3 MACHINE LEARNING TECHNIQUES FOR HEART DISEASE PREDICTION

Sai Manasa Vedantam

Abstract

With the advancement in technology and with the advent of Machine Learning Techniques, the focus of most of the researchers has shifted towards applying these models to make the hustle-bustle life of modern man easier. As health care is a fundamental matter of interest for everybody, the thought of implementing a Heart disease prediction system has took its birth. A recent study showed that 1 out of every 3 deaths is due to heart diseases or cardiovascular diseases. In order to reduce the large scale of deaths from heart diseases, a quick and efficient detection technique is to be discovered. This paper introduces a survey of various models based on algorithms and techniques and analyze their performance to determine their suitability for using them to develop the model. The models based on supervised learning are Naive Bayes, Support Vector Machine and Logistic Regression are found very popular.

INTRODUCTION:

Heart is one of the main organs of the human body. It pumps blood trough the blood vessels of the circulatory system. The circulatory system is extremely vital because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death.

Heart diseases have emerged as a prominent case around the world. As per World Health Organization (WHO), Heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global. Fortunately, one can identify the chance of getting attacked by heart diseases by observing some initial symptoms which has the ability to show that something is wrong with one's heart health. Such symptoms are as follows:-

1. Chest pain

It is the most common symptom of heart attack. If someone has a blocked artery or is having a heart attack, he may feel pain, tightness or pressure in the chest.

2. Nausea, Indigestion, Heartburn and Stomach Pain

These are some of the often overlooked symptoms of heart attack. Women tend to show these symptoms more than men.

3. Pain in the Arms

The pain often starts in the chest and then moves towards the arms, especially in the left side.

4. Feeling Dizzy and Light Headed

Things that lead to the loss of balance. They make us lose control over ourselves with continuous drowsy feeling.

5. Fatigue

Simple chores which begin to set a feeling of tiredness should not be ignored as they turn out to be life threatening if not observed.

6. Sweating

Some other cardiovascular diseases which are quite common are stroke, heart failure, hypertensive heart disease, Aortic aneurysms, Cardiomyopathy, Peripheral artery disease, Congenital heart disease and Venous thrombosis show sweating as a major symptom. The person might suffer from continuous sweating.

RELATED WORK:

In recent years, with the development in technology, methods of treating various diseases also took a higher step. It is evident that even heart diseases could be treated well with the improvements in the field of medicine no matter how complex the situation is. But the major challenge turned fact is that treatment is possible only when the issue is identified. The heart disease prediction system is aimed at identifying or predicting the chance of occurrence of heart disease to an individual by analyzing some related data. Earlier, this system has been developed using Data mining and Hadoop. In such systems, Knowledge discovery is the centre for the success of the system. The issue with such systems is that the data pruning techniques which could reduce the inaccuracies occurred due to over fitting weren't defined. This is a significant drawback as inaccuracy couldn't be tolerated in such life critical systems and hence, they are found to be unreliable.

[1]. L. Sathish Kumar and A. Padmapriya have given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone which gives a programmed and concealed way to deal with recognize designs that are covered up of coronary illness. The given framework utilize information mining methods, for example, ID3 algorithm.

[2]. M.A. Nishara Banu and B. Gomathy have given a paper named Disease Predicting system using data mining techniques. In this paper they talk about Maximal Frequent Item set Algorithm and K-Means clustering. Classification is important for prediction of a disease.

[3]. Mohammed Abdul Khaleel has given paper in the Survey of Techniques for mining of data on Medical Data for finding frequent diseases locally. This paper focuses on dissect information mining procedures which are required for medicinal information mining to find locally visit illnesses. In this, the algorithm used was Naive Bayes which indeed uses Bayes theorem. Hence Naive Bayes has a very high power to make assumption independently. The used dataset has information about more than 500 patients and is obtained from a diabetic research institutes of Chennai, Tamilnadu which is leading institute. WEKA tool is used and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%.

[4]. D.R. PatiI and Jayshril S. Sonawane have given a paper named Prediction of Heart Disease Using Learning Vector Quantization Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation.

METHODOLOGY:

1. Pre-Processing

Cleaning: Data upon which we want to work with may not be clean and ready for use. It may contain noise or it may contain some of the values missing. If we process such data, we can't get good results. Hence, it is necessary to perform data pre-processing in order to obtain good and perfect results. In this step, we will fill missing values and remove noise by using some techniques like filling with most common value in missing place.

