Stroke Prediction Using Machine Learning: A Review

Sandhya Gaikwad[1], Samina mulani[1], Ankita Shelke[1]

[1]Student, Department of Computer Engineering, vidya pratishthan kamalnayan bajaj institute of engineering and technology, Baramati, Maharashtra, India

Corresponding Author: saminamulani46@gmail.com

Abstract— A stroke occurs when blood supply to the part of brain is reduced or interrupted causes blockage in an artery which is serious issue. It is second major reason for deaths in worldwide. It is caused due to people lifestyle decision, high blood sugar, heart disease, obesity, hypertension. Due to this prediction of stroke becomes necessary and with the help of effective prediction algorithm which allow for early diagnosis and intervention. Several models are developed and evaluated to design a robust framework for long term risk prediction of stroke occurrence. Using different machine learning algorithm models namely gaussian naive bayes, logistic regression, decision tree, k nearest neighbour. The efficient data collection, data preprocessing, data transformation methods applied to provide reliable information for model to be successful. The performance of each classifier is estimated based on evaluation metrics such as accuracy, error rate loss function. It has possible to obtain accuracy of 98 %.

Index Terms—**Stroke Prediction, Machine learning Model, Decision Tree.**

1. Introduction

The most serious and deadly disease in humans is brain stroke, which is associated with a high death rate, poses risk and burden to healthcare systems worldwide. The brain is primary upper organ in human body. stroke which is medical disorder in which blood arteries in the brain are ruptured and causing damage to the brain. when blood supply to the brain is interrupted symptoms might develop. stroke causes long term neurological damage and complications causes neurological damage and often death. strokes symptoms include paralysis on one side of the body, difficulty in speaking or walking, dizziness, blurred vision, headache, vomiting, dropping mouth and loss of consciousness or coma. Stroke can impact on quality of life and putting load on public health resources. every year, more than 15 million people worldwide have a stroke. Therefore, the prediction of stroke becomes necessary and should be used to prevent damage by stroke.

Early detection of the numerous stroke warning symptoms can lessen the stroke severity. The main objective is to forecast the possibility of a brain stroke occurring at an early stage using machine learning techniques. using machine learning algorithm to identify risk variables is to early detection to prevent risk. Early brain stroke prediction yields a higher amount that is profitable for initiating time. Brain stroke is caused primarily by people's lifestyle decisions particularly in current situations. using different machine learning algorithm namely Naive Bayes, Logistic Regression, Decision tree classifier, k nearest neighbours, Random Forest classifier, AdaBoost, XG Boost classifier. the performance of each classifier is estimated by adopting performance evaluation metrics such as accuracy, error rate, root mean squared error, log loss. machine learning techniques are applied in identifying, classifying and predicting the stroke from medical information.

2. Stroke Prediction Theory

RPM (Remote patient monitoring) has the ability to significantly alter how patients are treated. If widely adopted, it will drastically alter the patient experience and make it possible for healthcare providers to manage their resources significantly more effectively than they presently are, reducing pressure on healthcare systems around the world. The patients admitted in hospitals are monitored by the nurses and are timely noted and reported to the doctor. In order to reduce man power we need specific devices that gives us an accurate measure of the patient's heart rate, temperature and etc. this brings out the idea of Quarantined patient monitor. The patient's complete report in displayed in the cloud server where the data is stored and linked to the application. Anyone can access the application through an authorized access to check the patient reports and monitor them from any remote location. When the patient's condition is critical the person who wish to monitor can get a message to phone number which is set for sending the alert notifications and alarms. And even the patient's location is tracked using GPS. And also, the patient can communicate with the microphone, as and when required. Smart health care is advantageous as the include efficient time usage, improves communication and better extended care. Through apps and a health information platform, patients can manage their condition on their own. Smart healthcare can help individuals

better manage their own health.

3. Literature Survey

Geethanjali et al. Stroke attack can be predicted accurately.they have used three classifier such as logistics regression .decision tree classifier and support vector machine for the prediction of stroke.

