ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 12, December 2019

Prediction of Heart Disease Using Machine Learning Techniques

Satish Chaurasiya [1], Dr.Neelu Nihalani [2]

Junior Research Fellow, University Institute of Technology, RGPV, Bhopal, India[1]

Professor, Master of Computer Application, University Institute of Technology, RGPV, Bhopal, India[2]

ABSTRACT: Heart disease can be detected prevented and diagnosed if predicted in a early stage. The predictive solution for cardiovascular risk estimation is extremely challenging can be performed by a medical expert based on his experience and patient's clinical data. The attempt to clinically screen the medical databases and predictive modelling through Machine Learning algorithm is regarded as a valuable and economical option for medical practitioners.The machine learning algorithm for heart disease prediction utilizes medical instances such as sex, blood pressure, cholesterol like 13 attributes as input and then these features are modelled to predict the likelihood of patient getting a heart disease. This paper provides an insight of various predictive machine learning algorithms like logistic regression, naive baye's, support vector machine and random forest tree used for prediction of heart disease with their respective accuracy.

KEYWORDS: Cardiovascular disease Coronary heart disease myocardial infarction, Machine Learning, Naïve Bayes, Support vector Machine, Logistic Regression, Random Forest Tree.

I. INTRODUCTION

The term "heart disease" is usually used interchangeably with the term "Cardiac disease". Different types of heart diseases are Coronary heart disease, Cardiomyopathy, Cardiovascular disease, Ischaemic heart disease, Heart failure ,Hypertensive heart disease, Inflammatory heart disease, Valvular heart disease. Cardiovascular disease (CVD) includes coronary artery disease generally refers to conditions that involve narrowed or blocked coronary-arteries (arteries that supply oxygen and blood to the heart) that can lead to a myocardial infarction commonly known as heart attack or stroke and angina (chest pain). Some of major symptoms of heart attack includes chest tightness, Shortness of breath, Nausea, Sweating and Fatigue, Indigestion, Heartburn, or stomach pain, Pressure in the upper back Pain that spreads to the arm.

World Health Organization report anticipates that 12 million people die every year because of heart diseases. Coronary artery disease and stroke account for 80% of CVD deaths in males and 75% of CVD deaths in females. The medical data of these heart patient is used by researchers to study the causes, affect, risk factors and preventive measures of the heart disease.

Among various Machine Learning techniques, the classification is one of the most important supervised learning techniques used for categorization of data patterns. The classification-based machine learning algorithms are used to predict membership function for labeling data instances. The Classification model is used for classifying future or unknown objects. The classification algorithm predicts instance categorical class (eg, negative and positive) and build classifier model based on the training set. The classification comprises two basic steps of learning and classification, in learning phase, training data are analyzed by classification algorithm and classification model is built. In the classification phase, test data are utilized to the model can be applied to classify data tuples whose class labels are unknown.

It is estimated that up to 90% of CVD may be preventable. On the basis of Machine learning (ML) proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. All these classification-based machine learning are using old patient record for getting predication about new patient. This predication system for heart disease assists doctors to predict heart disease in the early stage of disease resulting in saving millions of life.

II. RELATED WORK

Research conducted by ShusakuTsumoto [1] in year 2000 states that rapid growth in hospital information system to store laboratory examination in database it will be impossible for humans to deal with such huge amount of data. Thus it will be necessary to use some data mining techniques to draw same useful pattern from the database for reuse of stored data

Y. Alp Aslandogan, et. al. (2004), used three different classifiers namely K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and to conclude the decision of all above models into single outcome decision Dempsters' rule was used. This classification model based on the combined idea show increased accuracy [2].

Niti Guru, et. al. (2007), worked for forecasting of heart disease, Blood Stress and Blood Sugar by the aid of neural systems. Hearings were accepted out on example best ever of patients. The neural system is verified with 13 input variables such as blood pressure, period, angiography report etc. Controlled network was used for analysis of heart diseases. Back-propagation technique was accepted out as the support of training. The secretive data was nourished many times by the doctor; the acknowledged technique applied on the unidentified data since the judgments with trained data and caused a grade of possible ailments that the patient is inclining to heart disease. [3].

