e-ISSN: 2582-5208

International Research Journal of Modernization in Engineering Technology and Science

(Peer-Reviewed, Open Access, Fully Refereed International Journal)

Volume:04/Issue:06/June-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTION OF HEART DISEASE USING MACHINE LEARNING TECHNIQUES

Prasad Thombare*1, Madhvaj Ghalme*2, Saurabh Raut*3,
Narendra Dhakne*4, Ms. Poonam R. Dholi*5

*1,2,3,4Student, Department Of Computer Science And Engineering, Matoshri College Of Engineering And Research Centre, Nashik, Maharashtra, India.

*5Project Guide, Department Of Computer Science And Engineering, Matoshri College Of Engineering And Research Centre, Nashik, Maharashtra, India.

ABSTRACT

Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, KNN and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Keywords: Machine Learning, Heart Disease, Random Forest Classifier, Gradient Boost Classifier, K-Nearest Neighbors, Dataset.

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to heart disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. This project aims to predict future heart disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

II. LITERATURE SURVEY

With growing development in the field of medical science alongside machine learning various experiments and researches have been carried out in recent years releasing the relevant significant papers. The paper [1] proposes heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factors SMO (89KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms is still not satisfactory. So that if the performance of accuracy is improved more to give batter decision to diagnosis disease. [2]In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10- fold Cross Validation, considering 13 attributes, implementing 3 different algorithms, they concluded Gradient Boost and Random Forest gave the maximum accuracy of 74.0 percent. [3]Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87 Classifier: 84.

III. METHODOLOGY

The proposed system architecture will give an overview of the working of the system. Working of the system starts with the Gathering of data and selecting the important attributes that will give efficient accuracy. After that the required data is preprocessed into the required format that is suitable for machine learning algorithms.

After preprocessing of data, the data is then divided into two parts that are: training data and testing data. The algorithms are applied and the algorithm is trained using the training data. The accuracy is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Dataset Collection
2.) Attributes Selection
3.) Data Pre-Processing
4.) Data Balancing
5.) Disease Prediction

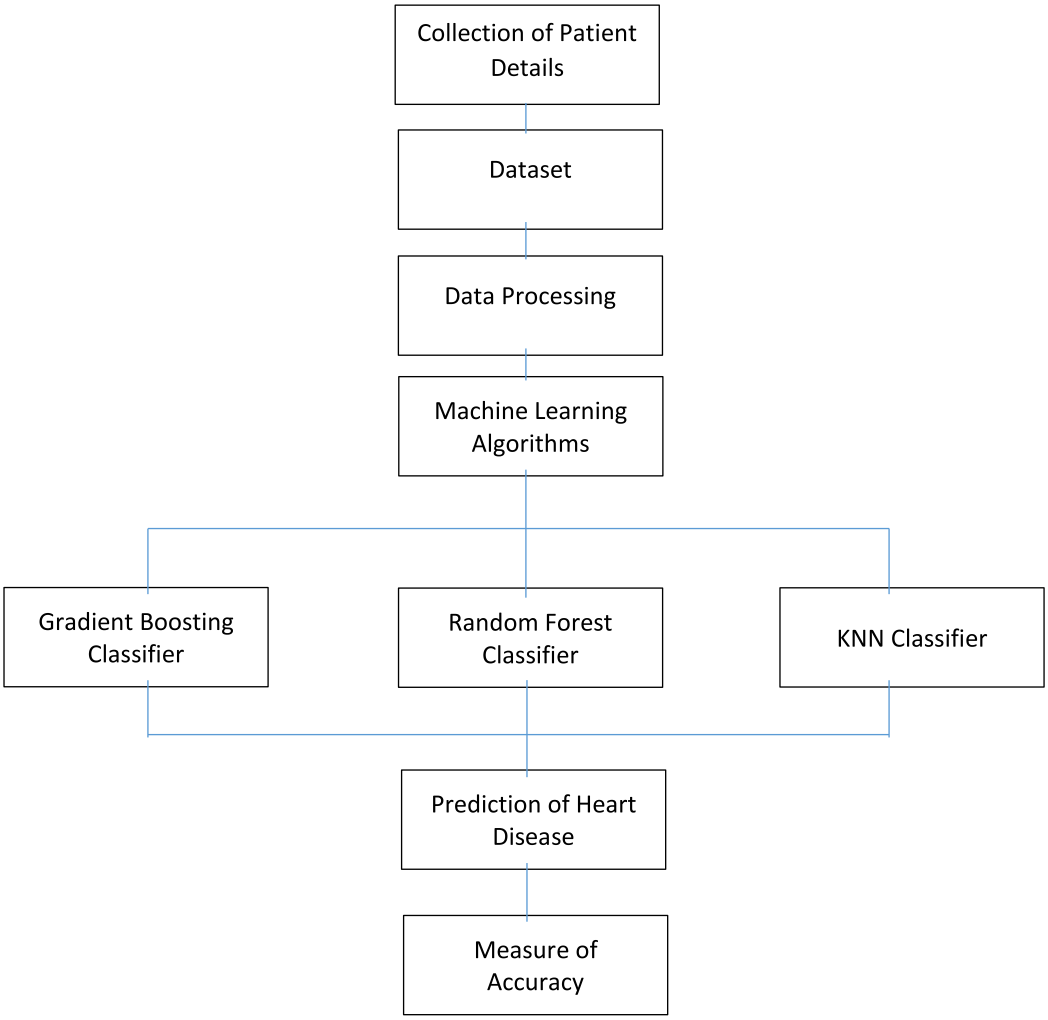
Collection of Patient Details
Dataset
Data Processing
Machine Learning Algorithms
Gradient Boosting Classifier
Random Forest Classifier
KNN Classifier
Prediction of Heart Disease
Measure of Accuracy

Figure 1: Flowchart of Proposed System.

