Machine Learning Based Approach Using XGboost for Heart Stroke Prediction

Sukhmanjot Dhillon[1], Chirag Bansal[2] and Brahmaleen Sidhu[3]

[1,2,3]*Department of Computer Science and Engineering, Punjabi University Patiala, Punjab*

Abstract

Many prediction methods are widely used in clinical decision-making to predict the prevalence of diseases, assess the prognosis or outcome of diseases, and help doctors treat diseases. However, traditional predictive models or methods are not enough to effectively collect basic data because they cannot simulate the quality of mapping the negative attributes of the medical field. The approach proposed in this paper uses. We use machine learning to predict survival of a heart patient. The approach uses patient's data like gender, age, hypertension, type of work, glucose level, body mass index, etc. to predict his/her chances of death due to heart failure. The dataset is retrieved from Kaggle. Machine learning based classification algorithms namely XGboost, Random Forest, Navies Bayes, Logistic Regression and Decision Tree have been implemented and their performance has been compared using parameters like precision, recall, F1-score and AUC.

Keywords

Machine learning, Stroke, Risk level classification, XGboost

1. Introduction

Heart diseases have seriously affected the world. Coronary artery disease is a common kind of heart disease. It is caused by buildup of plaque in the walls of the coronary corridors. Coronary corridors are answerable for providing blood to the heart and other body organs.Normal indications of the coronary illness are chest torment and distress. In some cases, heart attack is the first sign of the disease. This is accompanied by weakness, light-headedness, nausea, cold sweat, pain in the arms, and shortness of breath. The main causes of this disease are family history of disease, excess body weight, lack of activity, unhealthy eating, use of tobaccoetc.If not treated well in time heart disease can cause heart failure leading to death of patient.

In case of heart diseases, prevention is definitely better than cure. An early warning can be beneficial in saving the life of the patient. A data based system that provides timely indication of the risk of heart failure and is supported by medical information from patient's health data can be revolutionary. Great development has been achieved in the field of clinical and medical services using artificial intelligence, machine learning and data science approaches. Joining sensors with specialized gadgets can assist patients with getting input from all points, regardless of whether they are doing what they are doing. As of late, medical services has moved from the facility level to the patient-

International Conference on Emerging Technologies: AI, IoT, and CPS for Science & Technology Applications, September 06–07, 2021, NITTTR Chandigarh, India

EMAIL: banidhillon1@gmail.com (A. 1); chiragbansal254@gmail.com (A. 2); brahmaleen.sidhu@gmail.com (A. 3)
ORCID: Not Available (A. 1); Not Available (A. 2); 0000-0001-6519-7957 (A. 3)

driven level .In this speedy world, it is not difficult to direct a naturally directed person wellbeing test to get any individual in the tempest before a respiratory failure

The approach proposed in this paper uses machine learning to predict survival of a heart patient. The approach uses patient’s data like gender, age, hypertension, type of work, glucoselevel, body mass index, etc. to predict his/her chances of death due to heart failure. The dataset is retrieved from Kaggle ..AI based grouping calculations specifically XGboost , Random timberland , Navies Bayes , Logistic Regression and Decision Tree have been implemented and their performance has been compared using parameters like precision, recall, F1-score and AUC.

2. Related Work

The forms of machine learning used to predict heart attacks are very useful and have proven their importance in recent years. Manasa, Gupta[1] received a system that can be used to predict recurrent cardiovascular disease what's more, can be utilized by clients with coronary conduit infection. They utilized the Random Forest calculation which gave an exactness of 89%.

Rajliwall, et.al.[2] In this document, you need to plan a framework that supports supervised learning algorithms and package-level processes that use group isolation and channels dependent on sex, training, and age. Sentil Kumar Mohan[3], ChandrasgarTirumalai, GautamSrivachava. Utilize the mixture HRFLM strategy, which consolidates the elements of irregular timberland (RF) and straight technique (LM).

Nashif, Raihan,[4] A model is proposed, which might be a cloud-based coronary illness expectation model, which means to utilize AI calculations to recognize the following kind of coronary illness. Susmitha Manikandan conveys a twofold grouping model in the example paper module of this framework, which is utilized to foresee patients' irregular issues dependent on the patient's clinical information. They utilized a Random Forest With Linear Model that gave a precision of 88.7%.

