HEART STROKE PREDICTION USING MACHINE LEARNING TECHNIQUES

G.L.Sravanthi[1] M.Chandana[2] V.Lakshmi3 H.Swapna[4] V. N.L. Pavani[5] Department of Computer Science and Engineering. Vignan's Nirula Institute of Technology and Science for Women. Pedapalakaluru, AP, India.
Corresponding Author mail id: glsravanthi88@gmail.com

Abstract

Most strokes occur by an unexpected obstruction of courses prompting the brain and heart. Early awareness of different warning signs of stroke can minimize the stroke. In this paper, we propose early prediction of stroke diseases using differentmachine learning approaches with the occurrence of hypertension, body mass index level, heart disease, average glucose level, smoking status, previous stroke and age. Using these high features attributes, five different classifiers have been trained, namely: Logistics Regression, Decision Tree Classifier, Gaussian Classifier, Gradient Boosting Classifier, XGBoost Classifier for predicting the stroke. Afterwards, results of the base classifiers have been aggregated using the weighted voting approach to reach highest accuracy. And here this study has achieved an accuracy of 95%, where the weighted voting classifier performs better than the base classifiers.Thismodelgivesthebestaccuracyforthe prediction of stroke. The area under the curve value of the weighted voting classifier is also high. False positive rate and false negative rate of weighted classifierislowestcomparedwithothers.Asaresult, weighted voting is almost the perfect classifier for predicting the stroke that can be used by physicians and patients to prescribe and early detect a potential stroke.

1. Introduction

A stroke happens when the blood flow to various areasofthebrainisdisruptedordiminished,thecells in those regions do not get the nutrients and oxygen and start to die. A stroke is a medical emergency which requires immediate care. Early detection and proper management are required to minimize the further damage in the affected area of the brain and other complications in the body parts. According to theWorldHealthOrganization(WHO)ineveryyear fifteen million people are suffering from stroke worldwideandaffectedindividualsarepassingaway every 4-5 minutes. The two forms of strokes are ischemic and hemorrhagic. In the event of an ischemic stroke, drainage is blocked by clots, and in the event of a hemorrhagic stroke, a weak blood vessel explodes and bleeds into the brain. Strokecan be prevented by a healthy/balanced lifestyle that is wiping off the bad lifestyle like smoking and drinking, controlling body mass index (BMI) and average glucose level, maintaining good health of heart and kidney. The prediction of stroke is necessary and shall be treated to prevent permanent damage or death. In this paper, weconsidered hypertension, BMI level, heart disease, average glucose level as parameters for predicting stroke. In addition,machinelearningcanplayavitalroleinthe decision-making processes in this predictionsystem. In the literature, we found very few recorded works in which machine learning models were used to predict stroke. The machine learning algorithms are artificial neural network (ANN), stochastic gradient descent, decision tree algorithm, k-nearest neighbor (KNN), principal component analysis (PCA), convolutional neural network (CNN), naive bayes etc. We correlated a relation among the disease's attributes such as hypertension, BMI level, average glucose level, heart disease with stroke. A weighted voting classifier is proposed in predicting stroke using the diseases/attributes such as hypertension, body mass index level, heart disease, average glucose level, smoking status, previous stroke and age. The performance of the proposed weighted votingclassifieriscomparedwiththestate-of-the-art classifier such as Logistics Regression (LR), Decision Tree Classifier (DTC), Gaussian, Gradient Boosting Classifier (GBC), XGBoost(XGB).

2. Literature survey

Many researchers have already used machine learning based approaches to predict strokes. Govindarajan et al.

[11] conducted a study to categorize stroke disorder using a text mining combination and a machine learning classifier and collected data for 507 patients. For their analysis, they used various machine learning approaches for training purposes using ANN, and the SGD algorithm gave them the best value, which was95%. Amini et al. [4], [12] conducted research to predict stroke incidence, collected 807 healthy and unhealthy subjects in their study categorized 50 risk factors for stroke, diabetes, cardiovascular disease, smoking,hyperlipidemia,andalcoholuse.Theyused two techniques that had the best accuracy from the c4.5decisiontreealgorithm,anditwas95%,andfor the K-nearest neighbor, the accuracy was 94%. Chengetal.[13]publishedareportontheestimation oftheischemicstrokeprognosis.Intheiranalysis,82 ischemic stroke patient data were used, two ANN models were used to find precision, and 79% and 95% were used. Cheon et al. [14]– [16] performed a study to predict stroke patient mortality. In their study, they used 15099 patients to identify stroke occurrence. They used a deep neural network approach to detect strokes. The authors used PCA to extract medical record history and predict stroke. They have got an area under the curve (AUC) value of83%.Singhetal.[17]performedastudyonstroke prediction applied to artificial intelligence. In their research, they used a different method for predicting stroke on the cardiovascular health study (CHS) dataset. And they took the decision tree algorithm to featureextracttoprincipalcomponentanalysis.They used a neural network classification algorithm to construct the model they got 97% accuracy. Chin et al. [18] performed a study to detect an automated early ischemic stroke. In their study, the main purpose was to develop a system using CNN to automated primary ischemic stroke. They collected 256 images to train and test the CNN model. Intheir system image proposing to remove the impossible area that can't occur from stroke, they used thedata prolongation method to raise the collected image. TheirCNNmethodhasgiven90%accuracy.Sunget al.[5]performedastudytodevelopastrokeseverity index. They collected 3577 patient's data with acute ischemic stroke. For their predicting models, they used various data mining techniques and linear regression.Theirpredictionfeaturegotthebestresult from the k-nearest neighbor model (95% CI). Monteiro et al. [19] performed a study to get a functional outcome prediction of ischemic stroke using machine learning. In their research, they apply this technique to a patient who was passing three months after admission. They got the AUC value above 90%. Kansadub et al. [20] performed a study to predict stroke risk. In the study, the authors employed Naive Bayes, Decision Tree, and Neural Network to analyze data to predict stroke. In their study, they used accuracy and AUC as their pointer's.

