

Cardiovascular Disease Prediction using Machine Learning Algorithms

Nirmala M.B¹, Haarika Reddy K.R², Mohit Sah³, Shrinidhi Shastry⁴, and Veeksha V Murthy⁵

Dept. of CSE, SIT, Tumakuru, Karnataka, India

Email: mbnirmala@sit.ac.in, 1si17cs041@sit.ac.in, 1si17cs060@sit.ac.in, 1si17cs105@sit.ac.in, 1si17cs128@sit.a.in

Abstract—The world which we see today is getting advanced each year, this is due to some of the most promising factors such as development in the field of science and technology, industrialization, globalization, and various other aspects. Considering the remarkable development in surroundings we can also see a drastic change in the human's health. Humans are facing a lot of health issues because of the modern food and the lifestyle they have adopted. Junk food and sedentary working culture are the major causes of many diseases. Heart-related diseases are one of the hazardous diseases the world is facing today. 17 million deaths worldwide are due to Cardiovascular Disease (CVD) which contributes to 31%. With the increase in heart diseases, there are several pieces of equipment and sophisticated machines to predict the heart condition. This paper mainly concentrates on a system where a patient can easily check the condition of their heart. The proposed system considers some of the famous algorithms of Machine Learning which includes Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest. Cleveland dataset from UCI repository with 14 attributes and 303 instances is used. The patient just needs to provide some of the information as input and the system will perform predictions on the condition of the heart and other factors.

Index Terms— Machine learning, classification algorithms, prediction model, and heart disease prediction.

I. INTRODUCTION

Heart disease generally called cardiovascular disease has been one of the most hazardous diseases. The death rate due to heart-related diseases has been increasing for decades. By the statistics of WHO (World Health Organization), the number has reached around 17.9 million people die every year. This is due to the sedentary lifestyle and people's ever-changing food habits. Cardiovascular diseases comprise both heart and blood vessels disorder. These include rheumatic, cerebrovascular, stroke, and coronary heart disease can be classified as habitual and physiological risk factors. The habitual risk factors are smoking, overuse of alcohol and caffeine, stress. Tobacco smoking increases the risk of dying by 2-3 folds. Physical inactivity along with other physiological factors like raised blood pressure, obesity, lipids, overweight, glucose, hypertension, high blood cholesterol, and pre-existing heart problems affects the heart condition.

The data collected from various fields such as Manufacturing Engineering, Financial Banking, Corporate Surveillance, Criminal Investigation, and many more are processed using data mining for the extraction of required data from the datasets. Machine Learning is one of the booming domains in the field of AI (Artificial

Intelligence) by which the discrete information from various datasets can be handled and analyzed to predict the desired output. Prediction of patient's disease in early-stage and if treated well in advance can avoid menacing circumstances. Machine learning algorithms such as Naïve Bayes, K-nearest Neighbors (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM) are used to classify the attributes and analyze them to predict heart diseases. The factors affecting efficiency are biasness and variance of the dataset. KNN has a problem of overfitting caused due to high variance and low biasness. High biasness and low variance are always advantageous for small datasets as it takes low computing time for training and testing, considering Naïve Bayes for small datasets, it works efficiently with high biasness and low variance.

II. LITERATURE SURVEY

The authors of [1] have implemented algorithms such as KNN, SVM, Decision Tree, and Linear Regression. The Decision tree is a nonparametric algorithm technique that has an overfitting problem but can be resolved using removable methods. Support vector machine used for classifying the datasets as it has a statistical and algebraic background. Authors[1] used the Cleveland dataset and have considered 73% for training and 37% for testing. The accuracy of K-Nearest Neighbor is 87%, Support Vector Machine is 83%, and Decision Tree is 79%.

The aim of the Authors[2] is to provide the doctors with tools and techniques to diagnose heart disease at an early stage. Machine learning plays a vital role in identifying discretely hidden patterns and analyzing the available data. And the Authors[2] concentrated on providing an efficient algorithm with high accuracy for predicting whether the patient is suffering or not from heart disease. The input to the system is the patient's report of health which contains different attributes. The accuracy of Random Forest is 90.16%, Logistic Regression is 85.25%, and Naïve Bayes is 85.25%.

