2021 5th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, India. Oct 22-23, 2021

Feature Optimization Based Heart Disease Prediction using Machine Learning

D.P.Yadav
GLA University Mathura
dhirendra.yadav@gla.ac.in

Prabhav Saini
GLA University Mathura
prabhav.saini_bca18@gla.ac.in

Pragya Mittal
GLA University Mathura
pragya.mittal_bca18@gla.ac.in

Abstract— **Heart disease is a spontaneous, treacherous, and fatal disease. It is a group of several states that result in abnormal functioning of the heart. Based on the several pathology test report heart disease is identified by a doctor. The manual heart disease prediction is time consuming and error prone. Therefore, in the present study an automated system based on the performance analysis of several machine learning techniques has been developed. First, the well-known machine learning algorithm Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes and Random Forest applied on the dataset for the prediction of heart disease. To avoid bias performance 3-fold cross validation is applied. The highest average accuracy of 87.78% is obtained by the Naïve Bayes. The performance of the model is acceptable. Further, we have applied genetic algorithm on the dataset to optimize the features. After, optimization the highest average accuracy of 96% is achieved by the naïve Base.**

Keywords— Heart disease, Prediction, Optimization, Machine learning.

I. INTRODUCTION

Heart disease is also known for cardiovascular disease refers to any condition hitting the heart. This disease has been the leading cause of death and influences all genders and ethnic groups [1]. Due to damage of blood vessels oxygen level in the heart decrease and heart attack occurs. The major cause of heart attacks genetically inherited high blood pressure, high cholesterol, smoking, mental stress [2]. PTSD is the most acute and broadly studied form of stress related disorder, characterized by re-experiencing, avoidance, negative perception and the traumatic event. It has been reported that nowadays babies also suffer from several types of heart disease. The acquired heart disease mainly characterized as a coronary artery disease. Some coronary artery supply nutrients, blood and oxygen to our heart. If these arteries get damage than a plank is deposited to the vessels and blood supply in heart get effected [3]. Due to this congestive heart failure happens and a patient may die. Weak or ill heart tissue and abnormal heart valves are the two most common causes. Because the valves are too narrow, they may not allow enough blood to pass through. Alternatively, the valve may leak, allowing blood to flow back into the heart. This can cause the heart to beat too quickly or too slowly [4].

Extremely fast heartbeats may cause the heart to stop pumping blood. A normal heartbeat is required for the heart to effectively pump blood [5]. If the heart beats pumped too quickly, there may not be enough time for blood to enter the chambers, resulting in insufficient blood flowing through the heart with each beat; if the heart beats too slowly, there may not be enough reductions of the heart to supply the body with the blood that it requires. Pain in the chest, difficulty breathing, palpitations, swelling, and cyanosis are common symptoms of heart disease. There are the various complications of having heart disease like an unexpected loss of heart function cardiac arrest, Heart stroke, harm to a piece of the heart muscle, a brain stroke, harm to the brain providing arteries, thickening of the artery can also lead to internal bleeding, Blocking of blood supply to various parts of the body like hands or legs [6].

In the past several research related to heart disease have been conducted. In this regards, Venkatalakshmi and Shivsankar et al. [7] applied machine learning approaches like Random Forest, Naïve Bayes have been applied to predict the heart diseases. They trained both the models using 13 features. The Naïve Base model achieved highest classification accuracy compared to Random forest. Jindal et al. [8] also compared the performance of logistic regression, KNN and Random forest for the prediction of heart disease. They concluded that highest accuracy of 88.52% was achieved using KNN. McPherson et al. [9] described the risk factors of coronary heart disease or atherosclerosis using Neural Network. Their model is able to predict whether the test patient is undergoing due to heart disease or not. In the reasecrh of R. Subramanian et al. [10] predicted heart disease using 120 layers' deep neural network. Their model performance is optimal; however, model-training time is high and require large datasets. Medhekar et al. [11] applied machine learning technique called Naïve base to predict the heart disease. They train their model with five labels. The classification accuracy of the model is 89.58%.