3. Proposed System

Intheproposedsystemweusefivemachinelearning classifiers which we used to build stroke predictors. And this classifiers list is:LR, Decision tree, Gaussian, GBC, XGB. The Reason behind choosing these classifiers is that these are well known classifiers in building vulnerability predictors and used in several similar research works. We choose these five classifiers for building vulnerability predictors in our model. These well-known classifiersareusedinseveralresearchworks,similar to ours. Moreover, these models are evaluated by measuringtheconfusionmatrices.Nowpeoplear not consulting the doctor even if they have some health problems, because of their busy schedule or their negligence. It leads to finding dangerous diseases like heart diseases at early stages. One has to spend more money on scanning systems for predicting Heart stroke. Our existing models cannot predict accurately that the person can get the stroke or not in the future based on certain parameters like BMI, heart disease, smoking status, age, gender etc.

To overcome this problem, our project predicts accuratelywhetherthepersoncangetastrokeornot. Thisisthepurposeofourprojectandbythiswaywe can justify ourproject.

Advantages:

1. Accuracy ishigh.
2. Simplearchitecture.

We can call a Logistic Regression a Linear Regression model but the Logistic Regression usesa morecomplexcostfunction,thiscostfunctioncanbe defined as the '**Sigmoid function**' or also known as the 'logistic function' instead of a linear function. The hypothesis of logistic regression tends it tolimit the cost function between 0 and 1. Therefore linear functions fail to represent it as it can have a value greater than 1 or less than 0 which is not possible as per the hypothesis of logisticregression.

$$0 \leq h_\theta(x) \leq 1$$

Logistic regression hypothesis expectation.

4. Results: -

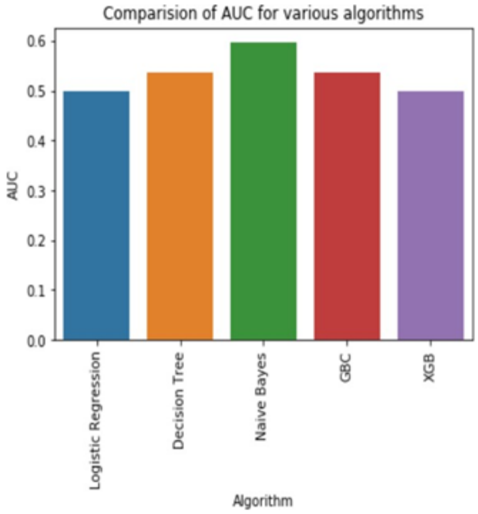

Fig: comparison of Algorithms under AUC

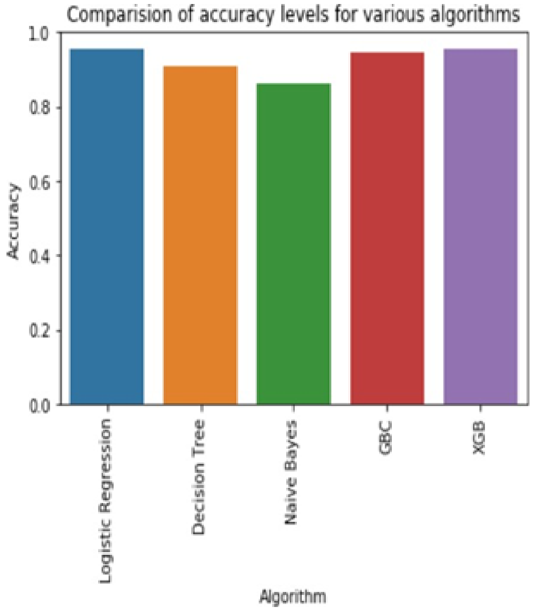

Fig: Accuracy levels of Algorithms

5. CONCLUSION:-

In this paper, we used ten classifiers to find out the performance of stroke occurrence of a person. The proposed weighted voting classifier considered gender, age, hypertension, heart disease, average glucose level, BMI, smoking status featureattributes topredictstroke.Theperformanceevaluationreveals that weighted voting provided the highest accuracy of about 95% compared to the commonly used other machine learning algorithms. As a result, the weighted voting can be considered for theprediction of stroke. We have evaluated the relationship between these diseases and the possibility of occurringstrokeinahumanindividual.So,ifwecan maintain this disease from an early stage then it will help to reduce stroke in our life. In the future, we wouldliketofusedeeplearningbasedimaging,such as brain CT scan and MRI, together with an existing model that will boost performanceindices.

REFERENCES

[1] M. Mahmud et al., “A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications,” Cognitive Computation, vol. 10, no. 5, pp. 864–873, 2018.

[2] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. Al Mamun, and M. Mahmud, “Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detectionofAlzheimer'sdisease,parkinson’sdisease and schizophrenia,” Brain Informatics, vol. 7, no. 1, pp. 1–21, 2020.

[3] M. Mahmud, M. S. Kaiser, and A. Hussain, “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, andN.ToghianFar,“Predictionandcontrolofstroke by data mining,” International Journal of Preventive Medicine, vol. 4, no. Suppl 2, pp. S 245–249, May 2013.

[5] S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.- W. Chen, and Y.-H. Hu, “Developing a stroke severity index based on administrative data was feasible using data mining techniques,” Journal of Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300, Nov.2015.

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

[7] S.M.Reza,M.M.Rahman,M.H.Parvez,M.S. Kaiser, and S. Al Mamun, “Innovative approach in web application effort & cost estimation using functional measurement type,” in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2015, pp.1–7.