STROKE PREDICTION USING ML CLASSIFICATION ALGORITHMS

Dr.M.Rajaiah, Dean Academics & HOD, Dept of CSE, Audisankara College of Engineering and Technology, Gudur.**Ms.M.Tejaswi**, B.Tech, Dept of CSE, Audisankara College ofEngineering and Technology, Gudur.
Mr.K.Rahul, B.Tech, Dept of CSE, Audisankara College ofEngineering and Technology, Gudur.
Ms.N.Mounika, B.tech, Dept of CSE, Audisankara College ofEngineering and Technology, Gudur.
Mr.N.Nitish, B.Tech, Dept of CSE, Audisankara College ofEngineering and Technology, Gudur.

ABSTRACT

Stroke is a medical disorder that harms the brain by rupturing the blood vessels there. It can also happen when the passage of blood and other nutrients to the brain is interrupted The World Health Organization (WHO) claims that stroke is the main globalcause of mortality and disability. The prediction of heart attacks has been studied, however likelihood of a brain stroke is depicted in very few works. Due to this evertheless, numerous machine learning models are created to forecast the potential for a brain stroke. This essay contains a variety of physiological variables with machine learning techniques, such as Decision Tree Classification, Random Forest, and Logistic Regression K-Nearest Neighbors, support vector machines, and classification likewiseNaive Bayes.

Keywords—Machine learning; logistic regression; decision tree classification; random forest classification;

1.INTRODUCTION

Stroke is the fifth-leading cause [1] of death in the US, according to the Centers for Disease Control and Prevention (CDC). Stroke is a non-communicable disease thataccounts for around 11% of all fatalities. Over 795,000 people in the US suffer with theaftereffects of a stroke on a regular basis [2]. In India, it ranks as the fourth most important cause of death.

Machine learning can be used to forecast the onset of a stroke thanks to advancements in medical technology. The algorithms used in machine learning are beneficial in providing accurate analysis and producing accurate predictions. The majority of the prior efforts on stroke concern heart stroke prediction. The research on brain stroke is rather limited. This study uses machine learning to forecast the likelihoodof a brain stroke. Nave Bayes has done the best among the five various classification algorithms utilised, gaining a higher accuracy metric, according to the essential elements of the techniques used and results achieved. This model's drawback is that it is being trained on textual data rather than real- time brain data.

To continue with this challenge, a dataset from Kaggle [3] is picked that has a variety of physiological features as its properties.

The final forecast is based on an analysis of these qualities. The dataset is initially prepared for the machine learning model's comprehension by being cleaned.

Data preprocessing

the procedure at this point. The dataset is searched for null values and filled up accordingly. If necessary, one-hot encoding is done after label encoding to encode string values into number

The dataset is divided into train and test data following data preprocessing. The new data is then used to create a model using a variety of classification algorithms. For each of these methods, accuracy is assessed and compared to obtain the most accurate prediction model.An HTML website and a Flask application are created when the modelhas been trained and accuracy determined. The user enters the values for the prediction in the online application. The web application and the trained model are linked through the flask application. The research comes to a conclusion about which algorithm is bestfor stroke prediction after thorough analysis

2.LITERATURE SURVEY

Using five machine learning algorithms, the Cardiovascular Health Study (CHS) dataset was used in [4] to predict strokes. The authors used the Decision Tree with the C4.5 method,Principal Component Analysis, Artificial Neural Networks, and Support Vector Machine to provide the best result. The CHS Dataset, however, which was used for this work, has fewer input parameters. In [5], stroke prediction was done using user-posted social media content. The DRFS approach was employed by the authors of this study to identify the various stroke-related symptoms. Natural Language Processing is used to extract text from social media posts, although this adds to the model's total execution time, which is undesirable.
According to research publication [7], the model was trained to predict strokes using decision trees, random forests, and multi-layer perceptrons. The three approaches' achieved accuracies were quite similar, with just minor variations. Decision Tree, Random Forest, and Multi-layer perceptron all had computed accuracy of between 74.31% and 74.53%.

including Decision Tree, Naive Bayes, and SVM, and then compared the results. The methods they utilised yielded a maximum accuracy of just 60%,which is rather low.

In [9], the authors forecast the likelihood of a stroke using several data mining categorization approaches. The Ministry of National Guards Health Affairs Hospitals in the Kingdom of Saudi Arabia provided the dataset. C4.5, Jrip, and multi-layer perceptron were the threeclassification methods employed (MLP). The model achieved an accuracy of about 95%using these strategies. Despite the fact that the study promises to achieve accuracy of 95%,the time.

Three distinct methods may be used, according toresearchpublished in [10], toforecast the likelihood ofhaving astroke. These algorithms include Neural Networks, Decision Trees,and Naive Bayes. This study found that, among the three algorithms, the decision tree had the best accuracy (about 75%). Nevertheless, based on the results from the

confusionmatrix, this modelcouldnotaccount for the casesfromthe actual world.