Authors[3] uses the computer to understand non-linear and complex interactions among various attributes or factors for prediction. Thus reducing the errors in prediction and producing truthful outcomes. Data mining explores large datasets to conclude past collected or repository data for future prediction. WEKA tool was used for pre-processing the data, which was in ARFF (Attribute Relation File Format) format. Out of 76 various attributes, only 14 different attributes were considered for better accuracy. The accuracy of Naïve Bayes is 88.15%, K-Nearest Neighbor is 90.78%, Decision Tree is 80.26% and Random Forest is 84.21%.

A large amount of hidden and raw data are collected from the health industry which can be used to make efficient decisions by applying data mining techniques. Also, there is a need for an effective analyzing tool to analyze the hidden and raw data, and data mining algorithms can be used to develop such systems for analyzing and classifying the data and to detect, analyze and predict the heart diseases from which the patient is suffering. To find the heart disease there is a need for some tests to be done on the patient. However, by using the data mining technique we can reduce the number of tests, doing this significantly adds to the performance and time. The dataset considered by Authors[4] is from the UCI repository which contains the Cleveland heart disease dataset. 14 out of 76 attributes with 303 instances were used.

The patient will be in a critical condition because of no early diagnosis of the disease, but this can be solved using Machine Learning. Machine learning algorithms used by Authors[5] are Support Vector Machine, Decision Tree, Naïve Bayes, and Neural Networks. The accuracy of Support Vector Machine (SVM) is 92.1%, Decision Tree is 89.9% and Naïve Bayes is 84.4%.

Prediction of the heart condition is one of the challenging tasks for Doctors. Authors[6] mainly concentrate on designing an automated system that can accurately predict the heart condition by using some attributes as input and also considering some of the famous Machine Learning algorithms. The accuracy of Support Vector Machine is 81.54%, Naïve Bayes is 80.43% and Logistic Regression is 82.89%.

In the field of medicine, the diagnosis of heart disease has been a crucial task for doctors and patients. The diagnosis of an individual is based on the patient's report of all previous health checkups and tests. The emerging improvement in the machine learning field has made the developing automated intelligent systems that improve the accuracy of prediction in the field of medicine which helps in making better decisions. The authors of [7] use a system to save individual life by consuming less time for prediction and treatment. Various classification techniques are used to achieve an accurate and better diagnosis. One of the classification methods is Neural Networks that can operate as the brain of humans and draw relationships among huge amounts of data in a dataset. The attributes were from ECG and symptoms used to perform prediction on the presence of heart disease. The accuracy obtained using the Cleveland dataset is as follows The backpropagation algorithm has 97.5%, the hidden layer with 20 neurons has 98.58% and the hidden layer with 5 neurons has 93.39%.

Data mining is one of the computer-based data processing in the field of medicine. It helps in finding the hidden different patterns of various related datasets. This can be used to diagnose heart disease. The data available will

always be in heterogeneous form. It is very difficult to draw a conclusion using such a huge volume of data. Hence organizing data is necessary. And an automated system from Authors[8] with different algorithms to diagnose effectively with more accuracy for individual clinical tests and reports. Authors[8] proposed an algorithm with an efficiency of 88.4% whereas Naïve Bayes has 75.8%, Support Vector Machine has 86.1%, Random Forest has 86.1%, and Decision Tree has 85.0%.

The diagnosis of heart disease can be done by using different machine learning and data mining techniques. Authors[9] used the feature selection method to improve the accuracy of the performance of the algorithm. Based on the diagnosis of the type of heart disease, many curable methods can be used for that patient. Using predictive model techniques can only give marginal success. Hence more complex and combinational models are needed for better accuracy in prediction at an early stage of heart disease. The system gets more intelligent as the amount of data fed into the system increases. There is a need for future work in improving accuracy and scalability. This can be done by using various discretization methods. The accuracy of Naïve Bayes is 82.4%, Support Vector Machine is 81.7%, and Random Forest is 77.9%.

Heart disease prediction is one of the quintessential fields in the medical field. Heart disease is also one of the most fatal in the world if not treated well. It describes the changes that occur in the heart that leads to this fatality. Another major challenge is Quality of Service (QoS). Data mining is the process of determining unknown hidden patterns of the data set. Authors[10] accuracy of Naive Bayes is 52.33%, Decision Tree is 52% and K-Nearest Neighbor is 45.67%.

