Biosc.Biotech.Res.Comm. Special Issue Vol 13 No 14 (2020) Pp-287-290

BBRC
Bioscience Biotechnology
Research Communications

Prediction of Heart Disease using Machine Learning Algorithm

Viraj S. Varale[1*] and Kalpana S. Thakre[2]

[1]*Student (Final Year B.E. IT), SCOE-Sinhgad College of Engineering Pune, India,*

[2]*Professor, Sinhgad College of Engineering Pune, India*

ABSTRACT

Human hearts suffer through various heart ailments. There are several diseases related to the heart like cardiomyopathy, Aorta diseases, Coronary Heart Disease (CHD) and arrhythmia which majorly contributes mortality and morbidity rates worldwide. One in 4 deaths in India is now because of cardiovascular disease with ischemic heart disease. The biggest challenge to overcome is the prediction of cardiovascular diseases via data analysis in the clinical domain. Now a day's large number of data is produced in health care and wellness industry. Finding meaningful data and patterns is the urgent need to make the proper regulations and forecasting. We proposed a framework for predicting a heart disease using three different algorithms: Random forest, Naive Bayes, and logistic regression. Proposed system uses Cleveland dataset from machine learning UCI repository for training and testing of the model. This model imbibes various significant features and classification techniques to predict the results. We also compare the results of proposed system with the algorithms that are existing in the literature, on the same dataset and it is observed that model produce an enhanced accuracy performance of 94.73%.

KEY WORDS: MACHINE LEARNING, HEART DISEASE, RANDOM FOREST, LOGISTIC REGRESSION, NAIVE BAYES, VOTING CLASSIFIER, CHI SQUARE METHOD, CLEVELAND DATASET, DECISION TREE, SUPPORT VECTOR MACHINE.

INTRODUCTION

Nowadays, whole world is suffering from Coronary artery and is a crucial source of death rate all over the world. According to a report prepared by various national and international health organizations, cardiovascular disease (CVD) caused 17.5 million (30%) of the 58 million deaths worldwide. A daunting increase is observed over last 20 years in the widespread presence of cardiovascular deaths in our country and other countries in Asia.

ARTICLE INFORMATION

**Corresponding Author: viraj.varale07@gmail.com*
Received 19th Oct 2020 Accepted after revision 29th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

Clarivate Analytics   crossref

NAAS Journal Score 2020 (4.31)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: http//www.bbrc.in/
Doi: http://dx.doi.org/10.21786/bbrc/13.14/67

In India, heart disease becomes the major cause of disease. Heart disease caused 17.8% of total demises. Demises due to heart diseases increased 15 lakhs in 2016 if compared with the statistics of year 1990. Heart disease is more deadly in males than in females. As per the report given by "The Wire" in 2016 more than half of the total heart disease demises in India were in people not older than 70 years. This condition was increased in the states of India, where the health care system is less equipped with cutting edge technology, which is a major challenge posed to the health systems. An urgent attention and action require to be provided in all the states of India for decreasing the deaths that are caused by cardiovascular disease in the economically productive age groups. Heart disease, stroke, diabetes and cancers are the main alarming causes that India has witnessed in last 25 years in India.

Statistical evaluation of the diseases like heart disease, cancer, chronic respiratory disease, and diabetes states that the prevalence of such disease increased in India from 1990 to 2016. A total death caused due to heart disease has almost double in the past 25 years and it is significantly increasing at alarming rate. Latest revolutionary technologies like machine learning, artificial intelligence and big data can be used to rehabilitate heart disease to reduce premature deaths. We have studied various machine learning techniques for heart disease. As per methods that is existing in the literature lot of algorithms are used to develop similar kind of systems but the results are not satisfactory. We have studied hybrid random forest and linear model (HRFLM) for prediction of heart disease. HRFLM model achieved 88.7% accuracy. HRFLM model proposed implementation of random forest and linear model algorithm for prediction. The proposed work suggests the prediction of heart disease using supervised machine learning algorithm using random forest, Naive Bayes and logistic regression algorithm by using voting classifier algorithm to integrate these algorithms into a single model.