Transformation: This involves changing data format from one form to other to make it more easily understandable by performing standardization, normalization, smoothing, generalization and aggregation techniques on data.

Integration: Data that we need to process may not be from a single source. Sometimes, it can be from different sources and we may have to integrate them which may be a problem while processing. So, proper integration is one of important phases in data pre-processing.

Reduction: When we work on data, it may be complex and it may be difficult to understand. Sometimes, in order to make them understandable to system, we will have reduce them to required format so that we can achieve good results.

The dataset used contains 4240 records with the attributes *Gender, Age, Education, CurrentSmoker, CigsPerDay, BPMeds, PrevalentStroke, PrevalentHyp, Diabetes, TotChol, SysBP, DiaBP, BMI, HeartRate, Glucose and Chd.*

Gender - A binary attribute representing male with 1 and female with 0
Age - Varies from 35 to 66
Education - Lies within the range 1 and 4
CurrentSmoker - A binary attribute representing YES with 1 and NO with 0
CigsPerDay - Denotes the count of the cigarettes a person smoke per a day
BPMeds - If the person is consuming BP medicines, it is 1. Otherwise, it is 0
PrevalentStroke - If the person has already suffered from heart disease in the past, it is 1. Otherwise, it is 0
PrevalentHyp - If the person is a sufferer of Hypertension, it is 1. Otherwise, it is 0
Diabetes - If the person is diabetic, it is 1. Otherwise, it is 0
TotChol - Denotes the total amount of cholesterol in a person's body.
SysBP - Gives the Systolic BP
DiaBP - Gives the Diastolic BP
BMI - Denotes the BMI of a person
HeartRate - A number representing the pulse of peron's heart
Chd - A binary attribute (label) showing the occurrence of heart disease with1 and non occurrence with 1

(a) NAIVE BAYES ALGORITHM:

It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes theorem. All Naive Bayes classifiers assume that *the value of any particular feature is independent of the value of any other feature*, given the class variable. Bayes theorem is given as follows:

$$\mathbf{P(A|B) = P(A).P(B|A) / P(B)}$$

where: A and B are events, $P(B) > 0$.

$P(B|A)$ is the probability of B occurring given that A is true

$P(A)$ and $P(B)$ are individual probabilities of A and B

Here, $P(A|B') = P(B1|A) \times P(B2|A) \times P(B3|A) \ldots\ldots \times P(A)$

Though it assumes an unrealistic condition that attribute values are conditionally independent, it performs surprisingly well on large datasets where this condition is assumed and holds.

Use of Naive Bayes for this system has its own advantages and disadvantages.

1. Advantages: It is very simple and easy to implement. Moreover, it works well with numeric as well as categorical data.

2. Disadvantages: It may lead to the probability zero as it is frequency based.

(b) ID3 Algorithm

To do this, we have many machine learning algorithms out of which, the most widely used methods are Naïve Bayes classification technique and Decision tree construction. In the construction of decision tree, we have many algorithms of which we look for this ID3 algorithm. The ID3 algorithm is one of oldest algorithms which is used for building decision trees. In the process of building decision tree, it handles missing values and removes outliers i.e; those data points which lead to noise. So, we can build this decision tree even the data is not cleaned well. Decision tree constructs classification or regression models as a structure which is similar to tree. It separates a dataset into fewer and fewer sub-sets while in the meantime a related decision tree is incrementally created. The last outcome is a tree with choice point and leaf point. A choice node has a minimum of 2 branches. Leaf nodes speaks to a group or choice. The highest choice hub in a tree which compares to the best indicator is called the root point. Choice trees can deal with various kinds of information.

ID3 is an algorithm which is used to build decision trees. It has some features like removing outliers, handling missing values but the major disadvantage is to handle over-fitting and it's not so easy to implement as that of Naïve Bayes algorithm.

Step 1: If all occasions in X are certain, then make YES node and end. On the off chance that all cases in X are negative, make a NO node and end. Generally select an element, B with qualities U1, ..., Un and make a choice node.

Step 2: Partition the preparation occasions in X into subsets X1, X2....Xn as indicated by the estimations of U.

Step 3: Apply the calculation recursively to each of the sets Ai.

However, this approach has its own advantages and disadvantages.

1. Advantages: Decision trees are capable to learn disjunctive expressions and their robustness to noisy data seem convenient for document classification.

