International Journal of Engineering & Technology, 7 (2.32) (2018) 363-366

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Prediction of Heart Disease Using Machine Learning Algorithms

Rajesh N[1], T Maneesha[2], Shaik Hafeez[3], Hari Krishna[4]

[1,2,3,4]Computer Science Engineering, K L E F, Guntur, India
**Corresponding author E-mail: nrajeshcse@kluniversity.in*

Abstract

Heart disease is the one of the most common disease. This disease is quite common now a days we used different attributes which can relate to this heart diseases well to find the better method to predict and we also used algorithms for prediction. Naive Bayes, algorithm is analyzed on dataset based on risk factors. We also used decision trees and combination of algorithms for the prediction of heart disease based on the above attributes. The results shown that when the dataset is small naive Bayes algorithm gives the accurate results and when the dataset is large decision trees gives the accurate results.

Keywords:*Decision tree, Data mining, Heart Disease Prediction, Naïve Bayes, K-means, Machine learning..*

1. Introduction

The main topic is prediction using machine learning technics. Machine learning is widely used now a days in many business applications like e commerce and many more. Prediction is one of area where this machine learning used, our topic is about prediction of heart disease by processing patient's dataset and a data of patients to whom we need to predict the chance of occurrence of a heart disease.

2. Literature Survey

[2]. 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 focus on dissect information mining procedures which are required for medicinal information mining particularly to find locally visit illnesses, for example, heart infirmities, lung malignancy, bosom disease et cetera. Information mining is the way toward extricating information for finding inactive examples which Vembandasamy et al. performed a work, to analyze and detect heart disease. In this the algorithm used was Naive Bayes algorithm. In Naïve Bayes algorithm they used Bayes theorem. Hence Naive Bayes has a very power to make assumption independently. The used data-set is obtained from a diabetic research institutes of Chennai, Tamilnadu which is leading institute. There are more than 500 patients in the dataset. The tool used is Weka and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%.

[3]. Costas Sideris, Nabil Alshurafa, Haik Kalantarian and Mohammad Pourhomayoun have given a paper named Remote Health Monitoring Outcome Success prediction using First Month and Baseline Intervention Data. RHS systems are effective in saving costs and reducing illness. In this paper, they portray an upgraded RHM framework, Wanda- CVD that is cell phone based and intended to give remote instructing and social help to members. CVD counteractive action measures are perceived as a basic focus by social insurance associations around the world.

[4]. L.Sathish Kumar and A. Padmapriya has given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone. This paper 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. This proposed method helps the people not only to know about the diseases but it can also help's to reduce the death rate and count of disease affected people.

[5]. M.A.Nishara Banu and B.Gomathy has given a paper named Disease Predicting system using data mining techniques. In this paper they talk about MAFIA (Maximal Frequent Item set algorithm) and K-Means clustering. As classification is important for prediction of a disease. The classification based on MAFIA and K-Means results in accuracy.

[6]. Wiharto and Hari Kusnanto have given a paper named Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm. In this paper they exhibit an expectation framework for heart infection utilizing Learning vector Quantization neural system calculation The neural system in this framework acknowledges 13 clinical includes as information and predicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.

[7]. 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 The neural system in this framework acknowledges 13 clinical includes as information and pre-

dicts that there is a nearness or nonattendance of coronary illness in the patient, alongside various execution measures.

3. Methodology

3.1. Data Pre-Processing

Cleaning: 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 cant 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.

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

3.2. ID3 Algorithm

To do this we have many machine learning algorithms out of which we the more widely used methods are Naïve Bayes classification technic and decision tree construction, in this decision tree construction we have many algorithms one which we took for this ID3 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[2]. 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[8]. 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[2]. 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, X2,…. , Xn as indicated by the estimations of U.

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

3.3. Naïve-Bayes Classification:

The Naïve-Bayesian classifier relies upon Bayes' speculation with autonomy suppositions among attributes [7-13]. A Naïve-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 [1].

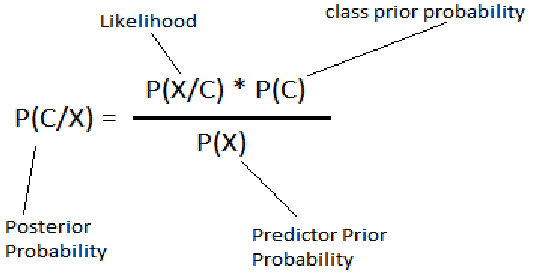

Fig. 1: Naïve Bayes proccess

- P(c|x) is the posterior probability of class (target) given predictor (attribute)
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

3.4. K-means

k-means clustering is one of clustering technic 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[6]. 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 technics 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[1]. By different combination of using this algorithms we can achieve good accuracy.

We reviewed different papers on heart disease prediction out of all prediction technics 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. We took a medical data of heart disease patients from UCI machine learning repository one of popular repository to get data for machine learning experiments it contains a record of nearly 300 patients we performed both this Naïve Bayes and ID3 technics on this training data using R tool. In R tool we used some 3rd party libraries like e1071 for implementing Naïve Bayes and rpart to construct decision tree. In the data set that we took for implementing this contains variables

