JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Heart Disease Prediction Using Machine Learning Technique and Risk Analysis

Ms.Wable Sarika Sanjay, Prof. Dr. N.R. Wankhede

Abstract- **-** cardiovascular diseases (CVDs), which are diseases of the heart, are the main cause of the large number of fatalities that have occurred over the course of the most recent few years and have become the most dangerous disease in India and throughout the entire world. In this approach, there might be a need for a precise, workable, and reliable tool to study these illnesses in time for effective therapy. Numerous clinical datasets were used in conjunction with machine learning methods and techniques to conduct extensive and complex information research. In recent years, several analysts have used a variety of methodologies to provide the health care industry and internal specialists with the prediction of heart-related disorders. This study presents a survey of many models that are entirely based on such algorithms and techniques and examines their effectiveness. Models using supervised learning techniques, such as Support Vector Machines (SVM), Naive Bayes, Decision Trees (DT), Random Forest (RF), and ensemble models are incredibly distinctive among the many researchers.

Index Terms- **Cardiovascular, datasets, supervised learning algorithms, Support Vector Machines, K-Nearest Neighbors, Naive Bayes, Decision Trees, Random Forest**

I. INTRODUCTION

Data Mining is a non-minor extraction of certain, beforehand obscure and potentially valuable information about data [1]. In short, it is a procedure of analyzing information from the substitute perspective of view and assembling the knowledge of it [2]. The discovered information can be used for various applications, for example healthcare industry. The Healthcare industry is "data rich", however lamentably not every one of the information is dug which is required for finding hidden patterns and effective decision making. Data Mining Techniques such as Propelled data mining techniques are Utilized to find learning in the database and for medicinal research, especially in the heart disease prediction. A major challenge facing the healthcare industry is the nature of the administration. Poor analysis can prompt appalling outcomes which are unacceptable. The datasets are overwhelming for human personalities to fathom, can be effectively investigated utilizing different machine learning techniques. Accordingly, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately. The doctors are embracing many scientific technologies. Our project's objective is to foresee the odds of heart disease based on the patient's dataset and the doctor's perspective in check-up using machine learning. By utilizing the patient's medical records, a new system is proposed to foresee the chances of heart attack Attributes such as Blood pressure (bp), age, thickness of the artery, etc. are sustained into the dataset and algorithm [3].

II. Literature Review

The motivation to do this problem comes from the World Health Organization estimated. As per the World Health Organization estimation till 2030, practically 23.6 million people will pass on due to Heart illness. Along these lines, to limit the threat, desire for coronary disease should be done. Investigation of coronary sickness has been regularly in perspective on signs, appearances and physical examination of a patient. The most troublesome and complex assignment in medicinal services area is finding of right ailment or right illness. In late patterns investigation on these broad datasets has been able to be fundamental because of monetary weights on medicinal services commercial enterprises. (business endeavors).Chen et al. Proposed the prediction system for heart disease. coronary illness using learning vector quantization algorithm. Another study probed on sample database of patients' records. The Neural Network is prepared, tested, trained with 13 input factors such as Age, Blood Pressure, Kumaravel et al. Have proposed automatic diagnosis framework for heart diseases using a neural network system with an accuracy of 63.6–82.9%. The heart is an important organ of human body part and it is similar to a pump, which circulates blood through the body [4] and If the circulation of blood in the body is inefficient, then vital organs like brain suffer and if heart quits working, demise happens within minutes. Life is totally reliant on the

successful working of the heart. The term heart disease alludes to disease of heart and blood vessel framework inside 51it. Several factors have been demonstrated that increases the chances of heart disease: Family history, Smoking, Poor diet, High blood pressure (Hyper Tension), High blood cholesterol (Caused by Obesity), Physical inactivity. True assurance of coronary sickness can't be conceivable by utilizing just human comprehension. There are stores of parameters that can impact the precise end like less exact outcomes, less experience, time subordinate execution, data up degree and substantially more.

