Heart Diseases Prediction using Machine Learning

Shaik Wajid[1], Gundla Sowmya[2], Maruthi Sravan Reddy[3], Mohammed Farooq[4]
[1,2,3]*CSE-AIML, Sphoorthy Engineering college, Hyderabad, India*
[4]*Asst Professor, CSE-AIML, Sphoorthy Engineering college, Hyderabad, India*

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

Keywords—Machine learning, Logistic regression, Heart disease, Support vector machine, accuracy

INTRODUCTION

Cardio-vascular diseases are the primary cause of death worldwide over the past decade. According to the World Health Organization it is estimated that over 17.9 million deaths occur each year because of cardiovascular diseases and out of these deaths 80% is attributed to coronary artery disease and cerebral stroke [1]. Many habitual factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are often deciding factors for heart disease. The efficient, accurate and early medical diagnosis of heart disease plays a pivotal role in taking preventive measures to avoid the complications that arise due to such diseases.

The major challenge faced in the world of medical sciences today is the provision of quality service and efficient and accurate prediction. The later problem can be solved by automation with the help of Data Mining and Machine Learning. Data mining is defined as a process used to extract usable data from a large set of raw data. It implies analyzing

patterns in large batches of data by making use of various software. It also involves effective data collection and warehousing coupled with computer processing. Machine learning which is subfield of data mining that deals with large scale well-formatted data efficiently. In the machine learning can be used for diagnosis, detection and prediction of various disease.

Various Machine Learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machine, KNearest Neighbour, Decision Tree, Random Forest and the ensemble technique of XGBoost are compared to find the most accurate model. Here the heart disease dataset from the UCI repository is used. In this research a discussion and comparison of the existing classification techniques is made. The paper also mentions scope of future research and different advancement possibilities.

RELATED WORK

There are numerous works has been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centres.

Avinash Golande and et. al.; studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared [1]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

T. Nagamani, et al. have proposed a system [2] which deployed data mining techniques along with the MapReduce algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due use of dynamic schema and linear scaling.

Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms [3]. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random Forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.

Anjan Nikhil Repaka, ea tl., proposed a system in [4] that uses NB (Naïve Bayesian) techniques for classification of dataset and AES (Advanced Encryption Standard) algorithm for secure data transfer for prediction of disease.

Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (KNearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analysed for different number of attributes [5].

Yu-Xuan Wang, et.al. Have explored different applications that demonstrated the significance of the ML methods in various areas [9]. They proposed a new technique for the designing of a working framework. The approach used the distinct machine learning procedures. After getting the proper result from the data miner, the whole information assembled from the structure was inspected. In light of the various tests, it was seen that proposed approach gave proficient results.

Zhiqiang Ge, et.al, (2017) proposed a work on analytics and data mining applications, which was done prior. These procedures were used in business area for various purpose of perspectives. Here they have explored 8 unsupervised and 10 supervised learning algorithms [10]. In their research, they showed an application work for the semi-supervised type learning algorithms. In industry method, it was seen that roughly 90%-95% applications utilized both the unsupervised and supervised machine learning procedures. Consequently, it was portrayed that the Machine Learning methods play an indispensable part in the planning of different novel applications for domains like medical services and industry.

Nagaraj M Lultimate, et al., has performed the heart disease prediction using Naive bayes classification and SVM (Support Vector Machine). The performance measures used in analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes [6].

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs as shown in Table 1. We analysed the classification algorithms namely Decision Tree, Random Forest, Logistic Regression and Naive Bayes based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

DATASET INFORMATION

The proposed work predicts heart disease by exploring four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart disease. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart disease. Fig. 1 shows the entire process involved.

The main objective of this research is to develop a heart disease prediction system using machine learning. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set. Heart disease prediction system using machine learning aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

