IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

Comparative Analysis of Classification Algorithm for Heart Disease Prediction

Rutuja Dukare, Nisha Malkhede, Shubham Shendkar, Ninad Nirawane,
Prof. H. E. Chaudhari, Dr. M. P. Wankhade
B.E Computer Engineering
Sinhgad College of Engineering, Pune, India

Abstract: *One of the most serious issues we face today because of lifestyle and health decisions is heart disease. The project's primary goal is to anticipate the likelihood of heart disease and the main risk factors for it. We made an effort to pinpoint the risk factors that heart diseases are caused by the most. With the help of machine learning techniques and five different classification algorithms—Logistic Regression, Random Forest, SVM, K-NN, and Naive Bayes. We analyzed the data sets to better understand our datasets and create classification models. We compared all the models and chose the best one. We then developed performance evaluation metrics to generate various parameters to evaluate the classifiers, using the boosting technique to further improve accuracy. On the Cleveland dataset, we noticed that SVM had higher accuracy, but after combining all the datasets, KNN and Random Forest showed similarly effective results. Considering the processing time, we came to the conclusion that KNN was a better option for our project. As a result, the suggested system can tell those who are healthy from those who have cardiac disease.*

Keywords: Heart Diseases, coronary heart ailment, design section, KNN, choice tree, Naïve Bayes, Logistic Regression.

I. INTRODUCTION

Every part of our body receives blood from the heart. The brain and several other organs will stop operating if it isn't functioning properly, and the person will eventually die. As people's lifestyles have changed, stress at work and poor eating practices have contributed to an increase in heart- related diseases. One of the leading causes of death in recent years has been heart attacks.

17.7 million people worldwide die from heart disease per year, or 31% of all fatalities, according to the World Health Organization. In India as well, heart conditions have been a major cause of death [19]. According to the Global Burden of Disease research, which was published on September 15th, 2017, 1.7 million Indians died from heart disease in 2016. India lost $237 billion between 2005 and 2015 as a result of heart- related or cardiovascular illness, according to WHO estimates [20]. Therefore, it is essential to make precise and realistic predictions about heart diseases. Medical organizations gather information on numerous heart-related issues from all across the world. These datasets, however, can be easily investigated using a variety of machine learning techniques because they are far too big for human brains to process. Recent studies have shown that a number of machine learning algorithms can accurately predict whether heart- related disorders would develop or not.

Heart disease early detection would save lives. Multiple heart disease risk factors should be considered and predicted in order to accomplish this goal. We gathered datasets with

14 features, including age, gender, the type of chest discomfort, blood pressure, cholesterol, and so on, to investigate these variables. We create prediction models utilizing various machine learning approaches based on these traits to forecast the occurrence of heart disease. Machine learning methods offer effective information extraction results by creating predictive models from diagnostic medical datasets. By taking information from such data, it might be feasible to forecast which people would get heart disease. During feature extraction, the features are transformed. As some, if not all, important information is lost during the transformation, it is frequently irreversible. In machine learning and statistics, classification is a supervised learning strategy in which the computer program first learns from the input data before applying the learning to classify new observations [22].We employ cross validation to make sure that our model

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

is extracting the correct patterns from the data and isn't picking up too much noise after using feature selection methods to enhance the selection of essential features. Many machine learning techniques can be used to predict heart disease. But choosing the best approach for making predictions based on these traits is very challenging. For the purpose of this study, we thus employ five widely used machine learning algorithms to forecast heart disease. We next employ Ensemble Techniques to increase the accuracy of the results. In the machine learning paradigm known as "ensemble learning," numerous models—often referred to as "weak learners"are trained to address the same issue before being integrated to get better results. The central claim is that we can obtain more trustworthy and/or resilient models by appropriately integrating weak models [23]. The final performance parameter for classifiers that we examine is the execution time. The other six are the classification accuracy, classification error, precision, specificity, and sensitivity. In terms of classification accuracy and execution time, the output of every classifier was assessed uniformly. For creating a high-level intelligent framework for heart illness that accurately distinguishes between those with heart disease and those who are healthy, it suggests which feature method is compatible with which classifier. The suggested machine- learning-based decision support system would be advantageous to physicians since it would improve the accuracy of heart patient diagnoses

II. AIM AND OBJECTIVE

Aim

The main aim of our project is to predict the binary target whether the person is having a heart disease or not .

