[9] have employed machine learning algorithms to predict stroke. Examples of machine learning techniques include artificial neural networks (ANN), stochastic gradient descent, the c4.5 decision tree algorithm, k-nearest neighbour (KNN), principal component analysis (PCA), convolutional neural networks (CNN), Naive Bayes, and others. Stroke has been associated with hypertension, BMI, average glucose, and cardiac disease [10].
The following is our contribution to this paper:
•The Random Forest model is created for stroke prediction using diseases/attributes such as age, smoking status, BMI, heart disease, average glucose level, and hypertension.
•The proposed Random Forest algorithm's performance is contrasted with that of cutting-edge classifiers like Logistics Regression (LR), K-Nearest Neighbor's (KNN), Decision Tree, and Naive Bayes.

The rest of the document is structured as follows. The second section reviews some previous study literature. The three components of section 3's description of the study methodologies are the data description, machine learning classifiers and assessment metrics, and implementation procedures. Section 4 describes the webpage. The correlation finding and performance analysis are thoroughly discussed in section 5, where the results and discussion are also presented. Section 6 then examines the conclusion.

II. Literature Survey

In the past, numerous scholars have used machine learning-based methods to predict strokes. To categorize stroke disorders, Govindarajan used data from 507 patients and a text mining and machine learning classifier combination. The SGD algorithm, which had a value of 95%, offered the best value out of all the machine learning techniques they utilised for training with ANN for their analysis.

Researchers including Amini looked at stroke prediction. Incidence investigated 807 healthy and ill individuals and categorised 50 risk factors for stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia, and alcohol use. The c4.5 decision tree algorithm and the K-nearest neighbour algorithm were the two techniques they used that had the highest accuracy, at 95% and 94%, respectively.

In order to forecast the mortality of stroke patients, Cheon conducted a study. In order to determine the frequency of strokes, they included 15099 people in their study. They used a deep neural network approach to identify strokes. PCA was used by the authors to extract medical record history and forecast stroke. Their area under the curve has a value of 83 percent (AUC).

To determine whether an automated early ischemic stroke detection system may be used, Chin conducted a study. Their main objective was to develop a system that would use CNN to automate primary ischemic stroke diagnosis. They collected 256 images for use in training and testing the CNN model. In order to remove the impossibly small area of stroke, they used the data prolongation strategy to elevate the collected image in their system image preparation. 90 percent of the time, their CNN method is accurate.

A stroke severity index was developed through research by Sung, 3577 patients who had experienced an acute ischemic stroke were the subject of their data collection. Their prediction models were developed using linear regression and data mining approaches. For their prediction feature, the k-nearest neighbours model delivered the best results (95 percent CI).

The functional outcome of an ischemic stroke was predicted by with the aid of machine learning. They applied this technique to a patient who passed away three months after being hospitalised for their investigation. More than 90% of the AUC was attained by them.

III. Research Methodology

This part is divided into three sub-sections: data description; machine learning classifiers and evaluation matrices; and implementation techniques. The three procedures are as follows:
A. Data Description:
The information used in this paper was obtained from a Kaggle as a Dataset. It's a document containing the

information of 5110 persons, and then all their parameters are listed :

age: This parameter describes the age of person. It's numerical data.
gender: This parameter conveys a person's gender. Here the data is categorical.
hypertension: This parameter describes that the person is having hypertension. Here the data is numerical.
work type: This parameter describes the person's profession type. This is categorical data.
residence type: This parameter represents the geographical area of a person. This is categorical data.
heart disease: This attribute describes whether the person has a heart disease or not. It's numerical data.
avg glucose level: This parameter represents the reading of glucose level from a person's blood. This data is numerical in nature.
bmi: This parameter describes the BMI (body mass index) of a person. This data is also numerical in nature.
ever married: This parameter describes marital status of a person. The data here is categorical.
smoking Status: This parameter describes the smoking habit of a person. The data here is also a categorical.
stroke: This parameter is the one which is to be predicted. The Model gives the output in the form of numerical data.