The researchers used the Cardiovascular Health Study (CHS) dataset to do stroke prediction in [11]. On the basis of their suggested conservative mean, they introduced a unique automatic feature selection technique that chooses robust features. For further effectiveness, they paired this approach with the Support Vector Machine technique. However, this led to the creation of manyvectors that have atendency to diminish

The prediction of thrombo-embolic stroke disease using artificial neural networks is suggested by research in [12]. The Back-propagation algorithm was employed as the prediction approach. This model was successful in obtaining an accuracy of about 89%.However, due of their complicated structure and growing number of neurons, neural networks take longer to train and demand more processing time.

3.SYSTEM METHODOLOGY

Various Kaggle datasets were taken into consideration in order to move on with the implementation. Anrelevant dataset was gathered fromamong all ofthe already-existing datasetsfor model construction.

The process of preparing the dataset to make it more comprehensible for machines

comes after the dataset

has been colected. Data preparationis the name given to this stage. This

Steps that are specific to this dataset include addressing missing values, handling data that is unbalanced,

and conducting label encoding.

1. Now that the data is preprocessed, it is ready for model building. For model building, preprocessed dataset along with machine learning algorithms are required. Logistic Regression, Decision Tree Classification algorithm, Random Forest Classification algorithm, K-Nearest Neighbor algorithm, Support Vector Classification and Naïve Bayes Classification algorithm are used. After building six different models, they are compared using five accuracy metrics namely Accuracy Score, Precision Score, Recall Score, F1 Score and Receiver Operating Characteristic (ROC) curve.

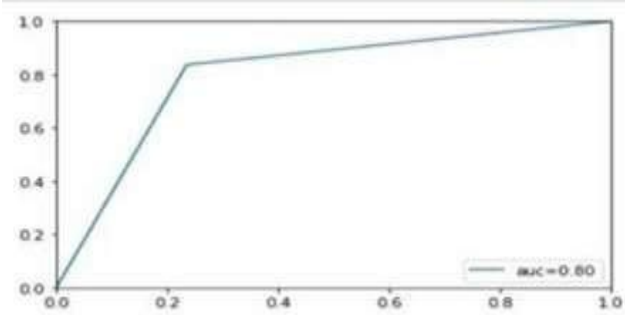
2. The comparison of the models gives the best model in terms of the accuracy metrics to proceed with the deployment phase. For deploying the model, an HTML page is developed to make it user-friendly for the user to enter the input parameters and get the result. The parametersentered by the user are sent to the model using a flask application which is basically a python framework that links the web application and the model together. The model takes the input parameters, predicts the output andreturns the result to the flask application. Now this flask willdisplay that result on theweb page for the user to check the result.

3.IMPLEMENTATION

The implementation of this project is as follows.A,.Dataset

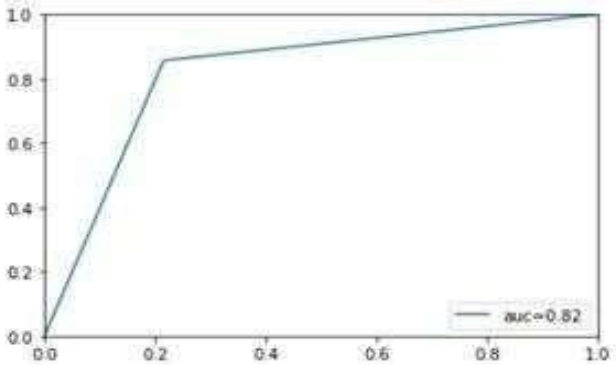
The dataset for stroke prediction is from Kaggle [3]. This particular dataset has 5110 rowsand 12 columns. The columns

have'id','gender','age','hypertension',heart_disease','ever_marri ed', 'work_type', 'Residence_type','avg_glucose_level', 'bmi', 'smoking_status' and 'stroke' as the main attributes. The output column 'stroke' has the value as either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possiblerisk of stroke.This dataset is highly imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column. Only 249 rows have the value '1' whereas 4861rows with the value '0' in the stroke column. For better accuracy, data pre-processing is performed to balance the data. The dataset discussed above is summarized in Table 1.



Naive Bayes classification is another method of supervised learning. An assumption made by a Nave Bayesclassifier [19] is that the existence of one feature in aclass has no bearing onthe presence of any other features. Its foundation is the Bayes Theorem. Every characteristic or attribute identified is independent of one another,according to the algorithm. The accuracy achieved with this method was 82 percent, with precision scores of

79.2 percent and recall scores of 85.7 percent. This method produced an F1 Score of 82.3 percent.According to Fig. 9, the Nave Bayes Classification Receiver Operating Characteristic (ROC) curve has an 82 percentaccuracy rate.



