Name: Chandrashekhar

Section: K18AP

Reg No: 11802214

Roll No: 37

Course Code: INT247

LOVELY
PROFESSIONAL
UNIVERSITY

Transforming Education Transforming India

Topic: Heart Disease Prediction

Submitted To: Usha Mittal

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyses huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

INTRODUCTION

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries.

Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. The silver lining is that heart attacks are highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use; eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high risk patients because of the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where machine learning and data mining come to the rescue.

Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches.

METHODOLOGY

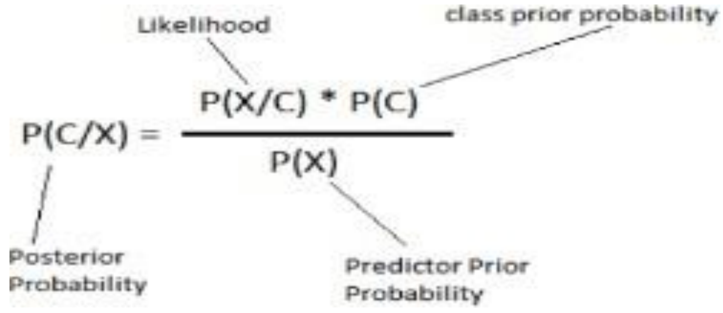
Data Pre-processing:

Data that we want to process will not be clean that is it may contain noise or it may contain values missing of we process we can't get good results so to obtain good and perfect results we need to eliminate all this, the process to eliminate all this is data cleaning. We will fill missing values and can remove noise by using some techniques like filling with most common value in missing place. Transformation: This involves changing data format to one form to other that is making them most understandable by doing normalization, smoothing, and generalization, aggregation techniques on data.

Integration: Data that we need not process may not be from a single source sometimes it can be from different sources we do not integrate them it may be a problem while processing so integration is one of important phase in data pre-processing and different issues are considered here to integrate.

Naive Bayes Classification:

The Naive-Bayesian classifier relies upon Bayes' speculation with Autonomy suppositions among attributes. A Naive-Bayesian output is definitely not hard to run, with no entrapped Repetitive parameter estimation which makes it particularly supportive for broad datasets in spite of its effortlessness, the Naive Bayesian classifier generally completes its job shockingly good and is broadly used in light of the fact that it frequently outflanks high order techniques which are complex. The Naïve Bayes treats every variable as independent which helps it to predict even if variables don't have proper relation.



K-Means Algorithms:

k-means clustering is one of clustering technique used to cluster Datasets based on nearest-neighbor here the data is clustered in k Clusters based on a similarity between them we are also fill missing values of data using this k-means. Once we clustered the data every dataset will come into any one of clusters by using this clusters if we have missing values in dataset we can fill those values as this are categorized into groups.

Now as this missing values are all cleared we can apply different prediction techniques on this for an example we can apply now as we know that for a dataset to be used for prediction in Naïve Bayes need to be pre-processed we can use this data for prediction in Naïve Bayes.

By different combination of using these algorithms we can achieve good accuracy. We reviewed different papers on heart disease prediction out of all prediction techniques and methods what everyone using when it comes to prediction is Naïve Bayes and decision trees we have different methods one which that we used here is ID3 algorithm.

Decision Tree Algorithm

The ID3 algorithm is one of old algorithm which is used for building decision trees in the process of building decision tree it handles missing values and removes outliers.

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 minimum of 2 branches Leaf nodes speaks to a grouping or choice. The highest choice hub in a tree which compares to the best indicator called root point. Choice Trees can deal with both all out and numerical information.

ID3 is algorithm which is used to build decision trees. ID3 has Some features like removing outliers, handling missing values and But there major disadvantage is to 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,

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

Proposed System

By the above experiment what we say is as Naïve Bayes results
And decision tree results may change so for every prediction we
Need not have a comparison of both the algorithms so get accurate
Results and in the same way if we use only a single algorithm
Which cannot pre-process data we even can't get good accuracy so
It's better to have combination of algorithms like k-means, ID3 and
K-means and Naïve Bayes.

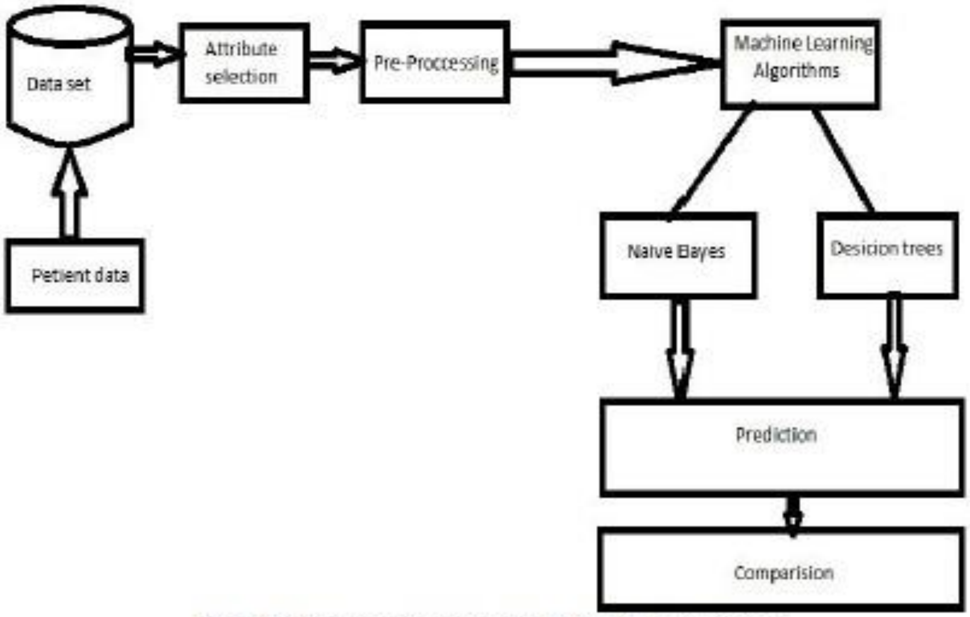


Fig. 3: System we suggest for the problem

Conclusion

In this what we found is during small datasets in some other cases Most of time decision trees direct us to a solution which is not Accurate, but when we look at Naïve Bayes results we are getting more accurate results with probabilities of all other possibilities But due to guidance to only one solution decision trees may miss Lead. Finally we can say by this experiment that Naïve Bayes is more accurate if the input data is cleaned and well maintained even though ID3 can clean itself it cannot give accurate results every time, and in this same way Naïve Bayes also will not give accurate results every time we need to consider results of different algorithms and by all its results if a prediction is made it will be accurate. But we can use Naïve Bayes consider variables as individual we can use combination of algorithms like Naïve Bayes and K-means to get accuracy.