Gavhane, et.al.[4] In this article, they designed a system in which they used NN formulas and hierarchical perceptrons to train and test data sets. Ravish, K. Shanti, Nayana R. Shenoy, S.Nisarg, and ECG data. Teach artificial neural organizations to precisely analyze and anticipate heart irregularities (assuming any). Utilizing the innocent Bayes strategy, calling the tree, K-closest neighbor, and arbitrary backwoods in 10-crease cross-approval, the exactness rate comes to 80%

Jae Woo Lee[5], The purpose of this article is to calculate and predict the probability of stroke within 10 years: "Computer strategies and procedures in biomedicine" Lee, Hyungsun Lim, et.al.Individual 5 Stroke-like probability

Stroke Probability[6]: A Risk Profile of the Framingham Study, Wolff,et.al. In this article, a health risk assessment function was developed to predict the incidence of stroke in the Framingham study cohort.

Formulate rules to assist sisters in predicting stroke[7]: National Health Insurance Information Survey Min SN, Pak SJ, Kim JJ, Subramaniyam M, Lee KS. -The purpose of this research is to derive the model equations used to develop preliminary recognition algorithms. Stroke with risk factors that may change.

3. Proposed Work

In this article, we have fostered a model that contains a double order of life and a sign of the danger of cardiovascular breakdown and is upheld by clinical data from individual information. The informational index we use comes from the machine Kaggle. Unstructured informational indexes are renewed in organized informational collections. The informational index has twelve ascribes, eleven of which are indicators. One chance is a double reaction variable. The outrageous incline of the slope is utilized in the order cycle. Calculations Involved-Few approaches utilized in our activities are:

Decision Tree: It is a choice help device that utilizes a tree-like chart or model of choices and their potential results, including chance occasion results, asset expenses, and utility. It is one approach to show a calculation that just contains restrictive control articulation

Naïve Bayes: It is a probabilistic machine-learning model that's used for classification tasks. The crux of the classifier is based on the Bayes theorem.

XGboost: XGboost is a good implementation of the gradient enhancement method. Even though there may not be new mathematical developments here, it is a gradient gain alternative that can be carefully designed for optimization and accuracy. It consists of a linear version, and the newborn tree may be a technique that uses various AI calculations to verify whether a fragile newbie will create a reliable newbie to improve the accuracy of the version. From (impulsive) and parallel learning (bagging), for example, random forest. Data collection can be a method that can be used to control the display of an AI version with advanced talent and precision processing is faster than enhancing gradients. These are built-in methods for closing the data gap.