B. Machine Learning Algorithms & Its Evaluation:
The five machine learning algorithms that were used in this study to develop stroke predictors are covered in this section. The algorithms on this list are as follows: Logistic Regression is the first method, followed by Decision Tree, K- Nearest Neighbors (KNN), Random Forests, and Naive Bayes. These algorithms were selected because they are well-known in the field of developing vulnerability predictors and have been used in a number of related studies. To create vulnerability predictors in our model, these five algorithms were selected; they are well-known algorithms that have been employed in similar research work. And finally, the metrics are evaluated for each classifier.

C. Implementation Procedure:

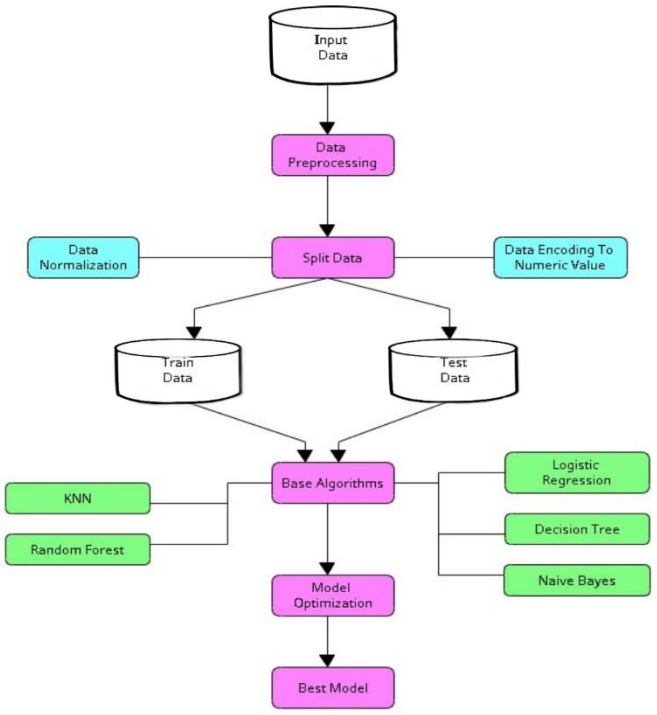

Fig. 1. Procedure for predicting stroke.

The process of implementation is described in this section. The Python and Sklearn (Scikit-learn) packages were used to finish the investigation, and Figure displays the complete strategy (Figure 1).

1) Input Data: Based on their various medical conditions, which may entail the probability of stroke, the 5110 patient's information was gathered. Kaggle was used to collect the data. The 1964 Helsinki declaration and its amendments or equivalent ethical standards were followed in all procedures utilised in studies involving human participants, as well as those mandated by the institutional and/or national research committee.
2) Data prepossessing: Before processing input, it checks for missing values and duplicate values. The other variables' means/medians were utilised to fill in any gaps in the data. There are several missing values for the smoking status parameter. These blank values can be filled in using the group by age property. Our dataset contains no values that are duplicates. As a result, it transformed our categorical data into a normalised data set with label encoding. After that, the data set will be discovered as a number value. Finally, the standard data set is acquired for additional processing.

3) Split Data: A dataset is split when it is split into training and testing groups. For training and testing, this research uses a split method. A dataset is split when it is split into training and testing groups. This paper uses a split strategy for training and testing.
4) Base Algorithms: As a basis algorithm, five methods are utilised to train and test the suggested method.
5) Model Optimization: The accuracy of each model is measured in this approach to determine the effectiveness of various types of algorithms and to select the best model.
6) Best Model: In this step, a Model is constructed utilising the specific machine learning algorithm based on the greatest accuracy found.

IV. Website

Based on the best model the website is developed. The Backend of website is designed using Flask which is a frame-work of python. Similarly, the front-end is developed using HTML, CSS and BOOTSTRAP.

