Heart disease prediction using machine learning algorithms

Hrudi Sai Akhil Bommadevara, Y.Sowmya, G.pradeepini

***Abstract:** Machine learning is the sub branch of artificial intelligence and it is making computers to learn from data without being explicitly programmed Heart disease prediction is used to determine the root cause of getting heart attack and the probability of getting a heart attack, group the people into different clusters based on getting heart attack or not There are five levels in heart attack from level 0 to level 4. There are 14 important attributes to be considered in analysis of heart attack namely age, BP, CHOL, gender, CP, CA, THAL.*

***Keywords:** Naïve Bayes, Decision Tree, Clustering, Linear Regression, Correlation.*

I. LITERATURE REVIEW

Different researchers have worked on heart disease prediction and used various machine learning algorithms and got different accuracy levels S.Indhumathi.etl used naïve Bayes for heart disease prediction at first they have pre-processed the data by using data cleaning methods next by using naïve Bayes function they have tested the data for prediction Detrano worked on logistic regression on Cleveland dataset and received an accuracy of 77% John Gennari worked on incremental concept formation and achieved 78.9% accuracy with clustering system.

II. INTRODUCTION

Machine learning is taking the input from the past or historical data and trying to predict the future. The output is discrete (true or false, yes or no) There are several machine learning algorithms like Naïve Bayes, Support Vector Machine, Decision Tree, Clustering etc The performance measure on a particular task improves with experience Heart disease diagnosis: We have total 74 attributes for diagnosis in which only 14 important attributes are considered, we have to predict the status of the patient in which there are 5 values from 0-4 The machine learning algorithms we considered are Naïve Bayes, Decision Tree, K-Means clustering and linear regression.

The attributes we considered are Age, Gender (0-female, 1-male), CP(chest pain)(1-typical angina, 2-atypical angina),
TRESTBPS(Bloodpressure),Chol(cholesterol)(mg/dl),
THALACH(max heart rate),ca, THAL

A. Decision tree

Decision tree is a classification algorithm, It is used to determine the root cause of getting heart attack, the root node of decision tree is the root cause of heart attack There are three important parts in decision tree

1.root node (entire population is considered as root node)

2. decision node (If there is further splitting of data)

3. leaf node (If there is no more splitting of data)

The two important things in decision tree:

1. Entropy:

$$Entropy(T)=\sum_{i=1}^{c} p_i \log_2 p_i$$

2. Information gain:

GAIN(T, X)=Entropy(T)-entropy(T,X)

Highest Information gain will be chosen as a root node

Decision tree can handle both discrete and continuous values and can easily generate IF-THEN rules based on the tree constructed The major disadvantage of decision tree is over fitting this can be resolved by pruning Pruning is the process of removing unwanted nodes in a tree, there are 2 types of pruning

1. Pre-pruning: remove the nodes during construction of the tree

2. Post-pruning: construct the tree completely and then identify the unwanted nodes and remove the nodes

Post pruning is preferred than pre-pruning because it is difficult to cut the trees during the construction of a tree

B. Naïve Bayes

In a probabilistic approach we have two kinds of probability:

1. Prior probability: Is the probability of an event before new information is collected, this probability doesn’t have any condition. Ex: P(Buys computer)

2. Posterior probability: calculating the probability under a certain condition

Ex: P(Buys computer/Age=35)

There are generally two algorithms in probabilistic approach

Manuscript published on 30 March 2019.
***Correspondence Author(s)**

Bommadevara hrudi sai akhil, Department of Computer Science and Engineering KL University, Vaddeswaram, AP, India.
Sowmya yalavarthi, Department of Computer Science and Engineering KL University, Vaddeswaram, AP, India.
G. Pradeepini, Department of Computer Science and Engineering KL University, Vaddeswaram, AP, India.

1. Bayes theorem
2. Naïve Bayes

Bayes theorem: Is the probability of an event based on prior knowledge, the major disadvantage of Bayes theorem is that we cannot classify new instances

P(A/B) = (P(B/A).P(A))/P(B)

Naïve Bayes: Is a conditional probability model, the instance which we have to classify is represented by a vector x= (x1, x2, x3.)