2. Disadvantages: Learning of decision tree algorithms cannot guarantee to return the globally optimal decision tree.

Improvement:

Many improvements can be done to Decision Tree algorithm itself, the learner and features also. These improvements can be modification / addition to the algorithm itself or extraction-selection/reduction of the features. Instead, if we use C4.5 algorithm, then, we can obtain the following advantages:

- Handling both continuous and discrete attributes:

- In order to handle continuous attributes, C4.5 creates a threshold and then splits the list in such a way that the attribute whose value is above the threshold and those that are less than or equal to it could be easily identified.

- Handling training data with missing attribute values:

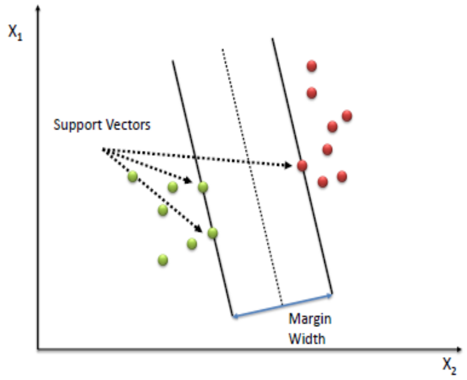
- C4.5 allows attribute values to be marked as _?' for missing. Missing attribute values are simply not used in gain and entropy calculations.

- C4.5 allows the attributes with different costs.
- Post Pruning - C4.5 creates first decision tree and after creation, it goes back through the tree and attempts to remove branches that do not help by replacing them with leaf nodes.

(c) Support Vector Machine (SVM):

It is a supervised machine learning algorithm which learns a linear model. The principle of SVM is finding a set of support vectors and building a classification or regression function upon them. The crucial aspect of SVM lies in the identification of a Hyper plane with maximal margin. Support vectors are those data points which the margin pushes against. A hyper plane which is as far as possible from the closest samples on either side classifies the best. Support Vector Regression (SVR) and Support Vector Classification (SVC) are two types of SVM for regression and classification respectively. To design the heart disease prediction system, we have used SVM linear classifier for a Binary Classification.

Using the Kernel trick, we can project data from a higher dimensional space to a lower dimensional space. The equation which defines a hyper plane is: $\boldsymbol{W^T X = 0}$ where W and X are two vectors. However, identifying the Optimal hyper plane is the objective of Support Vector Machine (SVM).



Support vector machine attempts to maximize the margin (distance between the hyper plane and the two closest data points from each respective class) to decrease any chance of misclassification. Some simple and popular implementations of support vector machine could be found in scikit-learn of Python and MATLAB. This algorithm deals mostly with the way the data is split and the hyper plane identification.

SVM has its own set of advantages and disadvantages.

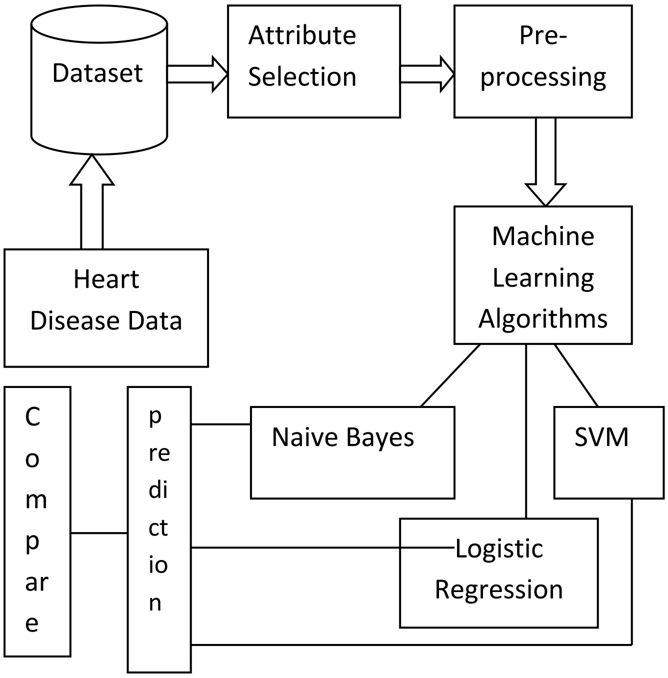
1. Advantages: Guaranteed optimality and abundance of implementations for both soft and hard margin data.

2. Disadvantages: SVMs cannot return a probabilistic confidence and also, the inferences becomes tedious with increase in the data size.