The Best Fitted Model is dumped/loaded into flask frame-work i.e. in the back- end. Before the loading of model, the scaler file is dumped in the backend. This Scaler file is used further to transform the input values. And then prediction algorithm is designed.

The Finally Designed back-end is further merged with the HTML files which are designed using CSS and Bootstrap.

The Website is hosted using the cross-platform programme ngork, which makes local server ports accessible over the Internet.

Any user of the website who has knowledge of the diagnostic is welcome to utilise it.

V. Result

A. Correlation Results:
The Pearson connection's consequences reveal the impact of feature qualities on the target attribute. Figure 2 shows the connection between the others attribute and the stroke attribute. The graph shows that no single parameter significantly affects stroke. Among the factors that have a big impact on stroke risk are gender, age, hypertension, heart disease, average blood sugar, body mass index, and smoking status. The least significant variables are work type, residence type, and ever married.

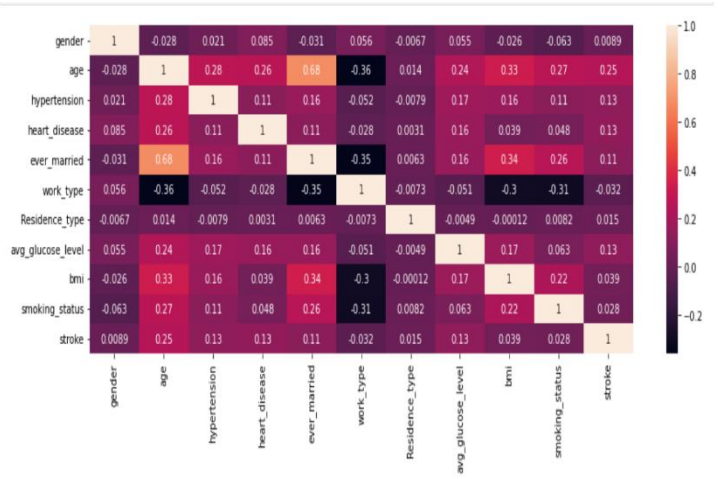

Fig. 2. Matrix correlations between sociodemographics, lifestyle, and disease.

B. Performance Evaluation:
The test dataset that was utilised to evaluate the efficacy of the machine learning approach for classifying the data will be covered in this section. A total of 1022 rows were used for testing purposes out of a total of 5110 rows of data set.
To evaluate the effectiveness of stroke prediction, Table I shows confusion matrices employing five different classifiers, namely Decision Tree, Logistic Regression, KNN, Random Forest, and Naïve Bayes.

Classifier Name	Predicted→ Actual↓	No Stroke	Stroke
Decision Tree	**No Stroke**	863	109
	Stroke	33	17
Logistic Regression	**No Stroke**	899	63
	Stroke	22	28
KNN	**No Stroke**	910	62
	Stroke	24	26
Random Forest	**No Stroke**	945	27
	Stroke	19	21
Naïve Bayes	**No Stroke**	868	104
	Stroke	17	23

TABLE I CLASSIFIERS FOR MACHINE LEARNING TO PREDICT STROKE USING CONFUSION MATRIALS

Figure 3 demonstrates how ML classifiers employ a range of current techniques to forecast stroke. As a result, the proposed study's accuracy is compared to that of a few state-of-the-art methodologies, and it is found to have a 95.6 percent accuracy.

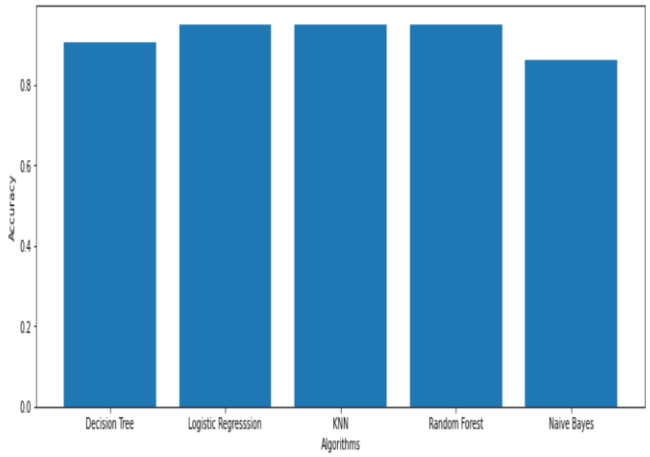

Fig 3. Accuracy of model with respect to various Algorithms.