Table :- Analysis of Existing System

Paper Title	Algorithm	Advantages	Disadvantages
Efficient Heart Disease Prediction System	Decision Rules	It uses the data mining system to predict the heart disease	Accuracy is upto 86.3 % percent.
Prediction and Diagnosis of Heart Disease by Data Mining Techniques	Naïve Bayes, J48	Prediction and Diagnosis of Heart Disease by Data Mining Techniques.	J48 gives better accuracy than other technique
Heart Disease Diagnoses using Artificial Neural Network.	ANN	Artificial Neural Network is Used for Predicting Heart Attack. By using ANN we get accuracy of 88%.	Accuracy is not enough
Heart Disease Prediction Using Weka Tool	Weka Tool	Prediction of Heart Disease using WEKA tool.	The performance analysis and its parametric metrics gives the not optimized detection in CVD.

III. Proposed System

Machine learning approach

Machine learning methods unit of measurement trained on datasets and a model is created for analysis.
Based on the accuracy of the model, the machine learning technique is suitable. The three methods in machine learning algorithms unit of measurement are supervised learning, unattended learning, and reinforcement learning. In supervised learning, the model is trained victimization tagged data that contains every input and result. Unsupervised learning methods do not use employment data or tagged data. It finds the hidden structures or patterns from unlabeled data.

Supervised Learning

Supervised learning desires a well-labeled dataset to educate. Supervised learning is of two types' regression and classification. In this system classification techniques facilitate hunting out the acceptable class labels which can predict the heart disease present. A machine learning model is developed that uses the labeled information to educate, classify the images and predict the disease status.

Unsupervised Learning

Unsupervised ways are supported by machine learning. The necessity of labeled datasets isn't needed in unsupervised learning. Image analysis was once done using unsupervised learning.

Random Forest

The classification contains information about trees from different sources. It takes a lot of time and effort to lay out the data. As a general rule, the more trees in a tree region, the better it is, but the more difficult it is to maintain. It would be preferable if classes could predict specific outcomes rather than anticipated outcomes. Each tree should have negligible assumptions. In contrast to other models, this requires less effort to arrange. According to it, the enormous structures that have been established work admirably in any capacity. It may be valid even without much data. In addition to scikit-learn, panda, matplotlib and other key libraries, the work is done in Python 3.6.4. The key control key has been used to eliminate the data from bitinformatics.com Modern data is incorporated into erased data. Eighty percent of the pamphlet is viewed as part of the train, while twenty percent is viewed as the test standard. As expected, timberlands and retreats have been estimated with smart estimations. Investigation of quality. Assaults are performed on the pre-handling line, and the cost is determined by grades given against the identification.

Figure.1 System Architecture

The proposed system shown in Fig1 predicts heart disease by exploring the 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.

Logistic Regression

Logistic regression is one of the Machine Learning algorithms that is most frequently employed in the Supervised Learning category. It is used to forecast the categorical dependent variable using a specified set of independent variables. Logistic regression is used to predict the output for a dependant variable that is categorical. The outcome must therefore be a discrete or categorical value. It offers the probabilistic values that lie between 0 and 1 rather than the precise values between 0 and 1. It can be either True or False, 0 or 1, or Yes or No. Logistic regression and linear regression are fairly similar, with the exception of how they are used. In contrast to linear regression, which is used to address regression issues, logistic regression addresses classification issues. Instead of fitting a regression line in logistic regression, we fit a "S" shaped logistic function that predicts two maximum values (0 or 1). The logistic function's curve demonstrates numerous possibilities, such as whether or not the cells are cancerous, whether or not a mouse is obese dependent on its weight, etc. Using both continuous and discrete datasets to classify fresh data, logistic regression is a crucial machine learning technique. Logistic regression can be used to quickly pinpoint the elements that will be effective when classifying observations using multiple sources of data.