Several studies on the heart disease has been carried out by using machine learning techniques. But the performance of these model is not optimal. In the present study based on performance analysis of several machine learning techniques heart disease prediction system has been developed. In this regards, first the well-known machine learning algorithm Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes and Random Forest applied on the dataset for the prediction of heart disease. The performance of these models are evaluated using 3-fold cross validation. We notice that the highest average accuracy of 87.78% is obtained by the Naïve Bayes. Further, we have applied genetic algorithm on the dataset to optimize the features. After, optimization the highest average accuracy of 96% is achieved by the naïve Base.

II. METHODOLOY

In the proposed work we have applied four machine learning techniques Naïve base, Random forest, SVM and KNN to predict heart disease. The performance of these models are acceptable. However, the real time heart disease prediction needs more robust and efficient system. Therefore, we have applied feature optimization technique. After applying Genetic Algorithm (GA) the performance of the

2021 5th International Conference on Information Systems and Computer Networks (ISCON) | 978-1-6654-0341-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISCON52037.2021.9702410

model has been increased. The details of the proposed method are shown in the Figure 1.

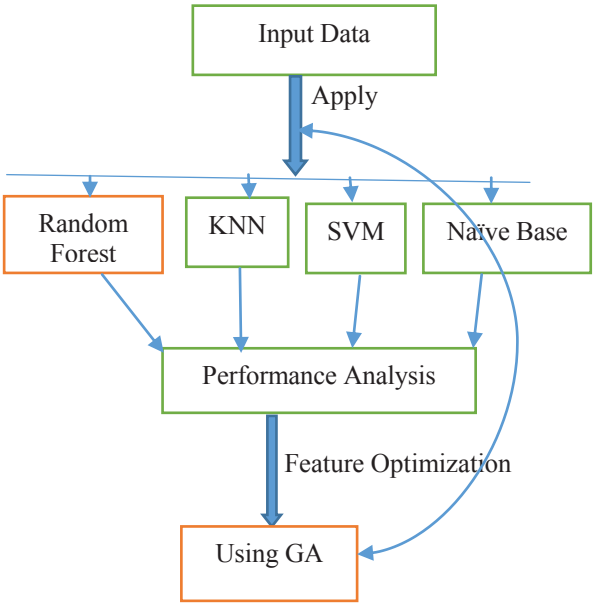


Fig. 1. The proposed method system flow diagram.

Naïve Bayes: This classifier processes each attribute's prospect in a class. The Bayes theorem is as follows: Let's consider $Y = \{y_1, y_2, y_3 \ldots\ldots\ldots y_n\}$ be a set of 'n' attributes. In Bayesian, Y is recognized as proof and H is some hypothesis means, the data of Y refers to particular class. We have to discover $P = \frac{H}{Y}$, the probability that hypothesis H holds provided evidence i.e. data sample Y. According to Bayes theorem, the $P = \frac{H}{Y}$ is represented as-

$$P(H / Y) = P(Y/ H)\, P(H) / P(Y) \quad (1)$$

Support Vector Machine (SVM): SVM is one of the classification methods used to identify patterns and data in a regression and group analysis. It separate data by attaining the most suitable hyperplane that departs all data points of one class from a different class [12]. SVM also uses a technique which is known as Kernel Trick to convert the data. Based on the conversion of data it obtains an optimal splitting line among the likely outputs. The boundary can be as easy as a narrow margin for binary classes, to a further difficult splitting that involves various classes [13].

Let's consider a two-dimensional space. To separate two-dimensional data a linear equation of line can is written, in which data points lying on either sides representing the respective classes.

The equation of the line is

$$x=ay+bx \quad (2)$$

Considering x and y as features, y_1, y_2…. y_n, it can be re-written as:

$$ay_1-y_2+b=0 \quad (3)$$

If we define $y = (y_1, y_2)$ and $w = (a, -1)$, we get:

$$w \cdot y+b=0 \quad (4)$$

KNN: K Nearest Neighbor algorithm befalls under the Managed Learning division and is used for sorting and regression. It is a versatile algorithm also used for attributing missing values and resampling datasets. As the title suggests it analyzes K Nearest Neighbors to prophesy the class or connected value for the original Data point.

