Heart Disease Prediction Using Machine Learning

Chikka Krishnappa T K[1], Roshani Gupta[2], Aishwarya M F Prabhakar[3], Shilpa N[4]
[1]Assistant Professor, Dept of Computer Science & Engineering
[2, 3, 4]Dept of Computer Science & Engineering
[1, 2, 3, 4] Atria Institute of Technology, Bangalore

Abstract- *Heart diseases are a major threat and concern in the lifestyle of people and in medical field. It has been detected as a disease that has one of the major mortality rates. Many areas of the word are affected by cardiovascular diseases. Taking these factors into consideration, the prediction of cardiovascular diseases becomes crucial and the methodologies that predict CVD's (Cardiovascular Disease) would help with reducing the mortality rate. Diagnosing the heart disease is a little complicated task to do. Therefore, there is a requirement for an automated system which is able to predict likelihood of heart disease of an person. Maintaining a prediction system ensures the quality of decisions made in medical field. A system that predicts the level of heart disease of a patient is developed using and collecting data on the attributes such as age, sex, chest pain type, blood pressure, fasting blood sugar etc. Python language is used as a coding language and jupyter notebook is used as the software platform. Four machine learning algorithms namely, decision tree, linear regression, naive bayes and random forest are used and the performance of all the algorithms are compared to get better results.*

Keywords- Machine Learning; Supervised Learning; Classification; Classifier; Hyperplane, Decision tree, Random forest, Naïve Bayes and Linear Regression.

I. INTRODUCTION

Thousands of people these days are affected with life threatening diseases. Cancer, stroke, diabetes etc. destroy ones physical well being and may result in death. Detection , diagnosis and treatment of these diseases during the first few stages can aid with the recovery or prevent further complications related to them. As a result, techniques like data mining and machine learning could be applied. Data mining identifies patterns in obtained data, allowing us to evaluate a variety of scenarios that may be used in the future to make decisions. Different ways for separating data or receiving arrangements based on learning and knowledge gain are referred to as data mining, with the purpose of using them for decision making and probability calculation prediction. The current health management systems capture a large amount of data on patient reports. Data mining techniques can be used to analyse massive amounts of data, and machine learning algorithms are utilised in the prediction process. As a result, the system's primary goal is to use data analytics to forecast the existence or absence of CVD and the severity of disease in patients.

A. Purpose

Hidden examples and connections are separated from other varies information sources using information and data mining. Data mining combines investigation, artificial intelligence, and dbms technology. Data mining has been linked to a few areas of therapeutic administrations, such as the revealing of linkages between analysis data and stored clinical data. Excellent therapeutic conclusion is a multi-step technique that necessitates accurate patient data, decades of professional knowledge, and a thorough understanding of therapeutic writing. The medical business has amassed a large amount of data, which has sadly not been exploited to uncover hidden links and patterns that would allow clinicians to make more efficient judgments.

Doctors regularly trust in their understanding and practice to make their selections. On the other hand, doctors having immense education in all divisions are extremely unavailable resource. The surgeons won't be able to analyse disorders precisely. Because of the challenging interdependence of multiple variables, effectively determining the symptoms of a disease at early stages is a strenuous task.

B. Dataset description

The data set comes from the University of California, Irvine (UCI) Data Mining Repository (Newman et al., 1998). The Cleveland data set is used to approve the system. In this datasets, there are Age, gender, type of chest pain, resting blood pressure, serum cholesterol in milligrammes per deciliter, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels, thal, and heart disease diagnosis are among the 14 attributes presented. [2].

1) Cleveland dataset: Robert Detrano, M.D., Ph.D., These data were gathered at the V.A. Medical Centre. All of the published

research focused on a subset of 14 of the 76 traits discovered after analysing the Cleveland heart disease database. Until now, ML researchers have relied solely on the Cleveland dataset. The seriousness of the sufferer's cardiac condition is specified in the num field by a number that ranges from 0 (no present) to 4. Experiments in the Cleveland database have focused on differentiating disorder existent (values14) from non-existent (rate 0). (Ephzibah, 2010). Six of the instances were eliminated because they lacked values. Heart disease is missing in 54 percent of the cases and present in 46 percent of the cases.

