IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

BRAIN STROKE PREDICTION USING SUPERVISED MACHINE LEARNING

[1] Kallam Bhavishya, [2]Shaik.Althaf Rahaman

[1] PG Student, [2]Assistant Professor

[1] Department of Computer Science,

[1]GITAM (Deemed to be University), Visakhapatnam, India

Abstract: A Stroke is a medical disorder that damages the brain by rupturing blood vessels. It can also happen when the brain's blood flow and other nutrients are interrupted. Stroke is the greatest cause of death and disability worldwide, according to the World Health Organization (WHO). The majority of research has focused on the prediction of heart stroke, while just a few studies have looked at the likelihood of a brain stroke. With this in mind, various machine learning models are being developed to forecast the likelihood of a brain stroke. This article employed machine learning techniques like K-Nearest and Nave Bayes Classification to model many physiological parameters for accurate prediction and discover the optimum approach.

Index Terms - Index Terms - Machine learning, K-nearest, Naive bayes, Brain Stroke

I. INTRODUCTION

According to the Centers for Disease Control and Prevention, stroke is the sixth greatest cause of death in the United States (CDC). Stroke could be a non-communicable disease that kills about 11% of the population. Every day, over 795,000 persons in the United States are affected by the effects of a stroke. It is the fourth largest cause of death in India. Because of developments in medical technology, Machine Learning can now predict the occurrence of a stroke. Machine Learning algorithms are useful for making accurate predictions and analyzing data. The majority of previous stroke research has focused on predicting heart attacks. Little or no attention has been paid to brain stroke. The algorithms that exist in machine learning are constructive to make accurate predictions and provide correct analysis. The work done so far on the topic of stroke mainly includes work on heart rate prediction. Little research has been done on stroke. This paper is based on using machine learning to predict the occurrence of stroke. The main component of the approach used and the results obtained are between two different classification algorithms. The limitation of this model is that it's trained with matter information instead of time period brain pictures. This paper shows the implementation of 2 machine learning classification algorithms. you'll be able to extend this paper more to implement all current machine learning algorithms. A dataset with various physiological characteristics has been selected as. These features will be analyzed later and used for final predictions. The dataset is first cleaned up and prepared to understand the machine learning model. This step is called data preprocessing. To do this, the zero values in the data record are checked and these are entered. Then label encoding is performed to convert the string value to an integer, followed by one-hot encoding as needed. After preprocessing the data, the dataset is split into training and test data. This new data is then used to build a model using various classification algorithms. The accuracy of all these algorithms is calculated and compared to obtain the best trained model for the prediction.

II. LITERATURE SURVEY

In stroke predictions were made of the vas Health Study (CHS) dataset victimization 5 machine learning techniques. As the optimal solution, the paper used a combination of decision tree and C4.5 algorithm, principal component analysis, artificial neural network, and support vector machine. However, the CHS dataset extracted for this task had a small number of input parameters. In stroke predictions were made from social media posts posted by people. In this particular study, the paper used the DRFS method to find a variety of symptoms associated with stroke disease. Extracting text from social media posts using natural language processing will increase the overall execution time of the model, which is not desirable. In this paper performed a stroke prediction task using an improvised random forest algorithm. It was used to analyze the risk level achieved with the stroke. As the paper suggests, this method is said to have performed better than existing algorithms. This particular study is limited to a very small number of types of strokes and cannot be used for new types of strokes in the future.

III. SYSTEM METHODOLGY

Various datasets were considered to advance the implementation. Suitable datasets for modeling were collected from all available datasets. After collecting the dataset, the next step is to prepare the dataset to make the data clearer and easier for the machine to understand. This step is called data preprocessing. This procedure involves handling missing values, handling imbalanced data, and performing label encoding that is specific to that particular dataset.

Now that the data has been preprocessed, you are ready to build the model. Model building requires a preprocessed dataset with machine learning algorithms. Logistic regression, the K-Nearest Neighbor algorithm, and the Naive Bayes classification algorithm are used. After two different models are built, they are compared using two precision metrics. Comparing the models reveals the best model in terms of accuracy metrics.

It shows the methodology flow chart of the proposed system as shown in **Figure 1**

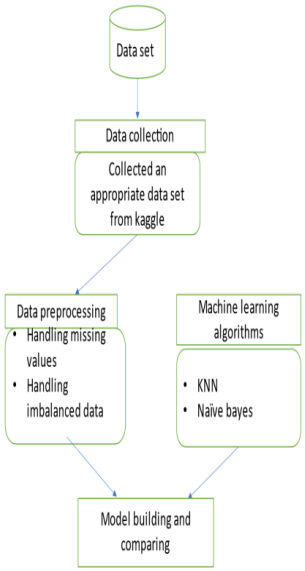

Fig.1 system methodology

IV. IMPLEMENTATION

The implementation of this project is as follows. A. Dataset for The stroke prediction data set is choosen . This particular dataset has 5110 rows and 12 columns. The columns include ID, gender, age, hypertension, heart disease, Always_Married, WorkType, ResidenceType, Average_Glucose Level, BMI, smoking status, and stroke. As the main attribute. The value of the output column "Hub" is "1" or "0". A value of 0 indicates that no risk of stroke was identified, and a value of 1 indicates a possible risk of stroke. This dataset is very imbalanced because the possibility of a "0" in the output column ("dash") is more important than the possibility of a "1" in the same column. The stroke column has a value of "1" in only 249 rows and a value of "0" in row 4861. To improve accuracy, perform data preprocessing to equalize the data.

