ISSN No. 0976-5697

Volume 11, Special Issue I, May 2020

International Journal of Advanced Research in Computer Science

CONFERENCE PAPER

Available Online at www.ijarcs.info

HEART DISEASE PREDICTION USING DATA MINING

Shamanth DL
School ofC& IT
REVA University, Bengaluru-560064, India
Shamanthdl6793@gmail.com

Sudhakar
School of C& IT
REVA University, Bengaluru-560064, India
sudhakarkurutahalli@gmail.com

Nishchay M
School of C & IT
REVA University, Bengaluru-560064, India
nischay32@yahoo.com

Vikeshreddy
School of C & IT
REVA University, Bengaluru-560064, India
Vikeshreddyrc1@gmail.com

Prof.Sohara Banu
School of C &IT
REVA University, Bengaluru-560064, India
soharabanu.ar@reva.edu.in

ABSTRACT--Heart disease is one of the notable causes of death in worldwide everyday divination of heart disease is a critical challenge in area of medical data analysis. Machine learning (ML) has been shown to be successful in helping in making decision and prediction from huge quantity of data produced by medical health care industry. Medical data mining is an important area of data mining and considered as one of the important research field due to the application in medical health care domain. In medical data mining classification and prediction of medical dataset process challenges. It is difficult to medical practice to predict the heart attack as it complex task. The health sector at present contain the information that are hidden and which are important in making decision. Data mining algorithms such as Naïve Bayes algorithm, decision tree algorithm and random forest are applied in the research for heart disease prediction the result show the comparison between the three algorithm and selecting the best one among three, the Random forest algorithm will provide the more accuracy can compared to all the three algorithms.

INDEX TERM:_ Naive Bayes Algorithm, Decision Tree Algorithm, Random forest.

I. INTRODUCTION

Heart disease any disorder that affects the heart. Sometimes the term "heart disease" is used narrowly and incorrectly as a synonym for coronary artery disease. Heart disease is synonyms with cardiac disease but not with cardiovascular disease which is any disease of heart or blood vessels.The incidence of heart failure is very high, the treatment of cardiac disease is complicated as it involves multiple important risk factors, such as asthma, elevated blood pressure, high cholesterol, irregular pulse rhythm and many others. The sternness of the disease is classified based algorithm like Decision tree(DT), Naïve Bayes (NB) and Random forest (RF). The disease has to be taken proper care not doing so, may end up to early death. The objective of medical science and Data mining are used to detect various aspects of the disease.Data mining is a method of analyzing vast pre-existing datasets in order to produce new knowledge, including the processing of massive data sets that find trends and create connections to solve problems by data analysis. For diagnosis of heart disease, we come up with Decision Tree(DT),Naive Bayes(NB) and Random forest (RF) where the results obtained by these three algorithms are analyzed and the most accurate among these are mentioned,Decision Tree is a classification algorithm that is a decision support method that utilizes a tree-like decision-making paradigm and its possible implications, namely the expense and usefulness of future case outcomes. This is one way to view an algorithm that includes only conditional control statements, Random forest (RF) or Random decision forest are an ensemble learning method for classification, regression and other tasks they operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. In machine learning, Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models. Naïve Bayes has been studied extensively since the 1960's.

II. LITERATURE SURVEY

The purpose of the paper is to examine the numerous data mining strategies popularized in recent years for the prediction of heart disease. Many experts use only one method to treat cardiac failure and others use more than one methodology.

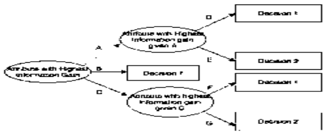
Anjan Nikhil Repaka, et.al. [1]:Work focuses on the treatment of cardiac failure utilizing past research and knowledge.Naïve Bayesian algorithm is used to build and achieve SHDP (Smart Heart Disease Prediction) and is used to diagnose risk factors linked to cardiac disease.In order to predict the likelihood of cardiac failure in a individual, characteristics such as age, cholesterol, sex, pulse, sugar etc. are obtained from diagnostic profiles.The data obtained are inputs to the Naïve Bayesian model to forecast heart attack.

Dhara B Mehta, et.al. [2]: In the research work,a data mining classification method for the data collection is introduced for a referral framework to assess the state of a person's heart and the risk level of cardiac failure.Data mining algorithms used include Logistic regression, Decision Tree, K Nearest Neighbor, Naïve Bayes and SVM dataset and the best accuracy obtained is 97.91 percent using SVM algorithm.

Sayali Ambekar, et.al. [3]: The work also deals with the issue of retrieving lost medical records, where records cleaning and imputation are used to transform partial data to full data. Prediction of heart attack is based on datasets using the Naïve Bayes and KNN algorithm. This research is generalized, disease incidence modeling is introduced utilizing organized data. Convolutionary neural network based unimodal disease risk prediction algorithm is used.The algorithm's precision is more than 65 percent.