TABLE I. ACCURACY PERCENTAGE OF THE MACHINE LEARNING ALGORITHMS FROM DIFFERENT RESEARCH PAPERS

references	Accuracy of algorithms
[1]	• SVM -83%. • Decision Tree -79%. • KNN -87%.
[2]	• Logistic Regression -85.25%. • Random Forest -90.16%. • Naïve Bayes -85.25%.
[3]	• Naïve Bayes -88.10%. • KNN -90.78%. • Random Forest -84.21%.
[4]	• Naïve Bayes -82.4%. • SVM-81.7%. • Random Forest -77.9%.
[5]	• Naïve Bayes-84.16%. • SVM -92.1%. • Decision Tree-89.9%.
[6]	• Naïve Bayes-80.43%. • SVM -81.57%. • Logistic Regression-82.89%.
[7]	• Logistic Regression -85.20%. • Naïve Bayes -90.20%. • Random Forest-85.20%.
[8]	• SVM -86.10%. • Decision Tree-85.0%. • Random Forest -86.10%.
[9]	• Naïve Bayes-82.4%. • SVM -81.7%. • Random Forest-77.9%.
[10]	• Naïve Bayes -52.33%. • Decision Tree -52%. • KNN -45.67%.

III. IMPLEMENTATION

Figure 1 describes the prediction model which follows a specific way of implementation starting from loading the dataset followed by pre-processing the data which includes importing libraries, finding missing data, and

encoding categorical data, next with splitting the dataset, applying the algorithms like Logistic Regression, Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, then analyzing the results, and modifying the model to obtain the highest possible accuracy.

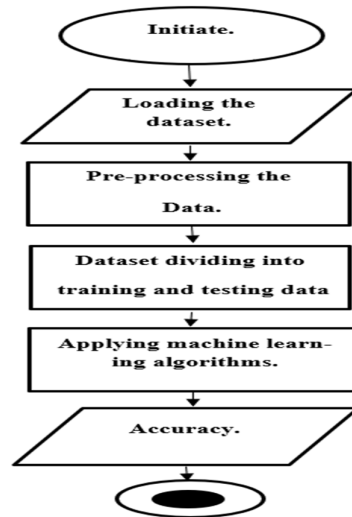


Figure 1: A prediction model

The dataset used for prediction is 'Cleveland' taken from the UCI repository, which has the following attributes:

- **Age:** Individual age.
- **Gender:** Gender of individuals following the format: 1 = male, 0 = female.
- **Resting Electrographic:** It has values of 0: normal, 1: abnormality of ST-T wave, 2: hypertrophy of left ventricle.
- **Serum Cholesterol:** Taken in terms of mg/dl (standard unit).
- **Blood Sugar:** Compared with 120mg/dl. 1: if it is >120mg/dl, else: 0.
- **Exercise-induced angina:** 1: yes, 0: no.
- **Chest pain:** Type of chest pain experienced by the individual and defined by 1: typical angina, 2: atypical angina, 3: non - anginal pain, 4: asymptotic.
- **Thalach:** Maximum heart rate achieved by an individual.
- **Slope:** Peak exercise ST segment, 1: up sloping, 2: flat, 3: downsloping.
- **The number of major vessels:** (0-3) colored by fluoroscopy.
- **Thal:** Thalassemia with 3: normal, 6: fixed defect, 7: reversible defect.
- **Old Peak:** ST depression induced by exercise relative to rest.
- **Rest Blood Pressure:** Taken in terms of mmHg.
- **Diagnosis of heart disease:** Outputting whether the individual is suffering from heart disease or not: 0 = absence 1,2,3,4 = present.

From Figure 1, Pre-processing the data considers importing the required libraries and processing the data, as few records of data might be missing. The missing data can be filled either by mean value or by ignoring it. The categorical data is converted into numerical data for prediction. From the dataset considered above, Gender is categorical data which is transformed into numerical by the following format: 0 for female and 1 for male and few other attributes such as defining the type of chest pain, slope.

The data is split as training, to feed to the algorithm and testing, to calculate the precision for the algorithms in the ratio of 80:20.