4.1 Data pre-processing

Before building the model, the data needs to be preprocessed to remove unwanted noise and outliers from the dataset, which deviates from proper training. Anything that prevents the model from becoming less efficient is done in this phase. After collecting the appropriate dataset, the next step is to clean up the data and make sure it is ready to build the model. The retrieved dataset has the 12 attributes listed in Table I. First, the "id" column is omitted because its presence does not make a big difference in model building. The record's null value is then checked and filled in as needed. In this case, the BMI column contains a null value filled with the average of the column data. After the null value is removed from the dataset, the next task is to encode the label.

4.2 Label Encoding

Label encoding encodes string literals in a dataset to integer values so that the machine can understand them. Machines are usually trained numerically, so you need to convert the string to an integer. The collected dataset has five columns with a string as the data type. When you perform label encoding, all strings are encoded and the entire record is a combination of numbers.

4.3 Handling Imbalanced Data

The dataset selected for the stroke prediction task is severely imbalanced. The entire dataset has 5110 rows, and 12 columns. Training machine-level models with such data improves accuracy, but other compliance indicators such as fit and recall are functional. If such imbalanced data is not processed, the results will be inaccurate and the predictions will be inefficient. Therefore, to get an efficient model, these imbalanced data must be processed first.

V.MODEL BUILDING

A. Splitting the Data

After pre-processing the data and processing the imbalanced dataset, the next step is to build the model. The subsampled data is divided into training and test data to improve the accuracy and efficiency of this task while maintaining a ratio of 80% training data to 20% test data. After splitting, train the model using various classification algorithms. The classification algorithms used for this purpose are logistic regression, decision tree classification algorithm, random forest classification, K-Nearest neighbor classification, support vector machine, and simple Bayes classification.

B. Classification Algorithms

i.K-nearest neighbors classification:

Another set of rules used for class is K-Nearest Neighbors (KNN). It is likewise a supervised gaining knowledge of technique. KNN [17] is a lazy set of rules that could now no longer teach at once on giving the dataset. Instead, it shops the dataset, and on the time of class, it acts at the dataset. The operating precept of KNN is to discover similarities among the brand new case (or facts) and to be had facts after which map the brand new case into the class this is maximum just like the to be had categories. The accuracy acquired is 93.2%.

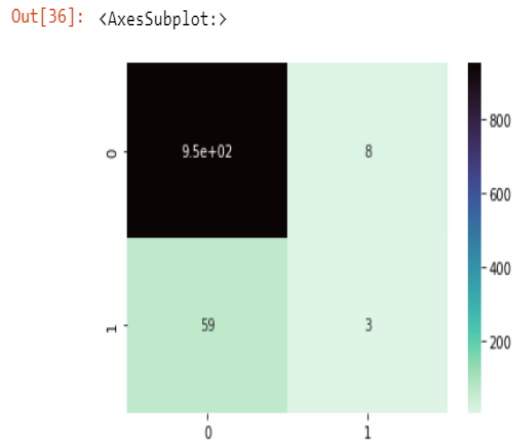

Fig-2: **K-NN Classification**

ii.Naïve bayes classification:

It is also a supervised learning method. The naive Bayes classifier assumes that the presence of a particular feature in a class is independent of the presence of other features. This is based on Bayes' theorem. This algorithm follows the principle that each classified function or attribute is independent of each other. The accuracy achieved with this algorithm was 86.8%.

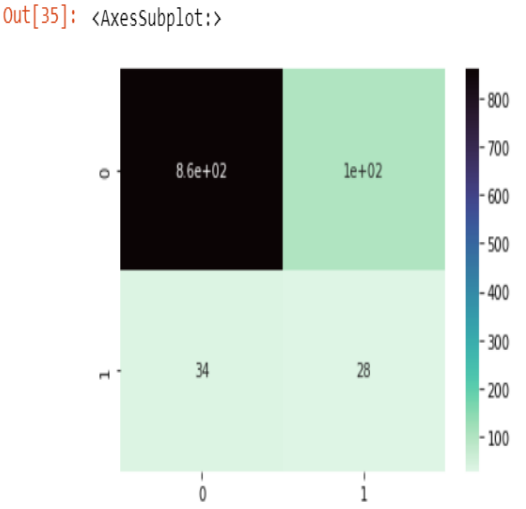

Fig-3: **Naïve bayes classification**

VI. CONCLUSION

Stroke is a serious condition that needs to be treated before it gets worse. Building machine learning models can help you predict stroke early and mitigate serious future impacts. This paper demonstrates the performance of various machine learning algorithms in successfully predicting stroke based on multiple physiological attributes. Of all the algorithms selected, the K-nearest Neighbors classification provides the best performance with 93% accuracy. Figure 1 shows a comparison of the accuracy of the various algorithms.

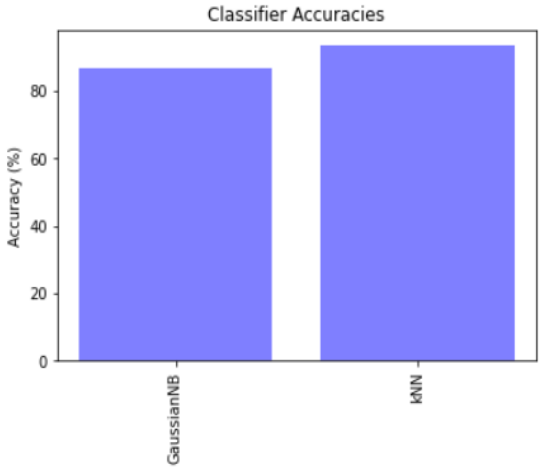

Figure-4: Classifier Accuracies

REFERENCES

[1] Concept of Stroke by Health-line.
[2] 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).
[3] Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy: Prediction of Brain Stroke Severity Using Machine Learning. In: International Information and Engineering Technology Association (2020).
[4] 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).
[5] Fahd Saleh Alotaibi: Implementation of Machine Learning Model to Predict Heart Failure Disease. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2019).
[6] Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018)