Min Chen, et.al. [4]: Specific areas display particular characteristics of such environmental diseases, which may hinder machine learning algorithms to accurately predict chronic disease.In work, machine learning algorithms, is used effectively to diagnose cardiac attack in regions where there is a recurrent danger by experimenting with a updated model of prediction over real-life data obtained from hospitals across central China between 2013 and 2015.The latent factor model was used to collect insufficient data to complete the data for the study to be effective.Using structured and unstructured hospital info, a new algorithm named the Convolutionary Neural Network Multimodal Disease Risk Prediction (CNN-MDRP) is suggested.

Archana L.Rane, et.al. [5]: Work involves assessments of heart disease prediction programs methodically and on the basis of hybrid strategies of categorization of the methodology collected, tabulated and evaluated.Such methods are also categorized into two major categories: separate and mixed, also identified as controlled, unsupervised, composite and miscellaneous. The conclusion shows that only one data mining methodology functions good, but dealing with hybrid data mining techniques produces positive and effective performance.


Attribute with Highest Information Gain
A
B
C
Attribute with Highest Information gain given A
Decision F
Attribute with highest Information gain given C
D
E
F
G
Decision 1
Decision 2
Decision 4
Decision Z

SanchayitaDhar, et.al. [6]: The usage of data mining is capable of raising the amount of checks that are needed to be done to diagnose heart disease.In order to minimize the amount of deaths induced by heart failure, a fast and effective screening procedure must be conducted with greater sensitivity and precision .The goal of this work is to introduce accessible and accurate strategies for forecasting cardiac disease using machine learning approaches.Thus, a combination approach utilizing a Random forest classifier and a straightforward K-means algorithm has been suggested.The dataset is also analyzed using two separate machine learning algorithms, namely the J48 tree classifier and the Naïve Bayes classifier and the effects are contrasted.

Senthilkumar Mehta, et.al. [7]: The most significant cause for mortality in today's world is heart disease. Data mining and machine learning has shown effective in making decision and predicting the heart disease using the data collected by healthcare industry. The proposal work is based on a novel method which aims in finding significant features using machine learning techniques which results in good accuracy in prediction of heart disease. The model is made used with different features and combination, als9 with several classification techniques which is previously known.

The accuracy gained by this proposed model reaches up to 88.7%.

Liaquat Ali, et.al. [8]:Many automation techniques rely on the pre-processing of apps only.Work focuses on both the improvement of functionality and the exclusion of problems presented by the predictive model, i.e. issues that are under-fitting and over-fitting, through eliminating such issues, the model will do well in all datasets, i.e. data preparation and data checking. The X2 regression model is used to delete redundant features and an extensive scanning technique is used to look for an optimally optimized deep neural network (DNN).The precision of the software is up to 93.33%.

III. PROBLEM DEFINITION:

Provide insight into the various data mining methods that can be used in the automatic cardiac disease prediction method. Heart disease is one of the leading causes of death worldwide, and early diagnosis of heart disease is very critical. The computer supported the doctor's cardiac attack predictive program as a method for diagnosing heart disease. A survey had proved that about sixty percentage of population are suffering from heart disease. Due to lack of knowledge people are unaware of earlier heart disease prediction. In several cases it is noticed at final stage, where health care industries are working on it very efficiently to find out heart disease at stage itself. Most effective problem here comes with cost where many individuals cannot afford the cost hence they fail to get diagnosed and detect the heart disease at early stage. Thus, predicting and preventing heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives.

IV. METHODOLOGY

In this research of prediction of heart disease is done using data mining and machine learning techniques.

Here 3 different algorithms have been used:

1) Naïve Bayes algorithm
2) Decision tree algorithm
3) Random forest algorithm.

Naïve baye's algorithm

In machine learning naïve Bayes classifier are a family of simple "probabilistic classifier" based on applying Bayes theorem with strong independent assumption behave the future. They are among the simplest Bayesian network model. Naïve Bayes classifier are highly scalable, requiring a number of parameters, linear in number of variables (features/prediction) in the learning problem. Naïve Bayes is a simple technique for construing classifier model. That assign class label to problem intense, represented as vector of future value.

$$\underbrace{P(c|x)}_{\text{Posterior Probability}}=\frac{\overbrace{P(x|c)}^{\text{Likelihood}}\overbrace{P(c)}^{\text{Class Prior Probability}}}{\underbrace{P(x)}_{\text{Predictor Prior Probability}}}$$

$$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$$

And advantage of naïve Bayes is that it only requires a small number of training data sets to estimate the parameter necessary for classifier. Naïve Bayes is a conditional probability model given a problem intense to be classified represented by a vector X = (x_1, x_2, …,x_n) represent some n features (independent features) it assign to this intense probabilities.