PROPOSED SYSTEM:

Logistic Regression:

Regression is a technique which is used to predict a dependent (target) variable from one or more independent variables. Logistic Regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used when the dependent or target variable is categorical. Hence, the outcome is usually a dichotomous variable i.e; there are only two possible outcomes. Binary Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables which may be continuous or discrete.



Logistic Regression uses Logistic Function, otherwise known as Sigmoid Function which was initially developed by statisticians to describe the properties of population growth in an ecology. It is a common **S** - shaped curve which can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits. When the target variable is binary or dichotomous, it is better to use Logistic Regression. The Logistic Function is given by the formula:

$$\boldsymbol{f(x) = \frac{L}{1 + e\hat{}(-k(x - x0))}}$$

where

L = Curve's max value

K = Curve's steepness

x0 = 'x' value of Sigmoid's mid point *f(x)* is said to be the STANDARD LOGISTIC FUNCTION if k = 1, x0 = 0 and L = 1. It is given by:

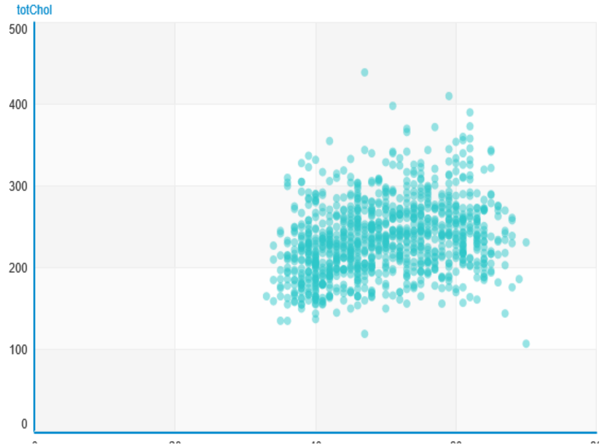
$$\boldsymbol{S}(\boldsymbol{x}) = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}\hat{} - \boldsymbol{x}}$$

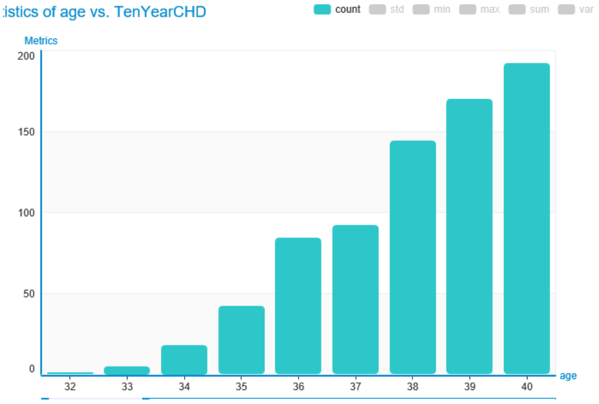
If *S(x)* = p, then the probability of getting attacked by Heart Disease is 'p'

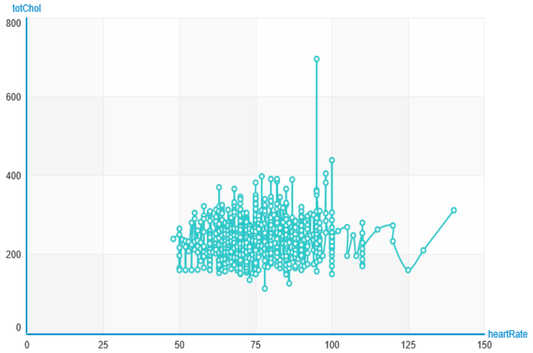
Eg: *S(x)* = 0.65 means there is 65% chance or there is a probability of 0.65 to get attacked by cardiovascular disease.

Like all regression analyses, the logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

The following sample graphs give a brief idea of how the visualization could be made.







CONCLUSION:

Based on the above outcomes, it can be concluded that there is a huge scope for machine learning algorithms in predicting heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. Alternating Naive Bayes with PCA performed extremely well but use of PCA for Logistic Regression reduced the accuracy. SVM performed extremely well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data much more efficiently. A lot of research can also be done on the discussed as well as on other algorithms to use for a particular type of data.

REFERENCES:

[1] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[2] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Med-ical Data Classification" in Annual Conference on New Trends in Information &

Communications Technology Applications - march-2017

[3] Deeanna Kelley “Heart Disease: Causes, Prevention, and Current Research” in JCCC Honors Journal

[4] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dhavachelvan, P., Venkatachalapathy, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.