After developing the model, it can be said that Naive Bayes has outperformed other methods. As a result, pickle is used to dump the model that was trained using Naive Bayes classification. The following step is to createweb and flask applications for entering the input parameters.The web page was created using basic HTML.This programme features an input form that collects user input values

and uses them to forecast the likelihood of a stroke occurring. The flask application receives the input parameters when the user hits the 'Check Here' button. In Fig. 10, a section of the HTML page is displayed.The link between the web page and the flask application, which is essentially pythoncode

Fig. 10. Input Form in HTML Code.

```
from flask import Flask, request, render_template
import numpy as np
import pickle

model=pickle.load(open('stroke_model.pkl','rb'))

app = Flask(__name__)
@app.route('/')
def home():
    return render_template("home.html")
```

Fig. 11. Flask Application Linking the Model and Web Page.

The output is predicted by the machine learning model after receiving the input parameters. The flask programme then updates the web page with the acquired result so that the user can view theforecast.

4.C0NCLUSION

Stroke is a serious medical illness that has to be treated right away to avoid getting worse. The creation of a machine learning modelcanaid in the earlydetection of strokeand lessen its severe effects. This studyexamines how well different machine learning algorithms predict stroke based on a variety of physiological characteristics. With an accuracy of 82 percent, Nave Bayes Classification outperforms all other methods. Fig. 12 depicts a comparison of the accuracy results from different methods. In terms of accuracy, recall, and F1 scores, Nave Bayes outperformed the others. In Figures 13, 14, and 15, the comparison of the Precision score, recal score, and F1 scoreis displaye

Fig. 12. Comparing the Accuracies of ML Algorithms.

Fig. 13. Comparing the Precision Scores of ML Algorithms

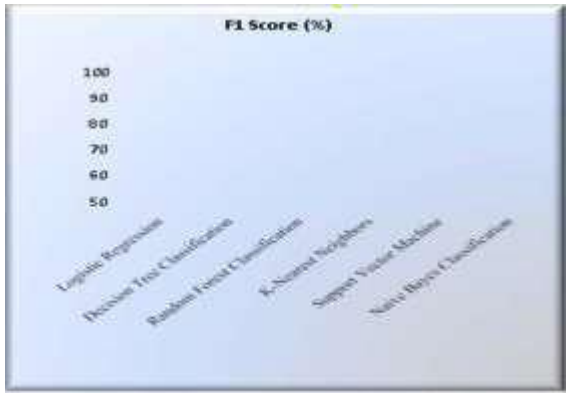


Fig. 15. Comparing the F1 Scores of ML Algorithms

This study recommends using different machine learning techniques to the dataset

used. By employing neural networks to train the model, this project may be further developed. More accuracy criteria can be taken into account while comparing the performance. This research is restricted to textual data, which may not always be reliable for predicting strokes. In the future, it would be more effective to compile a collection of pictures, such as brain CT scans, to forecast the likelihood of stroke

REFERENCES:-

Concept of Stroke by Healthline.

Statistics of Stroke by Centers for Disease Control and Prevention.

Dataset techniques.

Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. In Springer (2020).

Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy: Prediction of Brain Stroke Severity Using Machine Learning. In: International Information and Engineering Technology Association (2020).

Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from

electronic health records. In: 41st Annual International Conference of the IEEE

Engineering in Medicine and Biology Society IEEE (2019).

Fahd Saleh Alotaibi: Implementation of Machine Learning Model to Predict Heart Failure Disease. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2019).

Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018)

AUTHOR PROFILES

Dr.M.Rajaiah , Currently working as an Dean Academics & HOD in the department of CSE at ASCET (Autonomous), Gudur, Tirupathi(DT).He has published more than 35 papers in, Web of Science, Scopus Indexing, UGC Journals.

Ms.M.Tejaswi, as B.Tech student in the Learning department of CSE at Audisankara College of Engineering and Technology, Gudur.. His areas of interests are Networks, Mobile Wireless Networks, Big Data, Data warehousing and Data Miningand Deep

Mr.K.Rahul, as B.Tech student in the department of CSE at Audisankara Collegeof Engineering and Technology, Gudur.. His areas of interests are Networks, Mobile Wireless Networks, Big Data, Data warehousing and Data Miningand Deep Learning.

Mr.N.Mounika, as B.Tech student in the department of CSE at Audisankara College of Engineering and Technology, Gudur.. His areas of interests are Networks, Mobile Wireless Networks, Big Data, Data warehousing and Data Miningand Deep Learning.

Mr.N.Nitish, as B.Tech student in the department of CSE at Audisankara College of Engineering and Technology, Gudur.. His areas of interests are Networks, Mobile Wireless Networks, Big Data, Data warehousing and Data Miningand Deep Learning.