Literature Survey: Forecasting of heart disease is an important provocation in the area of health care industry. Many of the researchers have worked in this particular domain. Some of the research paper that are existing in the literature are explained here.

[Senthilkumar Mohan, 2019] have proposed a system for the prediction of heart disease using random forest and linear model algorithm. This research paper shows that use random forest and linear model gives better accuracy than any other algorithm like decision tree, genetic algorithm and neural network in these particular areas. [Resul Das, 2009] proposed a framework that uses SAS based technology for forecasting of heart disease. They developed an ensemble based neural network-based method by combining the posterior probability for the prediction of values. This neural network-based model obtained an accuracy of 89.01%, reactivity 80.95% and 95.91% specificity values. [Hamidreza Ashrafi Esfahani, 2017] proposed a cardiovascular disease detection system that uses a new ensemble classifier.

They proposed a hybrid algorithm to increase the accuracy of proposed method. They collected the patient's data from Cleveland dataset from UCI repository and applied discovery pattern algorithms like Decision tree, Neural Networks, Rough Set, SVM, Naive Bayes algorithm. They finally compared their accuracy and prediction. Based on results the hybrid method achieved an F-measure of 86.8%. [V. Krishnaiah, 2015] proposed a heart diseases detection system using Naive Bayes Algorithm. This algorithm is used to classify the data set. The F-measures values they observed are 74% accuracy, 71% precision, 74% recall and 71.2%. In this paper author [Suganya Ramamoorthy, 2016] used naïve Bayes and particle swarm optimization technique to design a forecasting system of heart disease. Particle swarm optimization technique is an efficient evolutionary computation technique which selects the most optimum features which contribute more to the result. This system obtained 87.91% of accuracy.

Proposed Method: In this study we have used Anaconda and Jupyter notebook to perform heart disease classification. It is easy to use and provides simple user interface. First step to create machine learning model is data pre-processing phase followed by feature selection using chi2 method. We train and test ML model using UCI Dataset. We used Random Forest, Logistic Regression, Naive Bayes and Voting Classifier algorithm from sklearn library. We used python libraries numpy for mathematical operation, pandas for data manipulation, sklearn for importing machine learning algorithms and seaborn, matplotlib for ploting graphs We evaluate performance of the model using confusion matrix.

Figure 1: System Architecture final prediction model

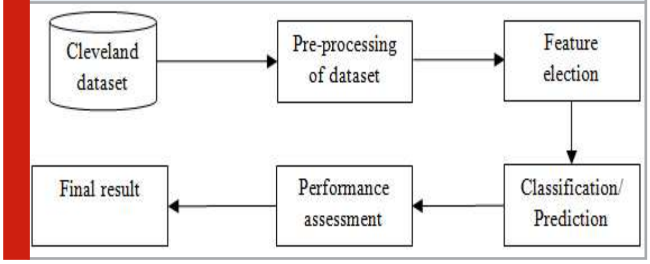


Figure 1 depicts the system architecture of final model. 6 steps in architecture diagram describes the development of final prediction model which are explained below.

UCI Dataset: Proposed method uses the patient's data form Cleveland dataset of UCI repository. This dataset contains total 13 features with class label in which 11 contains vital clinical records and 303 instances. We selected 8 features from 13 features to create model. 1. Chol 2. Thalach 3. Oldpeak 4. Thal 5. Cp 6. Ca 7. Exang 8. Age these 8 features are selected using chi square method.

Data Pre-processing: This is the first step for the implementation of proposed method in this step the Cleveland dataset is processed and anomalies are removed this dataset contains 303 patient records and we have used all instances of dataset for training and testing of the proposed model. During this process it is observed that 142 records shows the presence of heart disease by value 0 and 161 records shows the absence of heart disease by value 1. It is also observed that males are more prone to heart disease if compared with female.