$P(C_k/X)=(p(C_k)p(X/C_k))/P(X)$

C. *Linear Regression:*

Is generating a relationship between one dependent variable and one or more independent variables If there is only one independent variable, then it is called as simple linear regression, it is of the form Y=mX+C

If there is more than one independent variable, then it is called as multivariate linear regression With the help of linear regression, we can predict the values, Ex: If we give weight as input then we can predict the height of a person here weight is independent variable and height is dependent variable which depends on weight Linear regression will be useful when our goal is to predict or forecasting or error reduction In heart disease prediction it will be helpful in determining the BP/Chol of a person if you give age as input

D. *Clustering*

It is concept of grouping set of objects in such a way that all similar objects fall under one cluster The process is to choose the number of clusters and calculate the distance from the mean and the point and group them into clusters, calculate the new mean and keep repeating the process until previous mean and present mean are same. Here the method we use to calculate the distance is Euclidean distance Euclidean distance=sqrt($(x_2-x_1)^2+(y_2-y_1)^2$) There are many clustering algorithms available some of them are: K-Means Clustering, Mean-Shift Clustering, Density based Clustering, EM Clustering and hierarchal Clustering The major disadvantage of K-Means Clustering is if we choose less K value the accuracy will be less Ex: if we choose K=2 it will classify into two groups (getting heart attack, not getting heart attack) as there are 5 levels in prediction of heart disease we choose K=5 so it groups into 5 clusters based on level of heart attack .

III. PRESENT WORK

If there are any NULL values in the dataset or if any values are being missed they are replaced with -1 this stage is called as pre-processing stage and will be done after importing the dataset.

The machine learning algorithms we used are:

1. Decision Tree
2. Naïve Bayes
3. Linear Regression
4. K-Means Clustering

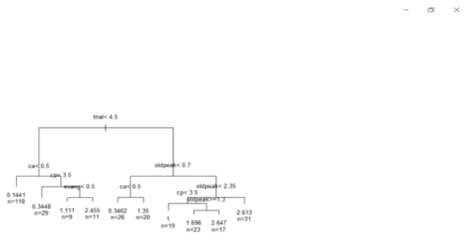
A. *Attributes*

Name	**Type**	**Description**
Age	Continuous	Age in years
Sex	Discrete	0-female 1-male
Cp	Discrete	Chest pain type 1.Typical angina 2.Atypical angina 3.Non-anginal pain.
TRESBPS	Continuous	Resting blood pressure (0-200)(mm/Hg)
Chol	Continuous	Cholesterol (0-603)(mg/dl)
FBS	Discrete	Blood sugar 0-false(<120 mg/dl) 1-true(>120 mg/dl)
RESTECG	Discrete	Resting electro cardio graphic results 0,1,2
THAL	Discrete	3-normal 6-fixed defect 7-reversable defect
THALACH	Continuous	Maximum heart rate
EXANG	Discrete	Exercise induced angina 0-no 1-yes
Num	Discrete	0-negative diagnosis 1-4 (from least serious to most serious)

This information data set is the part of a Heart Disease Data Set (it is obtained from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation), it contains a subset of 14 attributes. The task is to detect the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

IV. RESULTS

A. *Decision tree*



The root cause for heart disease prediction is THAL, 118 patients in our dataset have THAL<4.5 and ca<0.5 There are 29 patients who have cp<3.5 and ca>0.5 with THAL<4.5 in our dataset, 9 patients have EXANG value less than 0.5 with cp greater than 3.5 and calcium greater than 0.5 with

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

THAL less than 4.5 remaining 11 patients have EXANG greater than 0.5, there are 26 patients who have calcium content less than 0.5 and OLDPEAK less than 0.7 with THAL greater than 4.5 and the remaining 20 patients have calcium greater than 0.5 with OLDPEAK less than 0.7

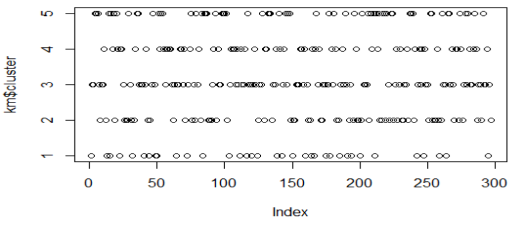
The root node in our decision tree is THAL and the decision nodes are calcium, chest pain and OLDPEAK the leaf nodes is EXANG remaining attributes are not considered because they are removed by pruning and from our results the major reason for heart attack is THAL<4.5 and ca<0.5