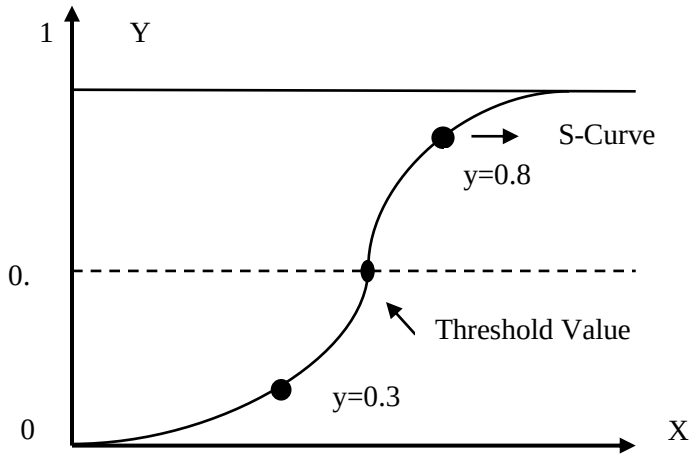

Figure2 Logistic Function

Factors Affecting for the disease.

Age:-Age is mostly used factor for developing the risk of heart disease but now a days many youngster also face the same due to less exercise, unhealthy lifestyle etc.

Older people are at greater risk of developing heart disease. Although the process of aging cannot be changed, leading a generally healthy lifestyle is recommended to help reduce the likelihood of developing heart and circulatory conditions.

Kindred: - Statistics suggest that people of South Indian, African or Caribbean descent have a greater risk of developing cardiovascular disease. Type 2 diabetes – a risk factor in itself for cardiovascular disease – also seems to be more prevalent among these groups. The reasons for this are difficult to define. However, leading a healthy lifestyle is generally recommended as a way for people from all backgrounds to help prevent heart and circulatory disease from developing.

Sex: - While it may have long been seen as a man's disease, the risk of cardiovascular disease in women has been underestimated, and symptoms may go unrecognized, complicating diagnosis and treatment. Though CVD risk factors are shared by men and women, some may be more prevalent and/or more significant for one sex or gender; for example, having diabetes may be a stronger risk for certain types of CVD in women.

Cholesterol: -High levels of low-density lipoprotein (LDL) cholesterol – also known as "bad cholesterol" – are linked to a range of cardiovascular diseases. Cholesterol is a fatty substance that is carried around the body by proteins. If too much LDL cholesterol is present, it can cause fatty substances to build up in the artery Walls and lead to complications.

Many more different attributes that can be added along with other attributes.

ANN (Artificial Neural Network)

In the past ten years, research interest in artificial neural networks, a type of artificial computer intelligence, has increased. Although they have been extensively employed to solve technical problems, medical difficulties, notably in the disciplines of radiology, urology, laboratory medicine, and cardiology, have only recently been subjected to their application. A distributed network of computing components called an artificial neural network is modelled after a biologic brain system and can be used in computer applications. It has the ability to find relationships in input data that are difficult to see using the standard analytical approaches now in use.

IV. Mathematical Formulation

In an early mistake, the assessed value (test or populace values) of a given an example or worth is often contrasted with the real value. RMSE stands for root mean square error

The equation number 1 shows the rmse formula.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} \quad (1)$$

Dataset Used

This reads the data from csv and performs preprocessing on the dataset. If any missing values in the dataset system will fill the missing values using preprocessing the dataset.

The dataset used was the
Heart disease
Dataset{https://www.kaggle.com/datasets/priyanka841/heart-disease-prediction-uci} It is a combination of 4 different databases, 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. Therefore, this system uses the already processed UCI Cleveland dataset available in the Kaggle website for our analysis. The fillna() function iterates through your dataset and fills all null rows with a specified value.