VI. Conclusion

To determine the performance of a person's stroke occurrence, the proposed research effort used five classifiers. The proposed Random Forest classifier used gender, age, hypertension, heart disease, average glucose level, BMI, and smoking status as feature parameters to predict stroke. In comparison to the regularly used other machine learning algorithms, Random Forest delivered the highest accuracy of roughly 95%, according to the performance evaluation. As a result, the Random Forest model is used to predict stroke.

VII. REFERENCES

[1]. M. Mahmud and colleagues, "A brain-inspired trust management model to provide security in a cloud-based IoT framework for neuroscience applications," Cognitive Computation, vol. 10, no. 5, pp. 864-873, 2018.

[2]. "Application of deep learning in diagnosing neurological illnesses from magnetic resonance images: a survey on the identification of Alzheimer's disease, Parkinson's disease, and schizophrenia," Brain Informatics, vol. 7, no. 1, 2020, pp. 1–21.

[3]. A. Hussain, M. S. Kaiser, and M. Mahmud, "Deep learning in mining biological data," arXiv preprint arXiv:2003.00108, 2020.

[4]. L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, "Prediction and control of stroke by data mining," International Journal of Preventive Medicine, vol. 4, no. Suppl 2, May 2013, pp. S245-249.

[5]. S. F. Sung, C Y Hsieh, Y H Kao Yang, H J Lin, and C H Chen Using data mining methods, it was possible to develop a stroke severity index based on administrative data in November 2015, according to Journal of Clinical Epidemiology, vol. 68, no. 11.

[6]. Low cost and portable patient monitoring system for e-health services in Bangladesh, 2016 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-4. M. C. Paul, S. Sarkar, M. M. Rahman, S. M. Reza, and M. S. Kaiser.

[7]. "Innovative technique in web application effort and cost estimation utilising functional measurement type," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE (2015), pages 1–7.

[8]. "Toward a heterogeneous mist, fog, and cloud-based framework for the internet of healthcare things," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4049-4062, 2018. M. Asif-Ur-Rahman, F. Afsana, M. Mahmud, M. S. Kaiser, M.

[9]. R. Ahmed, O. Kaiwartya, and A. James-Taylor.

[10]. Tamara Islam Meghala, Md., Minazuddin Emon, and Maria Sultana Keya Performance Analysis of Machine Learning Approaches in Stroke Prediction by Mahfujur Rahman in the Fourth International Conference on Electronics,

Communication, and Aerospace Technology (ICECA-2020).

[11]. Applications of deep learning and reinforcement learning to biological data, IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 6, pp. 2063–2079, 2018. [10] M. Mahmud, M. S. Kaiser, A. Hussain, and S. Vassanelli.

[12]. "Classification of stroke disease using machine learning techniques," Neural Computing and Applications, vol. 32, no. 3, Feb. 2020, pp. 817-828. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan.

[13]. In the 2014 International Conference on Electrical Engineering and Information & Communication Technology, S. M. Reza, M. M. Rahman, and S. Al Mamun presented "A novel concept for road networks-a car xml device partnership with big data." IEEE (2014), pages 1–5.

Cite this article as :

Rishabh Gurjar, Sahana H K, Neelambika C, Sparsha B Sathish, Ramys S, "Stroke Risk Prediction Using Machine Learning Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 4, pp. 20-25, July-August 2022. Available at doi : https://doi.org/10.32628/CSEIT2283121
Journal URL : https://ijsrcseit.com/CSEIT2283121