Sellappan Palaniappan, et. al. (2008), introduced IHDPS-Intelligent Heart Disease Prediction System by the use of data mining algorithm, i.e. Naïve Bayes, Decision Trees and Neural Network. Each process has its own authority to advance right outcome. The unknown designs and association amongst them have been used to paradigm this method. The IHDPS is web-based, user-friendly, mountable, trustworthy , stretchy and justifiable [4].

Chaitrali S. D., (2012), investigated a prediction of heart syndrome with the help of full amount of input characteristics. Thirteen medical attributes like blood pressure, sex, cholesterol etc. were recycled to predict the cardio vascular disease in a patient. He also introduced two new and different attributes like smoking and obesity. Data mining algorithms like Decision trees, neural networks and naïve baye's were used for analysing the heart disease from the clinical database. The concert of these prediction depends on the accuracy provided by the system. The accuracy provided by decision tree is observed 99.62%, neural network is 100% and naïve baye's is 90.74% respectively[5].

Monika Gandhi et.al, [6] used Naïve Bayes, Decision tree and neural network algorithms to analyse the medical dataset. There are a huge number of features involved in data set. So, there is a need to reduce the number of features, Feature selection was performed for this purpose. On doing this, they say that performance time is reduced. They made use of decision tree and neural networks.

Kiyong Noh et al. [7] introduced associative classification method for the extraction of multi-parametric features by assessing HRV (Heart Rate Variability) from Electro CardioGram, data pre-processing and heart disease pattern. The dataset consists of 670 peoples, distinguished into two groups, namely normal people and patients with heart disease, were employed to carry out the prediction for the associative classifier.

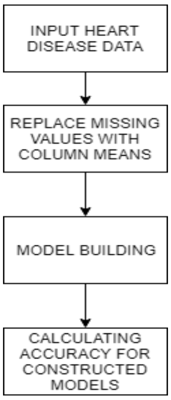
K. Polaraju et al, [8] proposed Prediction of Heart Disease using Multiple Regression Model and it shows that Multiple Linear Regression is appropriate method for predicting heart disease. The work is performed using training data set consists of 3000 instances with 13 different medical attributes which has mentioned earlier like age,sex, blood pressure etc. The data set is seprated into two parts that is 70% of the data are used for training and 30% used for testing purpose. Based on the results, it is clear that the classification accuracy of Regression algorithm is better as compared to other algorithms.

R. Sharmila et al, [9] proposed to use non- linear classification algorithm for heart disease prediction. It introduces to use big data tools such as Hadoop Distributed File System (HDFS), Mapreduce along with SVM to predict heart disease with optimized attribute set. This work made an investigation on the use of different data mining methods for predicting heart diseases. It suggests to use HDFS for storing large data set in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using Support Vector Machine. SVM is used in parallel fashion to yielded better computation time than sequential SVM.

A.Sudha et.al [10] discusses the data mining technology to predict heart disease. They also propose an architectural diagram which including given steps – dataset collection, normalization and pre-processing, dimensionality reduction using Principal Component analysis, feature subset selection, classification algorithm and result analysis. They performed the analysis using three classifiers decision tree, Naïve Bayes and neural networks. They conclude that neural networks perform well than other classifiers.

III. PROPOSED METHODOLOGY

In this paper, various machine learning methods are compared for predicting the risk of coronary heart disease in the patients based on their medical data. The following is the flowchart for proposed methodology:


INPUT HEART DISEASE DATA
REPLACE MISSING VALUES WITH COLUMN MEANS
MODEL BUILDING
CALCULATING ACCURACY FOR CONSTRUCTED MODELS

Data source We performed simulation on dataset called Heart dataset. The dataset is available in UCI Machine Learning Repository [11]. Dataset contains 303 records divided into two set training and test data in the ratio 80:20 each having 13 input variable and 1 predictor variable. A list of all those features is given below.

age	Age of patient	thalach	maximum heart rate achieved
sex	Sex, 1 for male, 0 for female	exang	exercise induced angina (1 yes)
cp	chest pain	oldpeak	ST depression induc. ex.
trestbps	resting blood pressure	slope	slope of peak exercise ST
chol	serum cholesterol	ca	number of major vessel
fbs	fasting blood sugar larger 120mg/dl (1 true)	thal	thalassemia (3 normal; 6 fixed defect; 7 reversable defect)
restecg	resting electroc. result (1 anomality)	num	diagnosis of heart disease (angiographic disease status)