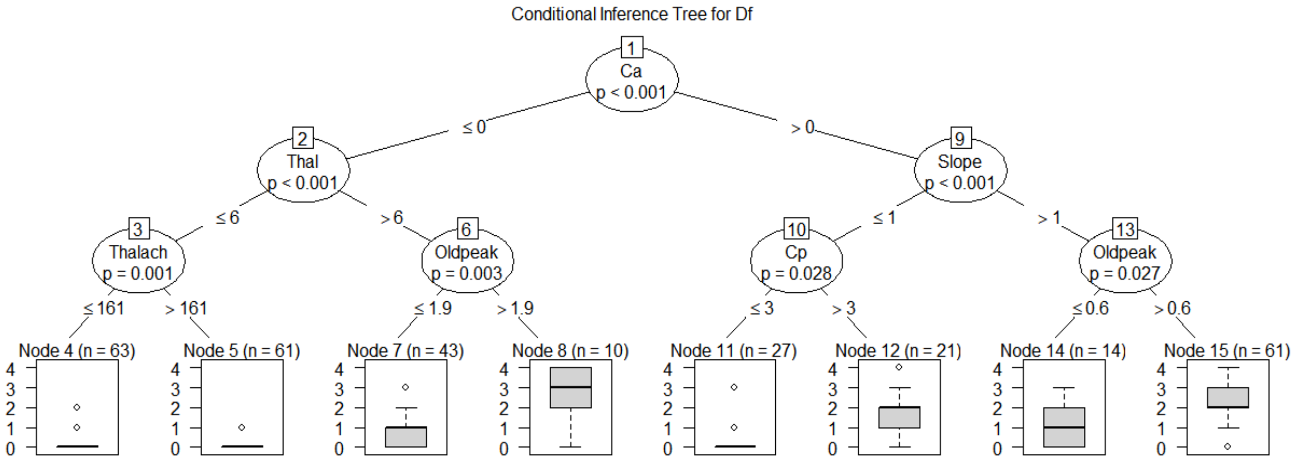

Table 1: data attributes for prediction

Attribute	Values and meaning
Age1	Age in year
Gender1	Value 1 and 0 for male and female
Cp1	Pain in chest Yes/No
Trest bps1	blood pressure during resting
Chol1	Cholesterol of serum in mg/dl
Fbs1	blood sugar during fasting
Restecg1	Resting electrocardiographic results
Oldpeak1	ST depression induced by exercise relative to rest
Slope1	The slope of peak exercise of ST segment. Value 1: upsloping Value 2: flat Value 3: down sloping
Ca1	Number of major vessels (0-3) colored by flourosopy
Thal1	3 = normal; 6 = fixed defect

Num1	Diagonal of heart disease
	Value 0: No Risk; Value 1: Low Risk;
	Value 2: Risk; Value 3: High Risk;
	Value 4: Higher Risk

Number which indicates rate of getting heart attack on a scale of 0 to 4. What we observed in the results are out of different executions on different data sets both algorithms predict with a good accuracy but when comparing both in most of cases decision tree is giving a result which is has less probability. We also observed that as Naïve Bayes treats all members of class as independent it can get all probabilities i.e., probabilities occurrence of different values of class members.

The result of decision tree:

```
> fltTree<-rpart( Num1 ~ Age1 + Sex1 + Cp1 + Trestbps1 + Chol1 +Fbs1 +Restecg1 + Thalach1 + Exang1 + Oldpeak1 + Slope1 + Ca1 + Thal1,df)
> predict(fltTree,df_new,type="vector")
 1
0.1478261
```

The result of Naïve Bayes is:

```
> model<-naiveBayes(Num1~.,data=df)
> predict(model,df_new,type="raw")

            0 1 2 3 4
[1,]   0.9549053   0.04493752   0.0001147186   1.119927e-05
3.123757e-05
```

This above results of decision tree and Naïve Bayes both are applied on same dataset taking same training set and the actual answer is 0. We can observe that in Naïve Bayes is showing the actual results with more probability that means we can predict 0 as answer but decision tree shows it as one which is approximately correct but not exactly correct.

Decision tree:

The above decision tree is constructed in R on the dataset that we used in this. To achieve that in R we used C50 library.

4. Proposed System

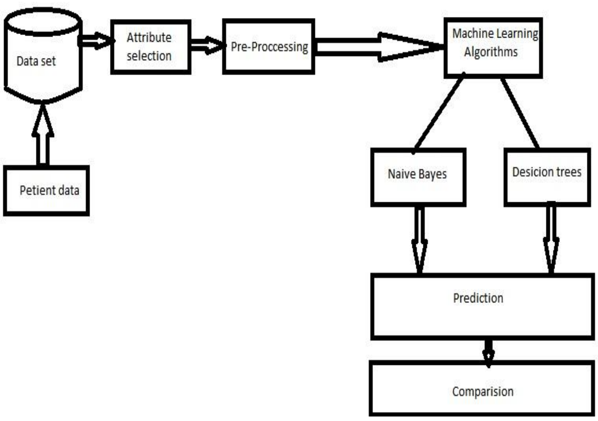

Fig. 3: System we suggest for the problem

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 its better to have combination of algorithms like k-means, ID3 and k-means and Naïve Bayes.

5. 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 it self 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.

References

[1] Sonam Nikhar, A.M. Karandikar “Prediction of Heart Disease Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June-2016 vol-2

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

[3] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics

[4] 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

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march-2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012

[7] 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.

[8] Alapatt, B.P., Kavitha, A., Amudhavel, J., "A novel encryption algorithm for end to end secured fiber optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amudhavel, J., Inbavalli, P., Bhuvaneswari, B., Anandaraj, B., Vengattaraman, T., Premkumar, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.

[10] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Bhuvaneswari Aadharshini, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] Amudhavel, J., Kodeeshwari, C., Premkumar, K., Jaiganesh, S., Rajaguru, D., Vengattatraman, T., Haripriya, R., "Comprehensive analysis on information dissemination protocols in vehicular ad hoc networks", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2058-2061.

[12] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges, opportunities and analysis on security based cloud resource virtualization", (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1458-1463.

[13] Amudhavel, J., Ilamathi, R., Moganarangan, N., Ravishankar, V., Baskaran, R., Premkumar, K., "Performance analysis in cloud auditing: An analysis of the state-of-the-art", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2043-2046.