Objective

- To select and validate algorithms for proposed approach.
- To analyze the performance and compare with existing systems.

III. SCOPE AND LIMITATIONS

A. Future Scope

Additionally, we can test out specific regression models and neural networks using the dataset below to see how they perform. Second, we can test this algorithm on a unique dataset to see how it performs and what problems it encounters during the model validation process. In order to better care for these patients at an early stage and prevent heart failure, this device may be specifically built to anticipate not only the presence or absence of coronary heart disease but also to anticipate the risk factor for heart failure. For the purpose of identifying patients with coronary heart disease and calculating the effectiveness of classifiers for more frequent detection, real-time data from specialized hospitals can be gathered analysis of patients with coronary heart disease.

B. Limitations:

Medical Diagnosis is considered as significant yet intricate task that needs to be considered out precisely.
The automation of the same would be highly beneficial.
Clinical decisions are often made based on doctors' intuition and experienced rather than on the knowledge rich data hidden in database.

IV. LITERATURE SURVEY

[1] Paper Name: Heart Disease Prediction using Machine Learning Algorithm . Author : Apurb Rajdhan , Avi Agarwal , Dr. Poonam Ghuli , Milan Sai , Dundigalla Ravi. Abstract : In recent times, Heart Disease prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart disease. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart disease dataset available in UCI machine learning repository. The proposed work predicts the chances of Heart Disease and classifies patient's risk level by implementing different data mining techniques such as Naive Bayes, Decision Tree, Logistic Regression and Random Forest. Thus, this paper presents a comparative study by analysing the performance of different machine learning algorithms. The trial results

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

verify that Random Forest algorithm has achieved the highest accuracy of 90.16% compared to other ML algorithms implemented .

[2] Paper Name: Comparative Analysis Of Data Mining Classification Techniques For Cardiovascular Disease Prediction. Author : Alpa Makvana , Devangi Kotak. Abstract : Cardiovascular diseases are the main cause of death around the word. Every year, more people die because of these diseases than any other disease. Data mining techniques are widely used for the analysis of diseases, including cardiovascular conditions from person's health data. For such analysis so many techniques are available in data mining. In data mining some of the classification techniques are used to predict the heart diseases. This can be taken as evidence that the proposed method can be used assertively as decision making support to diagnose a patient with cardiovascular disease.

Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms .They used Cleveland dataset and pre- processing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7% .

Santhana Krishnan. J ,et ,al proposed a paper "Prediction of Heart Disease Using Machine Learning Algorithms" using decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables. But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset.

Sonam Nikhar et al proposed paper " Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier.

Avinash Golande et al, proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self arranging guide and SVM (Bolster Vector Machine).

Lakshmana Rao et al, proposed "Machine Learning Techniques for Heart Disease Prediction" in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease. To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.

IV. METHODOLOGY

Data Collection and Preprocessing

The dataset used was the Heart disease Dataset which is a combination of 4 different database, but only the UCI Cleveland dataset was used. This database consists of a total of 76 attributes but all published experiments refer to using a subset of only 14 features [9]. Therefore, we have used the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The complete description of the 14 attributes used in the proposed work is mentioned in Table 1 shown below.

Classification

The attributes mentioned in Table 1 are provided as input to the different ML algorithms such as Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques [12]. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

performance is computed and analyzed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

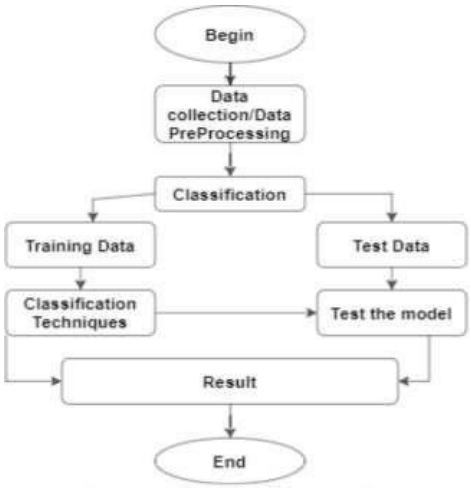

Fig. 1: Generic Model Predicting Heart Disease

Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

Decision Tree

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1. There are 13 independent variables which makes logistic regression good for classification.