The following are the algorithms applied

A. Logistic Regression

Logistic regression is one of the well-known Machine Learning algorithms which come under the Supervised Learning technique. It is mainly concentrating on predicting the dependent variables using the independent variables and the output is a probabilistic value i.e., 0 or 1. Logistic Regression is much similar to Linear

Regression but Regression problems are solved using Linear Regression whereas classification problems are solved using Logistic regression.

Figure 2 depicts a ROC (Receiver Operating Characteristic) curve which offers the overall performance of a class version for any class thresholds. Figure 3 is the regplot() characteristic gives the plot of fit by using true or false from the data provided and also permits to specify whether or not to estimate the logistic regression model. Figure 4 depicts the mxm (m-represents range of classes) confusion matrix is used for the assessment of the class model based on the known true values known.

B. Naïve Bayes

Naïve Bayes is a classification algorithm that assumes that the existence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes is quite simple and can be easily built for analyzing large datasets. It is fast in prediction and performs excellently on multiple classes.

The following equation represents the Naïve Bayes.

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

- $P(x|y)$ – class posterior probability.
- $P(x)$ – class probability.
- $P(y|x)$ – predictor class probability.
- $P(y)$ – predictor probability.

C. Support vector Machine

Support Vector machine is another algorithm of Machine learning which can be used for both Classification and Regression, but usually, it is used for Classification problem. SVM (Support Vector Machine) mainly concentrates on creating the best decision boundary that can separate n-dimensional spaces into classes. Support Vector Machine considers support vectors to point to the best decision boundary.

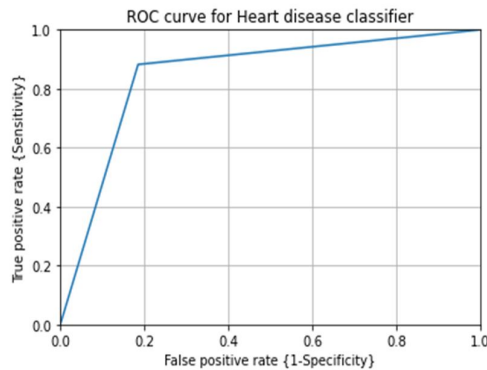


Figure 2: ROC curve

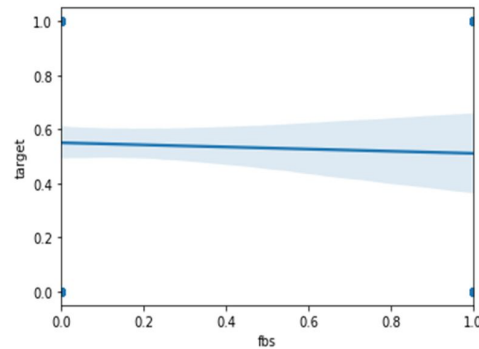


Figure 3: Regplot

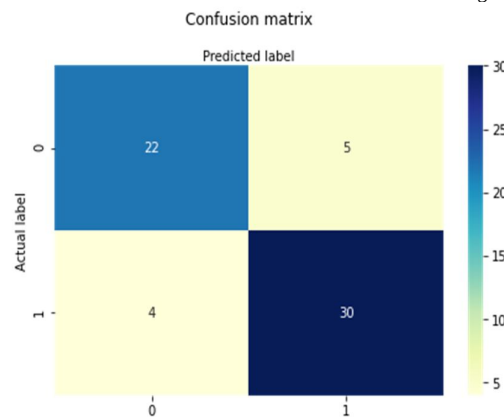


Figure 4: Confusion matrix

D. Decision Tree

The decision tree is a type of supervised machine learning algorithm which covers both Regression and Classification. A decision tree has a tree structure where the internal nodes represent a dataset. The decision rules are represented by the branches and the outcome is represented by each leaf node. A decision tree has two types of nodes i.e., decision node and leaf node. Decision trees are used to make decisions and the leaf nodes represent the output. The CART algorithm is used to build the tree. CART stands for Classification and Regression Tree algorithm.

E. Random Forest

Random Forest follows the supervised learning technique. It can be used for both Classification and Regression problems in Machine Learning. Random Forest work 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. The random forest has the least training time and it predicts the result with high accuracy and high efficiency.