1) Dataset Collection:

First, we collected a dataset for the prediction of heart disease. After collecting the dataset, we divided the dataset into training data and testing data. The training data of dataset is used for prediction model learning and testing data is used for evaluating and analyzing the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset that we used for our project is Heart Disease UCI. Out of 76 attributes that are present related to heart which we select 14 attributes to use for the system.

2) Attributes Selection:

Attribute or Feature selection includes the selection of appropriate attributes that will give efficient accuracy for the prediction system. This is used to increase the efficiency and accuracy of the system. Several parameters of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, etc., are selected for the prediction of our system.

3) Data Pre-Processing:

Data pre-processing is an important step for the creation of a system while using machine learning models. At first, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format that is compatible with model. It deals with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like data reduction and data transformation etc. Preprocessing of data is required for improving the accuracy of the model.

4) Data Balancing:

Data balancing is very important in system that dealt with large amount of data. So, we make sure our dataset is balanced. In case it is imbalanced, imbalanced datasets can be balanced in two ways. They are Under Sampling and Over Sampling 1. Under Sampling: In this sampling dataset balance is done by the reduction of the size of the ample class. This process is done when the data is adequate. 2. Over Sampling: In this sampling, dataset balance is done by increasing the size of the scarce samples. This process is done when the data is inadequate.

5) Disease Prediction:

Various machine learning models like KNN, Gradient Boosting and Random Forest are used for classification. Comparative analysis is performed among algorithms and the algorithm with highest accuracy is used for heart disease prediction in this proposed system.

IV. MACHINE LEARNING ALGORITHMS

Machine learning is a powerful technology that is a systematic study of various algorithms that provide the system with the potential to replicate human learning activities without being actually programmed. The machine learning models used in our project are:

1) Gradient Boosting Classifier:

Gradient boosting machines are a family of powerful machine-learning techniques that have shown considerable success in a wide range of practical applications. They are highly customizable to the particular needs of the application, like being learned with respect to different loss functions. In gradient boosting machines, or simply, GBMs, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principal idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squarederror loss, the learning procedure would result in consecutive error-fitting. In general, the choice of the loss function is up to the researcher, with both a rich variety of loss functions derived so far and with the possibility of implementing one's own task specific loss. This high flexibility makes the GBMs highly customizable to any particular data driven task. It introduces a lot of freedom into the model design thus making the choice of the most appropriate loss function a matter of trial and error. However, boosting algorithms are relatively simple to implement, which allows one to experiment with different model designs.

2) Random Forest:

Random Forest classifier algorithm is one among the supervised learning technique in machine learning algorithms. It is based on ensemble learning which is the method of combining various multiple classifiers to resolve a complicated problem to improve the performing nature of the model. Random Forest improves the prediction accuracy of the dataset which consists of several decision trees on various subsets of the given dataset by taking the average. Random forest not relay on one decision tree. Rather than looking forward to a one decision tree, the random forest acquires the prediction from each single tree, and supported the bulk of votes for predictions, it predicts the ultimate output by taking average. The upper the trees, the upper the accuracy. And also prevents the matter of over fitting. The ultimate output is taken based on using the bulk voting classifier for a classification problem wiyhin the case of a regression problem the ultimate output is relies on the mean of all the outputs.

3) K-Nearest Neighbor:

K-Nearest Neighbor algorithm is one among the simplest and easiest Machine Learning algorithms supported Supervised Learning technique. The K-NN algorithm will think about the identical content between the new data and already available data and place the new data into the category that's most just like the available categories. Based on similarity of data, KNN algorithm stores the data that is available and then classifies the data. This suggests when new data appears we are ready to easily classify it into a most suited category by using K- NN algorithm. The K-NN algorithm can also be used for Regression in addition as for Classification. But most of the days Classification problems are make use of it. ¬ K-NN is additionally called as non-parametric algorithm, which implies it doesn't make any particular assumption on underlying data. ¬ And it's also called as lazy learner algorithm because it doesn't learn from the training set immediately. The KNN algorithm in the training phase which is training data apart from testing data just when it gets the new data it just stores it, then it classifies that data into a category that's rather more just like new data.

V. RESULTS AND DISCUSSION

For Comparative study we are proposed to train 3 different models and test whether which model is given higher accuracy against the same dataset.

Table 1. Algorithm Accuracy Table

Algorithms	Accuracy
KNN	0.74
Random Forest	0.72
Gradient Boost	0.74

After observing the above results, we conclude that Gradient Boosting Algorithm and KNN has higher accuracy than Random Forest algorithm.

VI. CONCLUSION

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease. Average prediction accuracy probability of 74 Percent is obtained.

VII. REFERENCES

[1] A. H. M. S. U. Marjia Sultana [Analysis of Data Mining Techniques for Heart Disease Prediction 2018].

[2] M. I. K. A. I.S. Musfiq Ali [Heart Disease Prediction Using Machine Learning Algorithms].

[3] M. A. K. S. H. K. M. A. V. P. M Marimuthu, [A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach].

[4] K. Bhanot, toward data science.com, https://towardsdatascience.com/predicting-presence-of-heart-diseases-usingmachinelearning-36f00f3edb2c.

[5] Senthil Kumar Mohan, Chandrasekar Thirumalai and Gautam Srivastva, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access 2019.

[6] M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT 2019.

[7] M. A. Jabbar, P. Chandra, and B. L. Deekshatulu, "Prediction of risk score for heart disease using associative classification and hybrid feature subset selection," Int. Conf. Intell. Syst. Des. Appl. ISDA, pp. 628–634, 2012.