P(C_k/x_1, x_2, …… , x_n)

K - Parable outcome / classifier.

Decision tree algorithm

In data mining, machine learning and statistics usage prediction modelling approaches. Items target value is predicted by building the decision tree(as predictive model) ton go from observations about an item, model where target variable can take discrete sets of values is called classification trees. Decision tree with target variable are continues value are called regression trees.

The formula used

1) Entropy
   i. Using one attribute

$$E(S)=\sum_{i=1}^{c}-p_i\log_2 p_i$$

   ii. Using two attributes

$$E(T,X)=\sum_{c\in X}P(c)E(c)$$

2) Information gain

$$Gain(T,X)=Entropy(T)-Entropy(T,X)$$

Random forest algorithm

Random forest is a trade mark word for an assembly classifier that consists of several decision trees and a class output that is the default output default of individual trees. The Random Forest is a set of plants, somewhat different. It randomizes the code, not the training cycle.

V. IMPLEMENTATION**:**

Classification using Naïve Bayes classifier.

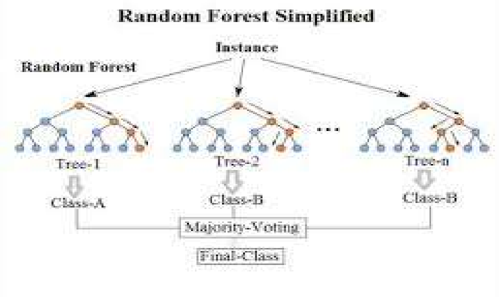
The Naïve Bayesian classifier is based on Baye's principle of independence assumptions between predictors.The Naïve Bayesian model is simple to construct, with no complex iterative parameter estimation, making it especially effective for very large datasets.It does relatively properly and is broadly used because it frequently out performs more state-of-the-art type strategies. Bayes theorem affords a manner of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naïve Bayes classifier assumes that the effect of the price of a predictor(x) on a given magnificence(c) is independent of the values of other predictors. This assumption is called elegance conditional Independence.

Classification using Decision tree algorithm.

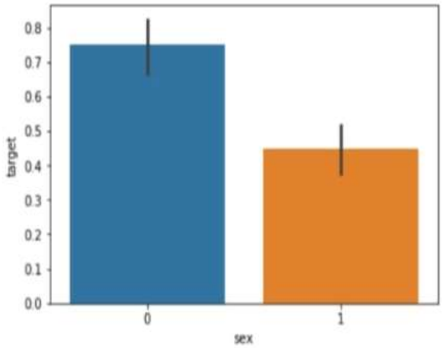
Decision tree constructs a classification or regression model in the context of a tree structure.It effectively splits the data set into smaller and smaller subsets, and at the same time the corresponding decision tree is incrementally created.The end product is a tree of judgment nodes and foliage nodes.There are two or three divisions of the judgment node.The Leaf node is a designation or judgment.In the decision tree leaving, ID3(Iterative dichotomies 3) was used to create a decision tree from a dataset.ID3 is a predecessor to the C4.5 algorithm which is usually used in the area of computer learning which natural language processing.

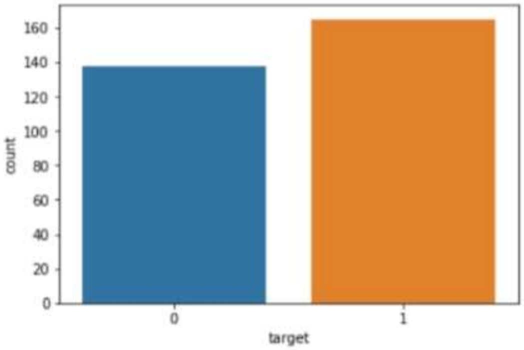
Classification using Random forest.

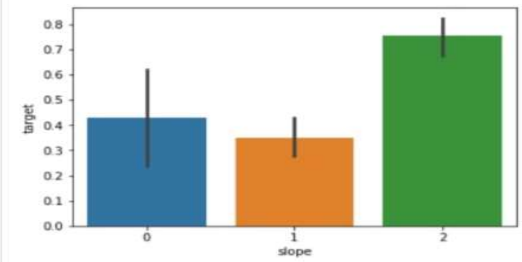
Random forest (or random forests) is a proprietary term for an ensemble classifier that is made up of several decision trees and outputs the class that is the mode of the individual trees output groups.