B. Naïve Bayes

0	1	2	3	4
0.533	0.179	0.117	0.117	0.051

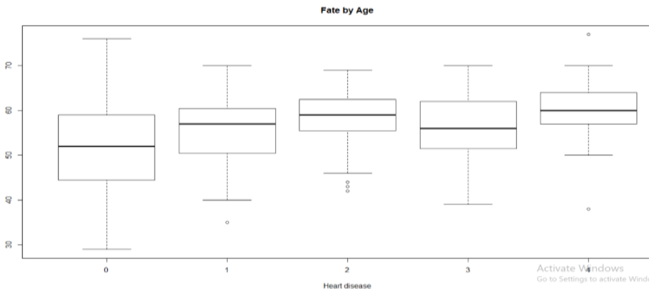
There are 5 levels in heart attack status 0th level indicates that there is no chance of getting heart attack which means 53% of total dataset doesn't get heart attack, 17% of the total dataset falls under level 1 category which has less chance of heart attack, 11% of the total dataset falls under level 2 category and only 5% has higher chances of getting heart attack, here in heart disease prediction we used naïve Bayes instead of brute force because brute force cannot classify new instances

C. Clustering



Here we used 5 levels of clustering and total 300 rows in our data set. And grouped all the similar levels into a cluster in our dataset row 2,4,6,8,10 have level 0 in the dataset so they are placed in cluster 0 and 3,5,13,18 have level 1 in the dataset so they are placed in cluster 1 In heart disease prediction we used K-Means clustering

D. Boxplot age vs heart disease



Age around 55 have heart status 0 which states that in our dataset people who have age around 55 doesn't get heart attack and age around 60 have heart status 4 which means that they have high chances of getting heart attack

E. Linear regression:

Age	BP
25	115.34
32	119.23
35	120.90
69	139.76
45	126.45

The person whose age is 25-45 has normal BP at a range of 115-126. Old people from age 55-70 has moderately high BP at a range of 139-150.So people at age 55-70 may suffer from heart diseases because they have high blood pressure

V. CONCLUSION

By using the machine learning algorithms on heart disease data set we have come to a conclusion that the root cause for heart disease is THAL and only 5% of the people have high chances of getting heart attack in our dataset. People having age around 50-60 have high BP which states that they have high chances of getting heart attack. As the age increases the cholesterol levels and BP keeps on increasing

REFERENCES

1. V. Kirubha and S. M. Priya, "Survey on Data Mining Algorithms in Disease Prediction," vol. 38, no. 3, pp. 124–128, 2016.
2. Mai Shouman, Tim Turner, and Rob Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients", International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.
3. K.Sudhakar, and Dr. M. Manimekalai, January 2014, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 1,pp. 1157-1160.
4. C. S. Dangare and S. S. Apte, "Improved study of heart disease predictionsystem using data mining classification techniques," InternationalJournal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.
5. Sairabi H. Mujawar, and P. R. Devale, October 2015,"Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.
6. Ashwini Shetty A, and Chandra Naik, May 2016,"Different Data Mining Approaches for Predicting Heart Disease", International Journal of Innovative Research in Science,Engineering and Technology(An ISO 3297: 2007 Certified Organization), Vol. 5, Special Issue 9, pp. 277-281.
7. Serdar AYDIN, Meysam Ahanpanjeh,and Sogol Mohabbatiyan,February 2016, "Comparison And Evaluation of Data Mining Techniques in the Diagnosis of Heart Disease",International Journal on Computational Science & Applications (IJCSA), Vol. 6,No.1, pp. 1-15.
8. Shadab Adam Pattekari,and Asma Parveen, 2012,"Prediction System for Heart Disease using Naive Bayes", International Journal of Advanced Computer and Mathematical Sciences, ISSN: 2230-9624,Vol. 3, Issue 3, pp. 290-294.