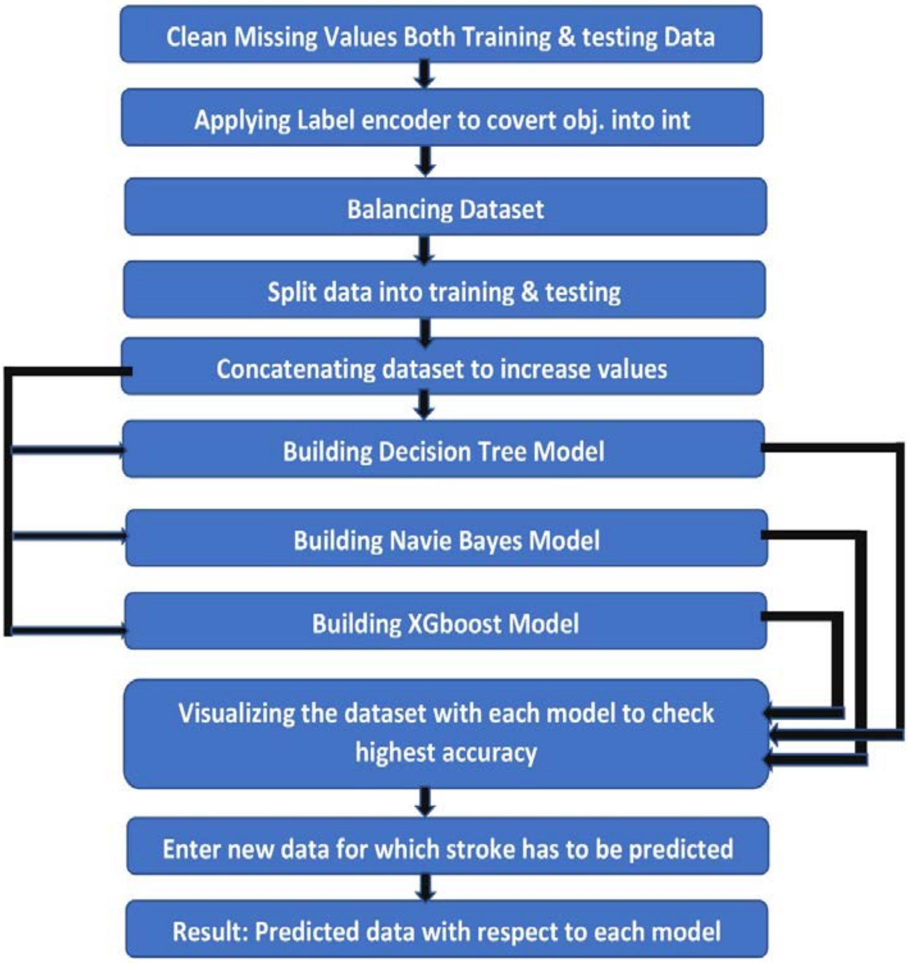
Clean Missing Values Both Training & testing Data
Applying Label encoder to covert obj. into int
Balancing Dataset
Split data into training & testing
Concatenating dataset to increase values
Building Decision Tree Model
Building Navie Bayes Model
Building XGboost Model
Visualizing the dataset with each model to check highest accuracy
Enter new data for which stroke has to be predicted
Result: Predicted data with respect to each model

Figure 1: Proposed Methodology

4. Implementation

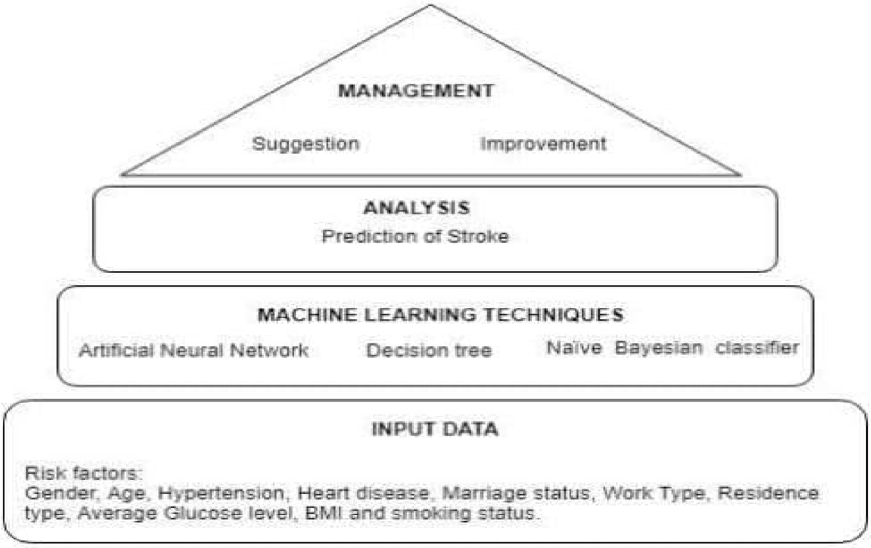
MANAGEMENT
Suggestion
Improvement
ANALYSIS
Prediction of Stroke
MACHINE LEARNING TECHNIQUES
Artificial Neural Network
Decision tree
Naïve Bayesian classifier
INPUT DATA
Risk factors:
Gender, Age, Hypertension, Heart disease, Marriage status, Work Type, Residence type, Average Glucose level, BMI and smoking status.