Machine learning algorithms can be used for regression and classification this paper mainly focus on the following classification techniques

Naive Bayes : This classifier is a powerful probabilistic classifier based on Baye's theorem and is used when the dimensionality of the input is very large. In this classification technique conditional independence is used, which means that an attribute value on a given class is independent of the values of other attributes. The Bayes theorem is as follows:

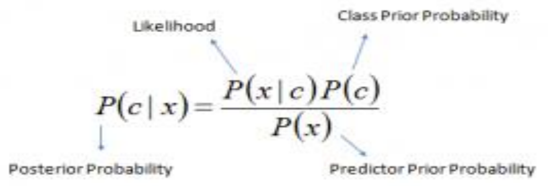
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 12, December 2019



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

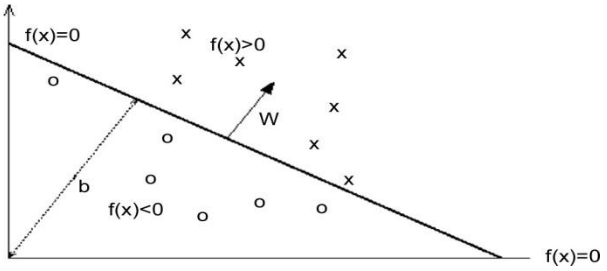
$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ 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*

Logistic Regression: Logistic Regression is a method used for classification when the dependent variable is dichotomous (binary), it analyses a dataset which has a one or more independent variable and gives prediction variable. Logistic Regression aims to predict the best relationship between the dependent and independent variables. logistic regression model uses the logit function to squeeze the output of a linear equation between 0 and 1. The logit function is defined as:

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Support Vector Machines: A Support Vector Mechanism (SVM) is a supervised classifier formally defined by a identifying the hyper plane which maximises the margin between two classes. In 2-D space, this hyper plane is a line dividing a plane into two parts wherein each class lay on either side of line. The points that lies in the separating Hyper plane is called Support Vectors. The distance between the Canonical and Separating hyper plane is known as Margin. Sequential minimal optimization a variant of SVM which breaks the problems in to sub-problems and then solve it analytically.



Random forest : Random forest is one of the most accurate and widely used machine learning algorithms, which ensembles large number of individual decision trees. Random Decision Tree firstly considers many decision trees before giving an output. Random forest algorithm uses a voting system for classification where it decides the class. RF is a combination of tree-structured classifiers {h(x, n)}, where for “x” data input and “n” are distributed random trees for the classification of data. Each one of the decision tree in random tree-structured forest, cast a vote that indicates the decision about class of data. RF uses the Gini-index for determining the final class in each tree. This algorithm chooses optimal attributes from “M” total number of input attributes at random for each tree. With this selected attribute, the best possible split is created using the Gini index to develop a decision tree model. This is an iterative process for each of the branches until the terminating nodes are too small to split further. For data set X having "n" classes, Gini-index, Gini(X) can be defined by:

$$Gini(X)=\sum_{i=1}^{n}(Rj)^{2}$$

where "Rj" is the relative frequency of class j in data set "X".

IV. SIMULATION RESULTS

The accuracy of the predictive models is evaluated by confusion matrix obtained by implementing predictive models on the test data. The confusion matrix show the correctly predicted as well as incorrectly predicted values by a classifier. The sum of TP and TN, from the confusion matrix, is the number of correctly predicted entries by the classifier.

	Heart disease 0	No Heart disease 1
Heart disease 0	TP	TN
No Heart disease 1	FP	FN

Table 4.1 Confusion matrix

	0	1
0	30	8
1	5	18

Table 4.2 Naive Bayes Confusion matrix

	0	1
0	32	9
1	3	17

Table 4.3 logistic regression Confusion matrix

	0	1
0	32	10
1	3	16

Table 4.4 Random Forest Confusion matrix

	0	1
0	32	9
1	3	17

Table 4.5 SVM Confusion matrix

The accuracy of the predictive machine learning algorithms discussed in this paper can be tabulated from the above confusion matrices

Predective Model	**Accuracy in %**
Naive Bayes	86.77
Logistic regression	80.32
Support Vector Machine	80.32
Random Forest	75.40