Only 14 attributes used:

- 3 (age) age in years
- 4 (sex) (1 = male; 0 = female)
- 9 (cp) chest pain type – Value 1: typical angina – Value 2: atypical angina – Value 3: non-anginal pain – Value 4: asymptomatic
- 10 (trestbps) resting blood pressure (in mm Hg on admission to the hospital)
- 12 (chol) serum cholestoral in mg/dl
- 16 (fbs) (fasting blood sugar >120 mg/dl) (1 = true; 0 = false)
- 19 (restecg) resting electrocardiographic results – Value 0: normal – Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) – Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- 32 (thalach) maximum heart rate achieved
- 38 (exang) exercise induced angina (1 = yes; 0 = no)
- 40 (oldpeak) ST depression induced by exercise relative to rest
- 41 (slope) the slope of the peak exercise ST segment – Value 1: upsloping – Value 2: flat – Value 3: downsloping
- 44 (ca) number of major vessels (0-3) colored by flourosopy
- 51 (thal) = normal; 6 = fixed defect; 7 = reversable defect
- 58 (num) (the predicted attribute) diagnosis of heart disease (angiographic disease status) – Value ranges 0-4, 4 being highly severe disease status. (in any major vessel: attributes 59 through 68 are vessels)

C. System Overview

In a sanatorium, decisions are frequently made largely on doctors' intuition and experience, rather than the database's wealth of knowledge records. This method leads to unfavorable biases, errors, and unnecessary health-care costs, all of which have an influence on the quality of services provided to patients. Experts and professionals will assist and make the diagnostic system more reliable by using machine learning to determine the automated end of diagnostic guidelines from previous descriptions, efficiently treat an affected person, and experts and professionals will assist and make the diagnostic system more reliable. Intelligent decision-making systems are defined as interactive laptop structures that assist in making decisions in the use of informational collections and models to find issues, resolve issues, and make intelligent decision-making systems are defined as interactive laptop structures that assist in making decisions in the use of records units and fashions to find issues, resolve issues, and make intelligent decision-making systems are defined as interactive laptop structures to assist in the utilization of records units and fashions to locate issues, resolve challenges, and create decisions. The proposed machine uses analysis to integrate and make the best decision on the health facility using a computer device. This patient file can lower the number of patients in a group, improving the safety of scientific decisions, reducing unfavourable changes in practice, and improving patient outcomes. This idea is appealing because modelling and assessment tools, such as information mining, have the power to create an information-rich environment that might significantly improve the quality of medical decisions.

Machine learning algorithms are used by the system to examine the data and train the models, which are then evaluated. Decision tree, naive bayes, random forest, and linear regression are some of the algorithms used for prediction. All four algorithms are used to create models, and their accuracy is compared. These models can be used to figure out what type of cardiac disease a patient has. The prediction is made by the system in the form of an expanded two-level categorization. Type 0 indicates the absence of cardiac disease, while type 1 through type 4 reflect the severity of the condition.

To train the model, algorithms use the Cleveland dataset. There are 12 attributes and 801 instances in the dataset. Age, sex, chest discomfort, blood pressure, serum cholesterol, fasting blood sugar, electrocardiograph, max heart rate, ST depression, slope, vessels, and diagnosis are all factors to consider. Table 1.1 shows some of the examples from the dataset that were used.

Table : Cleveland dataset

63	1	1	145	233	1	2	150	0	2.3	3	0	0
67	1	4	160	286	0	2	108	1	1.5	2	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	1
37	1	3	130	250	0	0	187	0	3.5	3	0	0
41	0	2	130	204	0	2	172	0	1.4	1	0	0
56	1	2	120	236	0	0	178	0	0.8	1	0	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3
57	0	4	120	354	0	0	163	1	0.6	1	0	0
63	1	4	130	254	0	2	147	0	1.4	2	1	2
53	1	4	140	203	1	2	155	1	3.1	3	0	1

II. PROPOSED SYSTEM

Goal of proposed algorithm is to ensure the accuracy level of prediction by using machine learning is at max level i.e.. Decision Tree Algorithm and Naive Bayes Classifier.

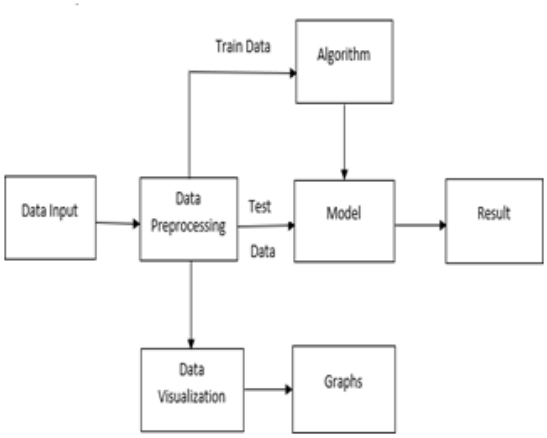
The dataset is introduced to the decision tree algorithm and Naive Bayes Classifier. In proposed system, dataset is trained and tested in the ratio of 70:30 with the help of decision tree. Naive Bayes Classifier is used here to classify each data based on the analysis of decision tree algorithm.