Figure 2: Steps of Implementation

4.1. Data Preprocessing

After collecting multiple files, process the information. This data set contains a large number of patient records. A total of 5110 + 43400 = 48510 files. 1663 The file is missing some values. The remaining 46,847 records are used for preprocessing. The factors of the informational collection boundaries are prepared. This variable can be utilized to check whether an individual has an extra/diminished danger of a respiratory failure. A cardiovascular failure is in progress, the worth is set to one (1), else, it is zero (0). The outcomes show that 37 of the 297 records have a worth of 1 demonstrating the commonness of focal dead tissue, and the excess 160 segments have a worth of 0, which is more averse to cause a coronary failure. The accompanying boundaries are remembered for the last mathematical informational index. The information record is in CSV design. There are twelve boundaries altogether recorded in underneath table:

Table 1
Dataset features

Feature Name	Description
id	Unique identification number
gender	Male or Female
age	Age of the patient
hypertension	Presence: 0 Absence: 1
heart_disease	Presence: 0 Absence: 1
ever_married	Yes or No
work_type	Children, Government job, Neverworked, Private sector job or Self-employed
Residence_type	Rural or Urban
avg_glucose_level	Patient’s level of glucose
BMI	Body Mass Index of patient
smoking_status	Smoked formerly, never smoked, smokes or unknown
Stroke	0 or 1

4.2. Feature Selection and Reduction

Two of the twelve boundaries are utilized to characterize patient information. 10 inverse boundaries are required. These ten boundaries are basic to the extraordinary and definitive condition of the heart. During the analysis, different types of AI were found, particularly basic numerical strategies like KNN, SVM, XGboost, and irregular timberland. Rehash the test by blunder taking care of many AI techniques with similar properties.

4.3. Classification and Modeling

Since our informational collection is prepared, many AI methodologies can be applied. Any place characterization results are acquired, numerous calculations are chosen, and their presentation is thought about, arrangement and recreation are a significant piece of the framework. In this load of calculations, XGboost gives us exceptionally precise outcomes.

5. Learning Method:

XGboost

The level of formula rule development performance includes accuracy, search, F measurement, and level accuracy. Such metrics are evaluated based on real transaction prices (TP), real negative values (TN), false-positive values (FP), and false negative values. (FN)

Accuracy
$$P = \frac{TP}{TP + FP} \quad (1)$$

Recall
$$R = \frac{TP}{TP + FN} \quad (2)$$

F-Measure
$$F = \frac{2PR}{P + R} \quad (3)$$

The Total Accuracy
$$F = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Table 2
Evaluated values

	XGboost	Random Forest	Navies Bayes	Logistic Regression	Decision Tree
Precision (%)	81.50%	51.34%	66.96%	31.54%	33.18%
Recall (%)	87.00%	92.81%	94.99%	94.52%	95.55%
F1-score (%)	95.20%	66.11%	78.55%	47.30%	49.26%
AUC (%)	97.48%	82.52%	96.64%	71.85%	71.15%

6. Conclusion

This paper presents the execution and correlation of AI based arrangement methods specifically XGboost, Random Forest, Navies Bayes, Logistic Regression and Decision Tree to foresee endurance of a heart patient. The dataset used contains information like patient's gender, age, hypertension, type of work, glucose level, body mass index, etc. to predict his/her chances of death due to heart failure. The performance of the algorithms has been compared using parameters like precision, recall, F1-score and AUC. By the usage of XG Boost, an accuracy of 97.56% is obtained.

7. References

[1] K. N. Manasa, PrinceKumarGupta,DiseasePredictionbyMachineLearningwiththe help of Big Data from Healthcare Communities, International Journal of Engineering Science And Computing (2017).

[2] Nitten S. Rajliwall, Rachel Davey, GirijaChetty. (2018), Cardiovascular Risk Prediction Using XGBoost. Institute of Electrical and Electronics Engineers (IEEE).

[3] SenthilKumar Mohan, ChandrasegarThirumalai, Gautam Srivastava. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Institute of Electrical and Electronics Engineers (IEEE).

[4] ShadmanNashif, Md. RakibRaihan (2018), Heart Disease Detection by Machine Learning Algorithms and Real-Time Cardiovascular Health Monitoring System, World Journal of Engineering and Technology.

[5] "Computer Methods and Programs in the Biomedicine" - Jae–woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho

[6] "Probability of Stroke: A RiskProfile from the Framingham Study" -Philip A.Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B.Kannel

[7] "Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study" - Min SN, Park SJ, Kim DJ, Subramaniyam M, Lee KS