Feature Extraction

We used chi square ($\chi 2$) method for feature selection
O - Observed number
E - Expected number
Step 1: for O = 1 TO N { (O – E).
Step 2: (O –E) 2.Step 3: [(O - E) 2 / E]}

Classification Using Machine Learning Algorithm: We used 3 different algorithms for forecasting of heart disease and a voting classifier algorithm that will give the final prediction results

1. Random forest (RF)
2. Logistic regression (LR)
3. Naive Bayes (NB)
4. Voting Classifier (VC)

Random Forest

Precondition: T:= (p1, q1), . . . ,(pn, qn), features F, and D are number of trees in forest.

```
RandomForest(T, F)
R ← null
for i = 1, . . . , D {
T (i) ← A bootstrap sample from T
Ri ← Decision_tree(T (i) , F)
R ← R ∪ {hi}
end for
return R }
function Decision tree(T , F) {
for each node:
f = subsection of F
divide on feature f
return the training model of decision tree}
```

Logistic Regression

X-Matrix of input features

Y- Class label 0/1 (yes/no)

β - Coefficient matrix of X.

B0- Y intercepts.

Step 1: Mathematical formula for Logistic regression

Z={ β0+ β1X1+ β2X2+... βm Xm.} Z= βX.

β = { β0 ,β1 ,β2...βm}. X= {X1, X2, X3...Xm}.

Step 2: Calculate β by using maximum likelihood estimation (MLE)

$$Ln(\beta)=\sum_{i=0}^{n}\{(y_i)LnP(X_i\mid\beta)+(1-y_i)Ln(1-(P(X_i\mid\beta))\}$$

Step 3: Sigmoid function

$$g(z)=\frac{1}{(1-e^{-z})}$$ where 0<=g(z)<=1

Naive Bayes

Step 1: find the prior probability for given class labels. P (yes) and P (no)

Step 2: Find Likelihood probability of each attribute for each class

P(B|A) where B-features and A- class labels.

Step 3 : Put these value in Bayes Formula and calculate posterior probability for class yes and class no

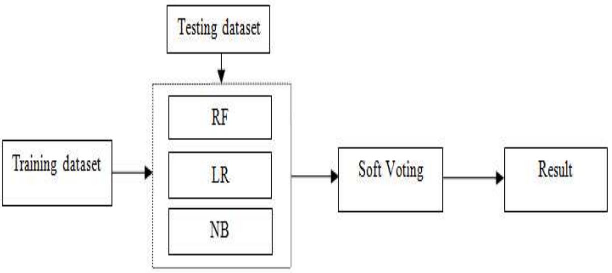
$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)}$$

A= {yes, no} B= Features

Step 4: result will be the class with higher probability.

Voting Classifier: Voting classifier is ensemble method used to aggregate the prediction of each classifier and predict the class that gets most votes. A voting classifier is used to find the result that is based on probability. As shown in figure 2.

Figure 2: Voting Classifier



RESULTS

Chi Square Test: Different performance evaluation terms such as accuracy, precision, recall and error in classification have been calculated for performance efficiency of this model. Accuracy calculates the percentage of correctly predicted true positive and true negative values. Precision calculates corrective predicted value in terms of percentage. We used confusion matrix to calculate accuracy, precision, recall and error rate.

Table 2. Chi square score

Features	Score
Age	23.29
Gender	23.91
Chest pain	81.68
Blood pressure	47.70
Cholesterol	173.10
Fasting Blood sugar	0.23
Ecg	10.02
Heart rate	110.13
Exang	57.79
Oldpeak	89.43
Slope	47.50
Ca(Fluoroscopy)	74.36
Thal	85.30

Table 2 shows the score of each feature calculated by using chi square method for feature selection. From the chi square result, we observed that the feature Chol scored 173.10 highest among all and feature Fasting blood sugar scored 0.23 lowest among all features as shown in figure 2. We removed irrelevant features selected only those features which are relevant and help to increase the accuracy of the prediction model.

Confusion Matrix: Cleveland dataset is used for tutoring and examine the data values. 75% of data is used for tutoring and 25% of for examining. We used 76 instances from Cleveland dataset for testing. Table 3 shows TP, TN, FP, FN values calculated by comparing Actual class label of 76 instances and value predicted by prediction model.