Table 4.6 Accuracy of predictive models

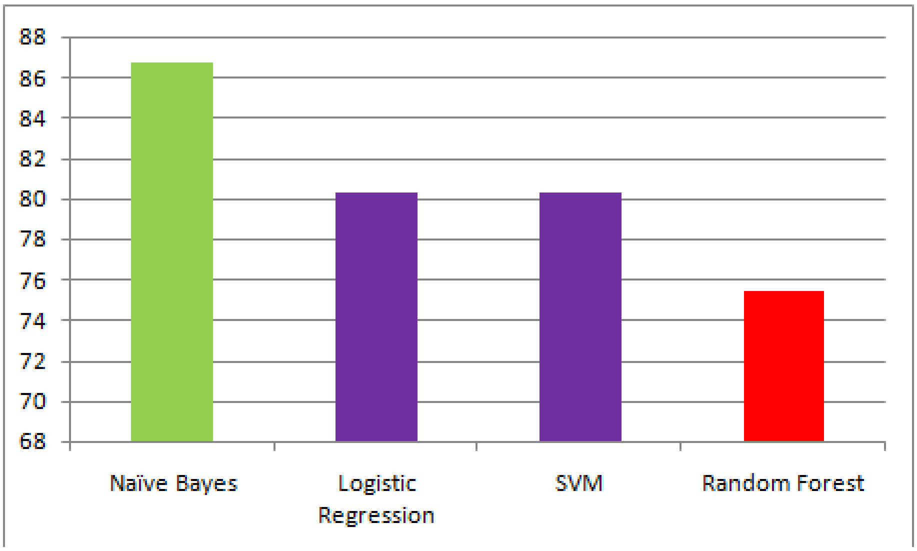


Figure 4.1 Accuracy Comparison

V. CONCLUSION AND FUTURE WORK

In this paper, we carried out an experiment to find the accuracy of different machine learning based predictive models. We selected four popular classifiers considering their qualitative performance for our experiment. We also choose one of the heart dataset available at UCI machine learning repository. Naïve bayes classifier is the best in performance. In order to compare the classification accuracy of four machine learning algorithms, classifiers are applied on same data and results are compared on the basis of correct classification rate and according to experimental results in table 4.6, it can be concluded that performance is closely competitive but Naïve base classifier is the best as compared to Support Vector Machine, Random Forest Tree and Logistic Regression. In future we would like to extend Our experiment on different data set to draw a more general conclusion and eliminate misclassification, as false prediction in health care could not be accepted.

REFERENCES

1. ShusakuTsumoto,” Problems with Mining Medical Data”, 0-7695- 0792-1 I00@ 2000 IEEE.
2. Y. Alp Aslandoganet. al.,” Evidence Combination in Medical Data Mining”, Proceedings of the international conference on Information Technology: Coding and Computing (ITCC’04) 0-7695-2108-8/04©2004 IEEE.
3. Niti Guru, Anil Dahiya, NavinRajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1, January - June 2007
4. SellappanPalaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (IJCSNS), Vol.8 No.8, August 2008
5. Chaitrali S. Dangare, Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques”, International Journal of Computer Applications (0975 888)Volume 47No.10, June 2012
6. Monika Gandhi,Shailendra Narayan Singh,Prediction in heart diseases using techniques of data mining (2015).
7. Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer, Vol:345, pp: 721- 727, 2006.
8. K. Polaraju, D. Durga Prasad, “Prediction of Heart Disease using Multiple Linear Regression Model”, International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017.
9. R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, International Journal of Computer Science and Engineering, May 2018.
10. A.Sudha, P.Gayathri, N.Jaisankar Effective analysis and prediction model for stroke disease using classification methods (April 2012)
11. Cleveland database: http://archive.ics.uci.edu/ml/datasetss/ Heart+Disease/
12. Sanjay Kumar Sen ,Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms ,ijecs, Volume 6 Issue 6 .
13. Jan M, Awan AA, Khalid MS, Nisar S, Ensemble approach for developing a smart heart disease prediction system using classification algorithm, https://www.dovepress.com/ensemble-approach-for-developing-a-smart-heart-disease-prediction-syst-peer-reviewed-article.