The classification model is based on 5110 records and they have provided the results with 95.49% of accuracy.

Veena Potdar et al. The different methods of prediction of stroke and concluded that each technique for prediction of stroke and concluded that each techniques has its own advantage and disadvantage, random forest techniques is considered as one of the techniques for analysing which shows promising result.

Amini et al. 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, and alcohol use. They used two techniques that had the best accuracy from c4.5 decision tree algorithm, and it was 95%, and for K-nearest neighbor, the accuracy was 94%

Singh et al. performed a study on stroke 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 feature extract to principal component analysis. They used a neural network classification algorithm to construct the model they got 97% accuracy.

4. Methodology

A. Dataset description

The dataset containing 43400 instances and 12 features. 9age,hypertension,gender,heart_disease,bmi,work_type,avg_g lucose_level,residence_type,ever_married.the stroke column represent the outcome where 0 indicates no stroke and 1 indicates stroke detected. The dataset contains 42617 non - stroke and 31962 strokes after balanced data ;783strokes detected before balanced data.

B. Dataset preprocessing

The preprocessing techniques are used to remove unwanted noise, missing values, outlier, label encoding, and so on. after data cleaning machine learning models are applied to the dataset. normalization techniques applied to prepare data which transform the value of numeric column to use a common scale lies between 0 and 1 and data standardization deviation of 1 for scaling of model.

C. Label encoding

Label encoder refers to encoding the categorical value to fit in the machine learning model smoothly. there are five columns named gender, ever married, work type, residence _type, and smoking status that comprise string values. all these string values are converted into a combination of numerical values.

D. Imbalanced data handling

The handling imbalance dataset using data balancing techniques random oversampling is applied to dataset. the main purpose of this method to randomly duplicate the examples from minority class .by repeating the original samples the random over sampler expands the dataset.

Applying model and Evaluation and performance metrics

The scope of stroke prediction includes various risk factors such as age, gender, medical history, lifestyle factors and biomarkers to develop models that can accurately predict an individual's risk of experiencing a stroke.it includes identifying and analyzing various medical risk factors associated with stroke, such as hypertension, diabetes, high cholesterol, smoking, obesity, family history of stroke. understanding how these factors interact and contribute to stroke risk is crucial for accurate prediction. these predictions can help healthcare professionals intervene early with preventive measures, lifestyle changes, or medical treatments to reduce the likelihood of a stroke occurrence. additionally, advancements in technology such as artificial intelligence and machine learning are being employed to enhance the accuracy and reliability of stroke prediction models.

5. Algorithm

A. Naive bayes

Naive bayes classifier which ensures probability maximization if the features are highly independent. decision making by naive bayes uses total probability value of all the diagnostic criteria on the existing database.it is supervised learning method .it assume that presence of a particular features in class is independent of the presence of other features. this algorithm follows the principle that each classified function or attribute is independent of each other.

B. K nearest neighbour

Another set of rules is used for k nearest neighbour.it is likewise a supervised gaining knowledge of technique. knn is a lazy set of rules that could now no longer teach at once on giving on the dataset. the operating percept of knn is to discover similarities among the brand new case and to be had facts after which map the brand new case into the class this is maximum just like to be had categories.it is commonly based on Euclidian,Manhattan and minkowski distance algorithm that evaluates distance from new data to all others.

C. Random forest

There are multiple decision trees in a random forest classifier. when these independent trees are combines in an ensemble through resampling, the results becomes subsets of instances that are used for classification and regression.in an random forest, the final output is a majority voting since each independent trees generates its own classification outcome.

D. Support vector machine

Support vector machine is a popular supervised learning algorithm used in medical fields for many years for predicting

the results and used for classification and regression problems. The classification performs most suitable hyperplane distinguishing the dataset between two classes. svm model to map features about patients' relevant information to stroke. svm can predict stroke effectively.