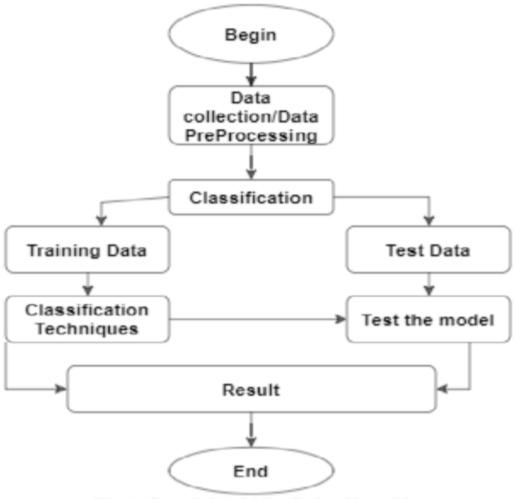
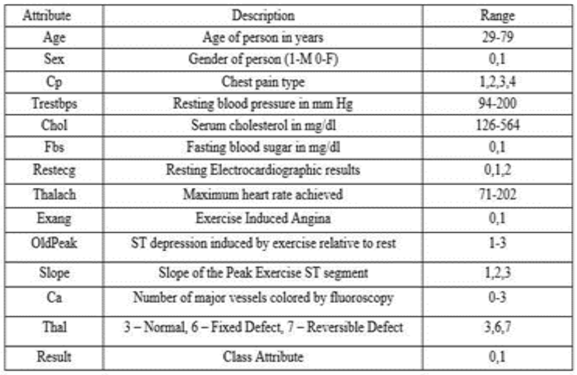


Fig. 1: Generic Model Predicting Heart Disease

Data Collection

In the data collection step for heart disease prediction using machine learning, relevant datasets containing patient information are gathered. These datasets may come from healthcare databases, research studies, or other sources. The collected data typically includes features such as age, gender, cholesterol levels, blood pressure, and other medical indicators. It's essential to consider ethical considerations and implement data privacy measures during this process. Once the data is collected, it undergoes exploratory analysis (EDA) to understand its distribution and characteristics before further preprocessing steps.

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 Table 1 shown below. Kaggle website for our analysis. The complete description in of the 14 attributes used in the proposed work is mentioned



Attribute	Description	Range
Age	Age of person in years	29-79
Sex	Gender of person (1-M 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the Peak Exercise ST segment	1,2,3
Ca	Number of major vessels colored by fluoroscopy	0-3
Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3,6,7
Result	Class Attribute	0,1

Fig.2:Feature

Data preprocessing

Data preprocessing is a crucial step in heart disease prediction using machine learning. It involves cleaning and transforming raw data to make it suitable for training a predictive model. These preprocessing steps are essential for building an accurate and robust heart disease prediction model, as they ensure the quality and integrity of the data used for training and testing the machine learning algorithm.

Training

The training phase extracts the features (independent variables) from the dataset and the testing phase (containing dependent variables) is used to determine how the appropriate model behaves for prediction. We have divided the dataset into two sections. These are the training and testing phases. We have split the dataset into 90% training are listed as below 1. For initializing the fixed internal random number generator, we use the random state parameter which will decide the splitting of data into train and test indices. Setting a random state will guarantee a fixed value that the same sequence of random numbers will be generated each time the code is being run. Setting random state, a fixed value will guarantee that the same sequence of random numbers is generated each time we run the code. Then we scaled the data using Standard scattered and fitted the training and testing data using 'fit. transform'.

Classification

The classification step in heart disease prediction using machine learning involves training model to categorize individual into different classes based on their risk of heart disease. The model is trained on a dataset with labelled examples, where each example include feature such as age, blood pressure, cholesterol levels, etc., and a corresponding label indicating whether the individual has heart disease or not. Once trained, the model can then predict features, aiding in early detection and intervention. 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] etc. The input dataset is split into 70% of the training dataset and the remaining 30% 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 performance is computed and analysed based on different metrics used such as accuracy, precision, recall and for measure scores as described further. The classification step is pivotal in developing a predictive model that can assist in identifying individual at risk of heart disease based on their unique health characteristics. The goal is to create an accurate and reliable model that can contribute too early detection and proactive healthcare intervention. The different algorithms explored in this paper

Random Forest: Random Forest algorithm is a supervised classification algorithmic technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the greater number of trees give higher accuracy. The three common methodologies are:

- Forest RI (random input choice)
- Forest RC (random blend)
- Combination of forest RI and forest RC

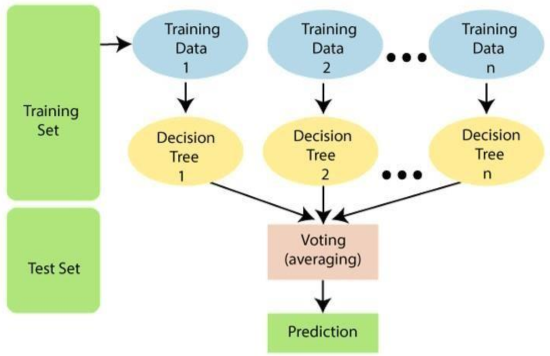


Figure : Working of the Random Forest algorithm

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.