Software requirement specification

- Python

Hardware requirement specification

- Laptop

V. RESULTS

1. Main Form GUI

Figure3. Main Form GUI

Figure3 shows the main form gui. In gui there are options for Load CSV, Finding Missing Values and feature extraction and algorithms

2. Gender Wise Distribution

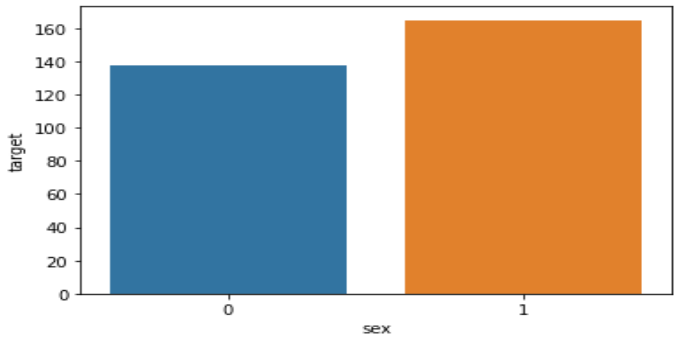

Figure4. Gender Wise Distribution

In the figure4 displays the gender wise distribution bar chart.

3. Attribute Wise Distribution

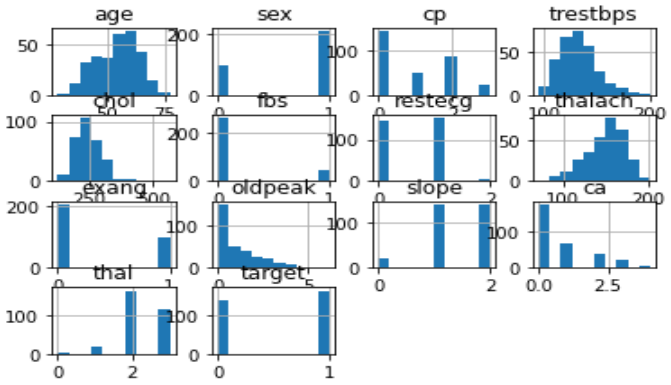

Figure5. Gender Wise Distribution

In the figure5 displays the gender wise distribution bar chart.

4. Attribute Wise Distribution

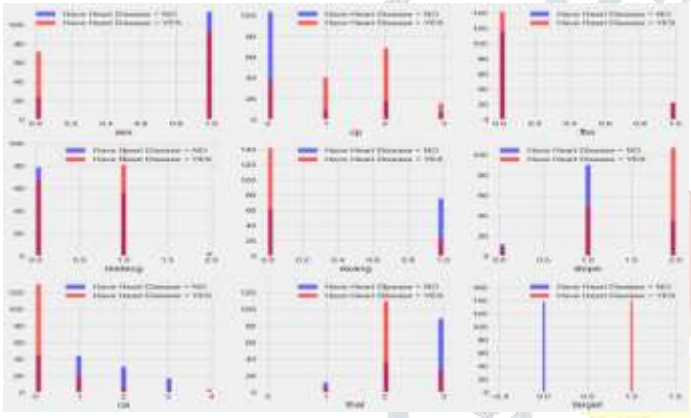

Figure6 Attribute wise distribution

In Figure6.displays the attribute wise distribution.

5. Feature Extraction

Figure7. Feature Extraction

In Figure7 shows the feature extraction from the CSV file.

6. Logistic Regression Confusion Matrix

```
Train Result:
================================================
Accuracy Score: 86.79%
------------------------------------------------
CLASSIFICATION REPORT:
                    0      1   accuracy  macro avg  weighted avg
precision        0.88   0.86       0.87       0.87          0.87
recall           0.82   0.90       0.87       0.86          0.87
f1-score         0.85   0.88       0.87       0.87          0.87
support         97.00 115.00       0.87     212.00        212.00
------------------------------------------------
Confusion Matrix:
 [[ 80  17]
 [ 11 104]]

Test Result:
================================================
Accuracy Score: 86.81%
------------------------------------------------
CLASSIFICATION REPORT:
                    0      1   accuracy  macro avg  weighted avg
precision        0.87   0.87       0.87       0.87          0.87
recall           0.83   0.90       0.87       0.86          0.87
f1-score         0.85   0.88       0.87       0.87          0.87
support         41.00 50.00       0.87      91.00         91.00
------------------------------------------------
```

Figure8. Confusion Matrix

The figure8 shows the confusion matrix.