6. Results And Discussion

In this paper, presented an integrated machine learning approach combining the elements of data imputation, feature selection and prediction. We provide an extensive comparison of machine learning methods with the Cox proportional hazards model and show that the machine learning methods significantly outperform the Cox model in terms of both binary stroke prediction and stroke risk estimation. Specifically, we propose the conservative mean heuristic for feature selection, which gives us the best performance as compared to other methods. In addition, we present a novel prediction algorithm, Margin-based Censored Regression, that achieves a better concordance index than the Cox model. Further, our method can be used for identifying potential risk factors for diseases without performing clinical trials. We hope that this paper will motivate the application of machine learning methods in healthcare data analysis.
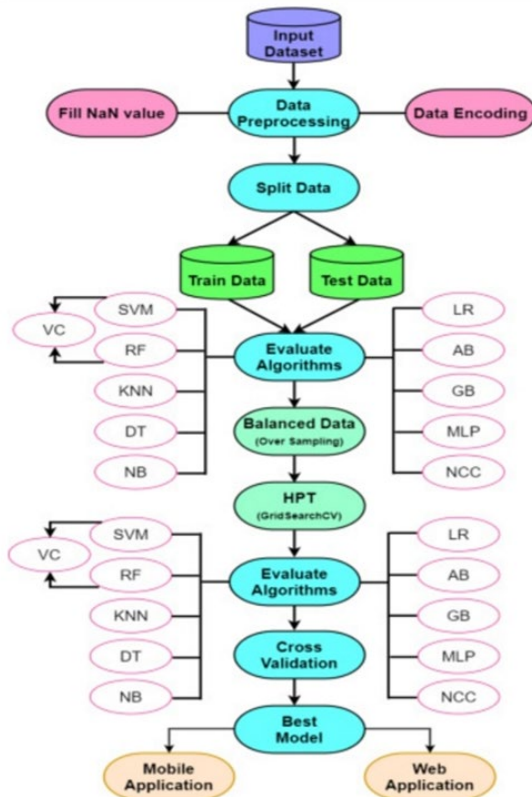

Fig.1. System Architecture of Stroke prediction system

The proposed system produces ranking decisions that were relatively highly consistent with those of the human experts. This system will enable a more effective way to short list submitted candidate CVs from a large number of applicants providing a consistent and fair CV ranking policy. The presented system automates the processes of requirements specification and applicant's ranking. This system can be used in many business sectors that may require expert candidates and also reduce workload of the human resource department.

7. Conclusion

A stroke constitutes a threat to a human's life that should be prevented or treated to prevented to avoid unexpected complications. Nowadays, with the rapid evolution of ai/ml, clinical providers, medical experts and decision makers can exploit the established models to discover the most relevant features. the future purpose is to enhance the ML framework and promising direction for CT scaning of image and evaluate predictive ability of model in stroke occurrence.

References

[1]. Naghavi, M.; Abajobir, A.A.; Abbafati, C.; Abbas, K.M.; Abd-Allah, F.; Abera, S.F. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017.

[2]. Hay, S.I.; Abajobir, A.A.; Abate, K.H.; Abbafati, C.; Abbas, K.M.; Abd-Allah, F. Global, regional, and national disability-adjusted life-years (DALYs) for 333 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet 2017, 390, 1260–1344.

[3]. Tatnall, Arthur (2005), 'Web Portals: The New Gateways to Internet Information and Services', Hershey, PA: Idea Group Publishing.

[4]. Sidharth Sankhye and Guiping Hu, Machine Learning Methods for Quality Prediction in Production, Ames, IA 50011, USA:Department of Industrial and Manufacturing Systems Engineering, Iowa State University, October 2020.

[5]. H Asadi, R Dowling, B Yan and P Mitchell, "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy", PLoS One, vol. 9, no. e88225, 2014.

[6]. Kaur Shubpreet and R.K. Bawa, "Review on Data Mining Techniques in Healthcare Sector", 2nd International conference on I-SMAC (IoT in Social Mobile Analytics and Cloud), 2018.

[7]. Sidharth Sankhye and Guiping Hu, Machine Learning Methods for Quality Prediction in Production, Ames, IA 50011, USA:Department of Industrial and Manufacturing Systems Engineering, Iowa State University, October 2020.