Support Vector Machine (SVM): Support
Vector Machine [11] is a classification technique of Machine learning to, which is used to analyse data and discover patters in classification and regression analysis. SVM is typically mull over when data is characterized as two class problem. In this strategy, data is characterized by finding the best hyper plane that isolates all data points of one class to the other class. The higher separation or edge between the two classes is, the better is the model, considered. The data points lying on limit of the margin are called as support vectors. The actual basis of SVM is mathematical methods used to design complex real-world problems. We have chosen SVM for this experiment because our dataset - Cleveland Heart Disease Dataset CHDD has multi class to predict based on various parameters. In SVM, the mapping of training data is to be done with a function called kernel (Kernels of SVM), these are linear kernel, quadratic kernel, polynomial kernel, Radial Basis Function kernel, Multilayer Perceptron kernel, etc. Apart from the kernel's functionalities in SVM, few more methods are available such as quadratic programming, sequential minimal optimization, and least.

Decision Trees: Decision Tree algorithm [12] in Machine Learning is used to develop the Classification models. This classification model is based on the tree-like structure. This comes under the category of supervised learning, where the target result is already known. Both the categorical and numerical data can be applied on Decision tree algorithm. Decision tree consists of root node, branches and leaf nodes. Data is evaluated on the basis of traversing path from the root to a leaf node. For our dataset - CHDD, a total of 283 tuples were assessed down the decision tree. They potentially came to a positive or negative assessment for the heart disease prediction. These were compared to the actual parameters to check for the false positives/false negatives which show the accuracy, specificity, and sensitivity of the model.

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 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:

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

Testing: In the context of heart disease prediction using machine learning, the testing step involves evaluating the performance of the trained model on a separate dataset that it has not seen during the training phase. This step is crucial for assessing the model's ability to generalize well to new, unseen data. The testing step helps validate the model's efficacy, reliability, and ability to make accurate predictions on new cases of heart disease. It is crucial stage in the machine learning pipeline to ensure that the developed model is ready for deployment in practical healthcare scenarios.

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

$$\text{Precision} = (TP) / (TP + FP)$$

$$\text{Recall} = (TP) / (TP+FN)$$

$$\text{F–Measure} = (2 * \text{Precision} * \text{Recall})/(\text{Precision}+\text{Recall})$$

- TP True positive: the patient has the disease and the test is positive
- FP False positive: the patient does not have the disease but the test is positive
- TN True negative: the patient does not have the disease and the test is negative
- FN False negative: the patient has the disease but the test is negative

CONCLUSION

The overall purpose is to describe the various ML techniques that are useful in predicting heart disease. Effective and accurate predictions with a small number of characteristics and evaluation of the purpose of this study. The data was previously processed and used in the model. K-Nearest Neighbor with 89%, Random Forest with 82% and Support Vector Classifier with 81% algorithms working very well. However, the Decision Tree Classifier provides a slight accuracy of 79%. We can continue to expand this research that integrates other ML strategies such as time series, integration rules and integration with other integration strategies. Considering the limitations of this study, there is a need to use complexity and combination of models to achieve high accuracy in predicting early heart disease. The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health's diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research.

REFERENCE

[1] Avinash Golande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, pp.944950,2019.

[2] T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with MapReduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, Design and Implementation Heart Disease Prediction Using Naives Bayesian, International Conference on Trends in Electronics and Information (ICOEI 2019).

[5] Theresa Princy R, J. Thomas, Human heart Disease Prediction System using Data Mining Techniques, International Conference on Circuit Power and Computing Technologies, Bangalore, 2016

[6] Nagaraj Lutimath, ChethC, Basavaraj S Pol., Prediction of Heart Disease using Machine Learning, International journal Of Recent Technology and Engineering,8, (2S10), pp 474-477, 2019

[7] UCI, Heart Disease Data Set. [Online]. Available (Accessed on May 2020): https://www.kaggle.com/ronitf/heartdisease-uci.

[8] Sayali Ambekar, Rashmi Phalnikar, Disease Risk Prediction by Using Convolutional Neural Network,2018 Fourth International Conference on Computing Communication Control and Automation.

[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, Medical Data Mining for Heart Diseases and

the Future of Sequential Mining in Medical Field, in Machine Learning Paradigms, 2019, pp. 7199.

[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series,2018.

[11] Fajr Ibrahem Alarsan., and Mamoon Younes Analysis and classification of heart diseases using heartbeat features and machine learning algorithms, Journal of Big Data,2019;6:81.