6. Accuracy Metrics

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	87.26	81.32
1	K-nearest neighbors	76.89	65.93
2	Support Vector Machine	100.00	54.95
3	Random Forest Classifier	100.00	81.32

Figure9 Accuracy Metrics

The Figure9 Shows the Accuracy Metrics

7. ANN Classifier Result

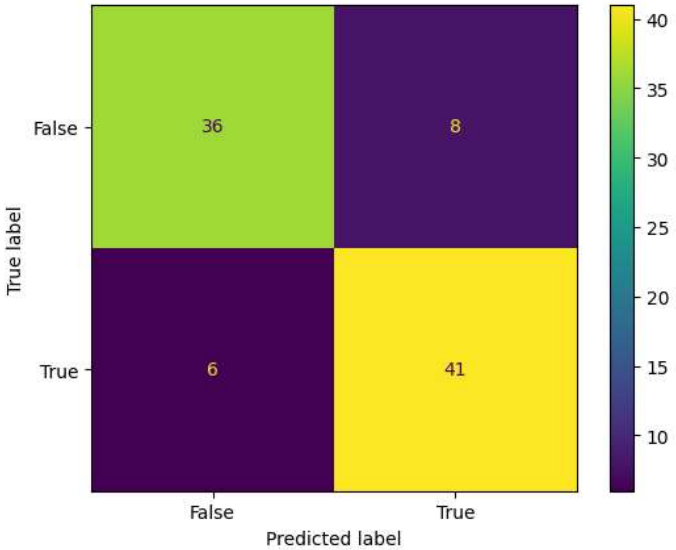

Figure10 Confusion Matrix Result

CONCLUSION

Heart illnesses are fatal and difficult to treat, taking many lives each year. A sedentary lifestyle and excessive stress are two main factors that have made things worse. Early disease detection makes it easier to treat the condition. Once a patient has been diagnosed with a sort of heart disease. When choosing a course of therapy from such suitable machine learning approaches, datasets, and algorithms, data mining can be of great assistance. While using the same machine learning method, datasets addressing the same medical condition like coronary artery disease (CAD) may show different results. The effectiveness of the medical diagnosis and analysis is the primary factor on which the classification results and accuracy score result for the chosen important aspects are based. CNN is utilized to predict the patient's diagnosis with more accuracy than was possible before qualities were reduced. A smaller dataset will perform better than a larger one. Before building the model, inconsistent and missing values were fixed, however in real life, this is not the case.

REFERENCES

[1] K. Sudhakar, Dr. M. Manimekalai Study of Heart disease prediction using data mining ISSN: 2277 128X

[2] M. Revathi Review of Heart Disease Prediction using Data mining techniques

[3] Sellappan Palaniappan, Raah Awang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008

[4] Ankita Pimputkar, Jitendra S. Dhobi A Survey on Heart Disease Prediction using Hybrid Technique in Data Mining IJARIIE-ISSN(O)-23954396

[5]Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee Heart Diagnosis and prediction using machine learning and data mining : Review

[6] Burak Kolukisa , Hilal Hacilar, Gokhan Goy, Mustatfa Kus, Burcu BakirGungor, Atilla Aral, Vehbi Cagrigungor Evaluation of classication algorithms, Linear discriminate Analysis, and a New hybrid Feature selection methodology for the diagnosis of Coronary Artery Disease

[7] S. B. Patil and Y. S. Kumaraswamy, Intelligent and effective heart attack prediction system using data mining and articial neural network, European Journal of Scientic Research, 2009, p. 642-656.