Figure 2: Chi square score

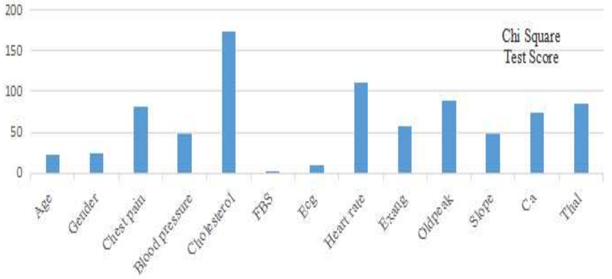


Table 3. Confusion matrix result

Algorithm	TP	FP	FN	TN	Instances used for testing
Decision tree	31	7	8	30	76
SVM	20	18	5	33	76
Random forest	32	6	2	36	76
Naïve Bayes	35	3	5	33	76
Logistic regression	34	4	2	36	76
Voting classifier	36	2	2	36	76

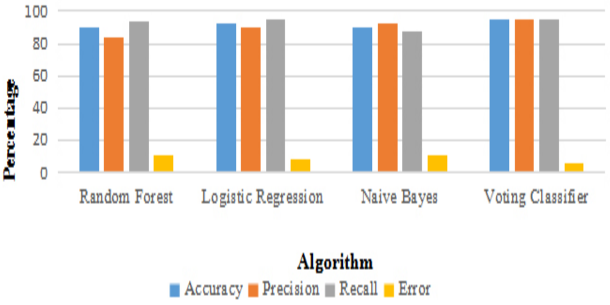
Table 4. Comparing different algorithms

Algorithm	Accuracy	Precision	Recall	Error
Decision tree	80.26	81.57	79.48	19.47
SVM	69.73	52.63	80	30.27
Random forest	89.47	84.21	94.11	10.52
Naïve Bayes	89.47	92.10	87..50	10.52
Logistic regression	92.10	89.47	94.44	9.21
Voting classifier	94.73	94.73	94.73	5.26

Table 4 compares result of Decision tree, Support vector machine, Random forest, Naive Bayes, Logistic regression and voting classifier (final prediction model). The maximum percentage of accuracy is achieved by our prediction model in comparison with existing heart disease prediction models.

System Performance Result: The prediction model is train and test using 8 features and the accuracy is evaluated for modelling techniques. The prediction model obtained 94.73% accuracy. Features are selected using chi square test. Table 4 shows accuracy, classification error, precision, and recall of various machine learning algorithms in percentage. Figure 3 is graphical representation of Accuracy precision recall and error rate of random forest, naive Bayes, logistic regression and voting classifier.

Figure 3: Confusion matrix result



REFERENCES

Chun-An Cheng, Hung-Wen Chiu, "An artificial neural network model for the evaluation of carotid artery stenting prognosis using a national-wide database", IEEE, ISBN NO: 978-1-5090-2809-2, 11-15July 2017, pp 2566-2569.

Hamidreza Ashrafi Esfahani, Morteza Ghazanfari, "Cardiovascular disease detection using a new ensemble classifier", IEEE, ISBN NO:978-1-5386-2640-5, 22-22 Dec. 2017, pp 1011-1014.

Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, "Effective diagnosis of heart disease through neural networks ensembles", ScienceDirect, Vol. 36, ISSN: 0957-4174, May 2009, pp 7675-7680.

Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE, Vol. 7, ISSN: 2169-3536, 2019, pp 81542-81554.

Suganya Ramamoorthy, S. Rajaram, A. Sheik Abdullah, V. Rajendran, "A Novel Feature Selection Method for Predicting Heart Diseases with Data Mining Techniques", Vol. 15, ISSN: 1682-3915, January 2016, pp 1314-1321.

Uma N. Dulhare, "Prediction system for heart disease using Naive Bayes and particle swarm optimization", ResearcGate, Vol. 29, ISSN: 0970-938X, January 2018, pp 2646-2649.

V. Krishnaiah, G. Narsimha, N. Subhash Chandra, "Heart Disease Prediction System Using Data Mining Technique by Fuzzy K-NN Approach", Springer, Vol. 337, ISBN NO: 978-3-319-13728-5, 2015, pp 371-384.