Naive Bayes is probabilistic model that uses Bayes Theorem for classification. The key insight of Bayes' theorems that the probability of an event can be adjusted as new data is introduced. An advantage of the naive Bayes classifier s that it requires a small amount of training data to estimate the parameters (means and variances of the Variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

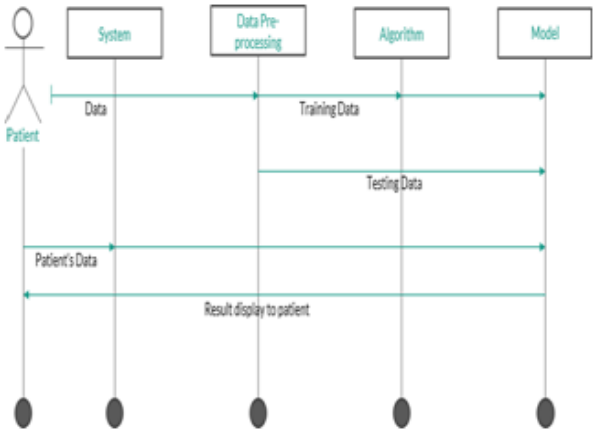
III. SYSTEM DESIGN

A. BLOCK DESIGN

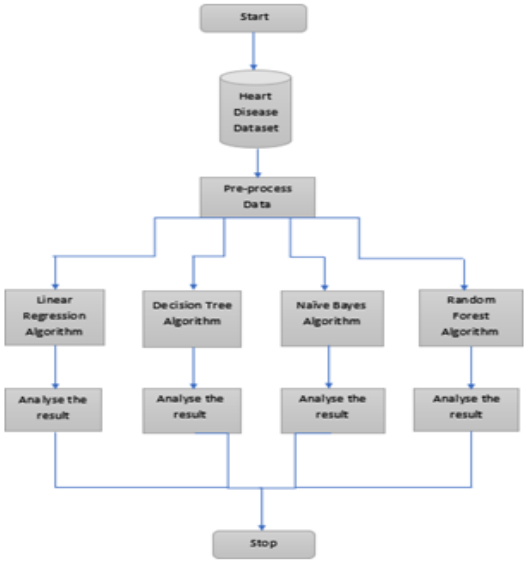


A block diagram is an outline of a framework in which the main parts are represented using blocks that are connected by lines that demonstrate the connections of each of the blocks.

B. SEQUENCE DIAGRAM



C. DATA FLOW DIAGRAM



IV. RESULTS

This process can be classified using a variety of Machine Learning methods. The methods used were linear regression classifier, decision tree method, naive bayes classifier, and random forest method, and the accuracy graph shown below is a bar graph.

According to Figure 6.1, accuracy of random forest algorithm is highest followed by decision tree, naive bayes and linear regression. The accuracy scores of the algorithms are:

- Random forest: 92-95%
- Decision tree: 90-92%
- Naive Bayes: 59-61%
- Linear regression: 58-61%

Patient's details whose probability of suffering from heart disease has to be predicted should be entered. The model will use this data to make the prediction represents the prediction process followed by linear regression algorithm.

Prediction by Linear regression

```
Enter value for  age
58
Enter value for  sex
1
Enter value for  chest_pain
4
Enter value for  blood pressure
150
Enter value for  serum_cholestoral
270
Enter value for  fasting_blood_sugar
0
Enter value for  electrocardiographic
2
Enter value for  max_heart_rate
111
Enter value for  induced_angina
1
Enter value for  ST_depression
0.8
Enter value for  slope
1
Enter value for  vessels
0

 ['58', '1', '4', '150', '270', '0', '2', '111', '1', '0.8', '1', '0']

Patient has heart disease type 2
```

Prediction by Random forest

```
Enter value for  age
58
Enter value for  sex
1
Enter value for  chest_pain
4
Enter value for  blood pressure
150
Enter value for  serum_cholestoral
270
Enter value for  fasting_blood_sugar
0
Enter value for  electrocardiographic
2
Enter value for  max_heart_rate
111
Enter value for  induced_angina
1
Enter value for  ST_depression
0.8
Enter value for  slope
1
Enter value for  vessels
0

 ['58', '1', '4', '150', '270', '0', '2', '111', '1', '0.8', '1', '0']

Patient has heart disease type 3
```

V. CONCLUSION

Cardiovascular disease (CVD) is a condition that affect because of the greater fatality rate, CVDs have become a major issue. Cardiovascular illnesses affect large swaths of the universe. Prediction of cardio vascular diseases becomes incredibly important among all of these basic elements, and models that forecast CVDs would have a considerable impact on lowering the fatality rate. In compared to merely two-level classification, expanded two-level classification has a greater priority since patients obtain a better picture of their health in terms of CVD situations. Avoidance is always an essential step in the prevention of any condition, thus it should be addressed as a top priority and worked on. Machine learning algorithms generate applicable software programs for prediction and disorder type prediction, respectively. When compared to the other three algorithms of decision tree, nave bayes, and linear regression, Random Forest predicts heart disease with a maximum accuracy of 95%.