IV. RESULTS

The accuracy obtained by applying the algorithms is stated in Table 2 and the pictorial representation in Figure 5.

TABLE II: ACCURACY OBTAINED

Algorithm	The accuracy achieved (in %)
Logistic Regression	85.25
Naïve Bayes	85.25
Decision Tree	81.97
Support Vector Machine	81.97
Random Forest	95.08

From Table 2, the accuracy of Logistic Regression is 85.25%, Naïve Bayes is 85.25%, Decision Tree is 81.97%, Support Vector Machine is 81.97% and Random Forest is 95.08%.

V. CONCLUSION

The main objective of this research work is to find a better model to predict relatively accurate results regarding heart disease prediction to help in decreasing the death rate due to the non-availability of an appropriate automated system. This research work uses the Cleveland data set from UCI for training and testing the model. It performs a relative analysis among Logistic Regression, Naïve Bayes, Support vector machine, Decision tree, and Random forest algorithms to find the most accurate prediction. The Random forest algorithm has provided a comparatively precise result for the study with an accuracy of 95.08%.

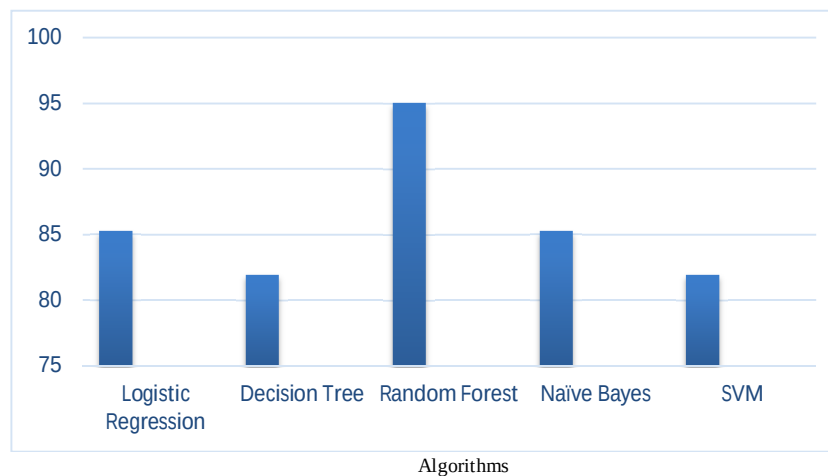


Figure 5: Bar Graph representing the percentage of accuracy

REFERENCES

- [1] Archana Singh, Rakesh Kumar "Heart Disease Prediction Using Machine Learning Algorithms", International Conference on Electrical and Electronics Engineering, 2020.
- [2] ApurbRajdhan, Milan Sai, Avi Agarwal, Dundigalla Ravi, Dr. Poonam Ghuli, "Heart Disease Prediction using Machine Learning", International Journal of Engineering Research & Technology, Vol.9 Issue 4th, April-2020.
- [3] Devansh Shah, Samir Patel, Santosh Kumar Bharti, "Heart Disease Prediction using Machine Learning Techniques", published on 16th, October-2020.
- [4] SarangamKodati, Dr. R Vivekanandam, "Analysis of Heart Disease using in Data Mining Tools Orange and Weka", Global Journal of Computer Science and Technology, Vol18. 2018.
- [5] C. Beulah Christalin Latha, S. Carolin Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked, 2019.
- [6] Rishabh Magar, Rohan Memane, Suraj Raut, Prof. V.S. Rupnar, "Heart Disease Prediction using Machine Learning", Journal of Emerging Technologies and Innovative Research, Vol7, June 2020.
- [7] K Subhadra, Vikas B, "Neural Network-Based Intelligent System for Predicting Heart Disease", International Journal of Innovative Technology and Exploring Engineering", Vol8. Issue 5th, March-2019.
- [8] Senthilkumar Mohan, Chandrashekar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, 3rd July-2019.
- [9] Chaitrali S. Dangare, Dr. Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, Vol47. 10th June 2012.
- [10] Adil Hussain She, Dr. Pawan Kumar Chaurasia, "A Review on Heart Disease Prediction Using Machine Learning Techniques", Research Gate, April-2019.