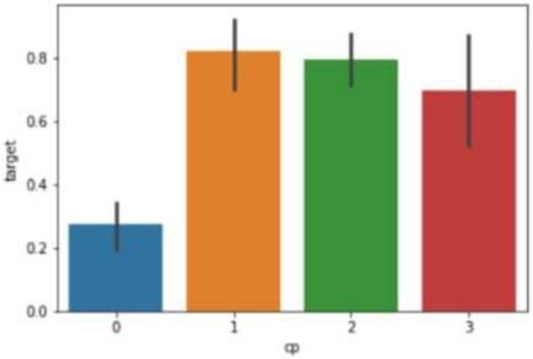


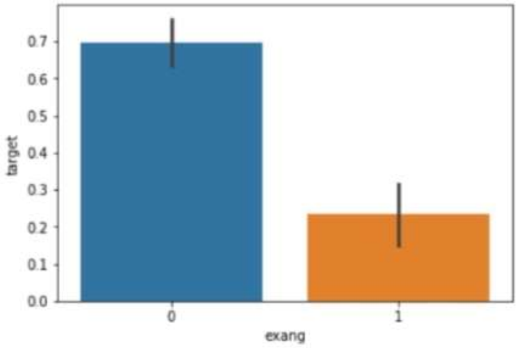
VI. OUTCOMES:

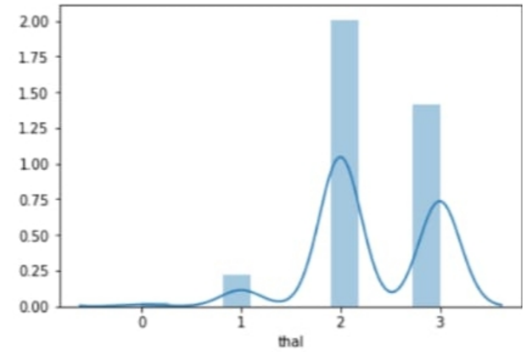


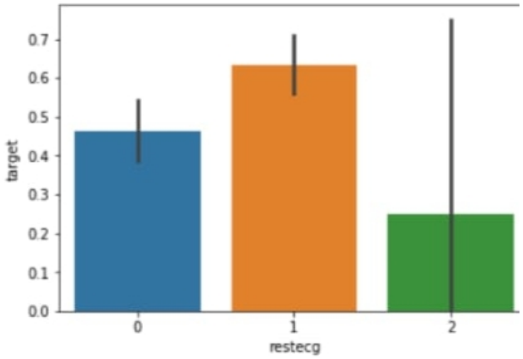


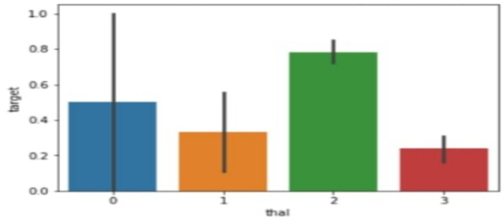


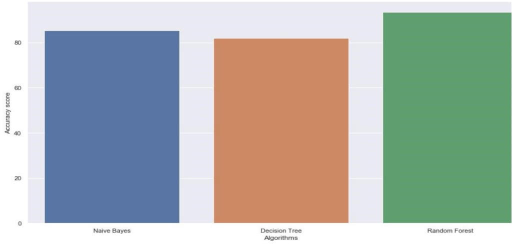












VII. CONCLUSION:

The main aim of this paper is to eradicate the difficulty of predictingthe heart disease. There are many means in which the heart disease is predicted but some are not that efficient which produces the late results which might be major cause to the death of the victim hence early and accurate prediction is very important also the algorithm used previously in other researches gives less accuracy where this exact problem wouldn't be found hence , here in this research we use 3 algorithms which are Naïve Bayes, Decision Tree and Random Forest,all these algorithms are used separately and with the same dataset and then the graphs are plotted further these are compared and the algorithm with the best result is produced.All these are done using data mining and machine learning techniques.

VIII. REFERENCES:

[1]. Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin(2019), "Design and Implementing Heart Disease Prediction Using Naïve Bayesian".

[2]. Dhara B Mehta, Nirali C. Varanagar(2019), "Newfangled Approach for Early Detection and Prevention of Ischemic Heart Disease Using Data Mining".

[3]. Sayali Ambekar, Rashmi Phalnikar (2018), "Disease Risk prediction by Using Convolutional Neural Network".

[4]. Min Chen, Yixue Hao, Kai Hwang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities".

[5]. Archana L. Rane (2018), "A Survey on Intelligent Data Mining Techniques used in Heart Disease Prediction".

[6]. SanchaitaDhar, Krishna Roy, Tanusree Dey, Pritha Datta, Ankur Biswas (2018), "A Hybrid Machine Learning Approach for Prediction of Heart Diseases".

[7]. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava (2019), "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques".

[8]. Liaquat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, Javed Ali Khan (2019), "An Automated Diagnostic System for Heart Disease Prediction Based on x^2Statistical Model and Optimally Configured Deep Neural Network".