Naive Bayes

Naïve Bayes algorithm is based on the Bayes rule[]. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event
(A) given some prior probability of event B represented by P(A/B)[10] as shown in equation 1 :

$$P(A|B) = (P(B|A)P(A)) / P(B) \quad (1)$$

V. RESULT AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree, Naive Bayes and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (4)] tests accuracy.

Precision = (TP) / (TP + FP) (2) Recall = (TP) / (TP+FN) (3)

F– Measure=(2 * Precision * Recall)/(Precision +Recall) (4)

TP True positive: the patient has thedisease and the test is positive.

FP False positive: the patientdoes not have the disease but the test is positive.

TN True negative: the patient does not have the disease and the test isnegative.

FN False negative: the patient has the disease butthe test is negative.

In the experiment the pre-processeddataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 2. The accuracy score obtained for Random Forest, Decision Tree, Logistic Regression and Naive Bayes classification techniques.

VI. ACKNOWLEDGMENT

It gives us great pleasure in presenting the project report on 'Comparative Analysis And Classification Algorithms For Heart Disease Prediction.' I Would to like to take this opportunity to thank my internal Guide Prof. H. E. Chaudhari for giving me all the help and guidance I needed. Their valuable suggestions were very helpful. I am grateful to Prof. M. P. Wankhade, Head of Computer Department, for providing healthy environmental facilities in the department. He allowed us to raise our concern and worked to solve it by extending his cooperation time to time.

VII. CONCLUSION

This paper compares the performances of the class algorithms inside the prediction of coronary heart ailment. It tries to find out the satisfactory classifier for this task. within the experimental dataset, 14 attributes had been used. but all of the attributes aren't similarly emphasized for detecting coronary heart disease. because of this, a characteristic selection technique become presented that gets rid of the inappropriate attributes which are not fairly correlated with the alternative features used for type. each type set of rules offers a noticeable overall performance even as the usage of the selected 10 attributes instead of 14 attributes inside the prediction of coronary heart disease. a number of the studied classifiers, Logistic Regression performs better than other classification algorithms. Binary elegance hassle is solved to perceive whether the affected person has coronary heart disorder or no longer. It's miles endorsed to solve the multiclass problem for detecting coronary heart disorder via dividing coronary heart disorder patients into various classes.

REFERENCES

[1]. Apurb Rajdhan, Avi Agarwal, Milan Sai, Dundigalla Ravi, Dr. Poonam Ghuli, Heart Disease Prediction using Machine Learning, International Journal of Engineering Research Technology (IJERT) 04, April-2020.

[2]. Anagha Sridhar, Anagha S Kapardhi, Predicting Heart Disease using Machine Learning Algorithm, International Research Journal of Engineering and Technology (IR- JET) 04 Apr 2019.

[3]. Nawal Soliman ALKolifi ALEnezi, A Method Of Skin Disease Detection Using Image Processing And Machine Learning, 16th International Learning Technology Conference 2019.

[4]. Neha Prerna Tigga, Shruti Garg, Prediction of Diabetes using Machine Learning, ICCIDS 2019.

[5]. M. Kamber and P. J. Han, Data Mining Concepts, and Techniques, 3rd ed., 2012.

[6]. M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Computational Intelligence Technique for Early Diagnosis of Heart Disease," 2015 IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015

[7]. M. Islam, Y. Elgendy, R. Segal, A. A. Bavry and J. Bian, "Risk prediction model for in-hospital mortality in women with ST-elevation myocardial infarction: A machine learning approach," Journal of Heart & Lung,pp. 1-7, 2017.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 7, May 2023

[8]. P. C. Austin, J. V. Tu, J. E. Ho, D. Levy, D. S. Lee, "Using Methods from Data Mining and Machine Learning Literature for Disease Classification and Prediction: a Case Study Examining Classification of Heart Failure Subtypes," Journal of Clinical Epidemiology 66 (2013) pp. 398-407, 2013.