The distance equation is-

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (5)$$

Random Forest: Random forest is a managed machine learning algorithm that can be used for solving both analyzing and regression problems. However, often it is favored for classification. It is named a random forest because it connects multiple decision trees to make a forest and feeds random features to them from the given dataset. Rather than depending on a single decision tree, the random forest takes prophecy from all the trees and picks the most favorable result through the polling process [14].

Let D be a training set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ (6)

Let p = $p_1(x), p_2(x), \ldots, p_k(x)$, where, p=set of weak classifiers

If each p_k works as a decision tree and is defined as

$$\Theta = (\theta_{k1}, \theta_{k2}, \ldots, \theta_{kp}) \quad (7)$$

Each decision tree k leads to a classifier

$$p_k(X) = p(X|\theta_k) \quad (8)$$

Final classification f(x) = Majority of $p_k(X)$ (9)

III. RESULT AND DISCUSSION

A. Dataset

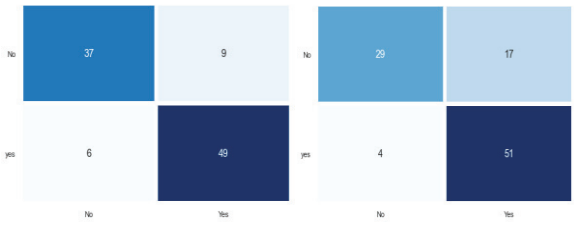
This study utilizes experimental data of 303 candidates gathered from UCI Machine Learning Repository [15]. The dataset is open access data repository and it can be used for the analysis of the heart disease.

In the presented study we have performed analysis of each model before and after optimization so that best model can be decided based on the performance measures. Since the dataset is small a 3-fold cross validation has been applied to avoid bias performance of the models.

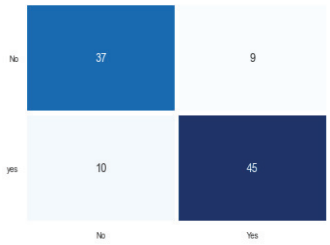
B. The confusion matrices of each model before optimization

1) KNN

The confusion matrices of KNN model is shown in the Figure 2 .



(a) (b)

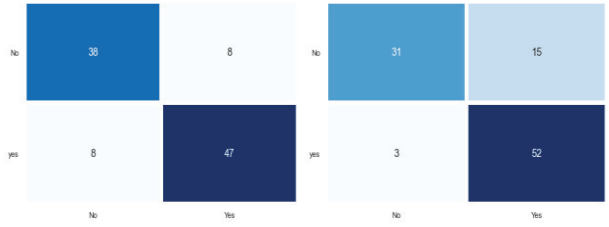


(c)

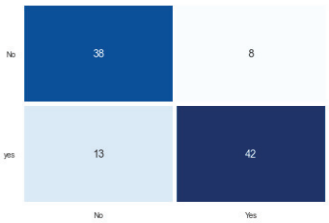
Fig. 2. CM (Confusion matrix) 2(a), 2(b) and 2(c) of KNN model

2) Random Forest

The confusion matrices of Random forest is shown in the Figure3.



(a) (b)

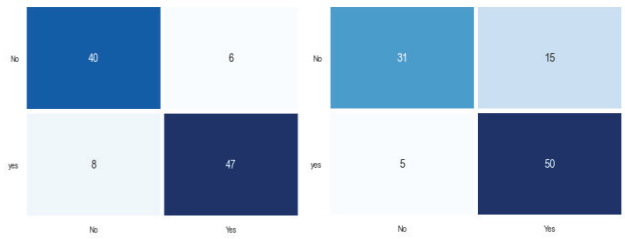


(c)

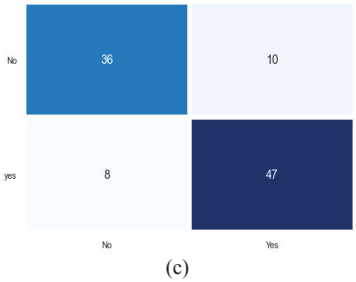
Fig. 3. CM (Confusion matrix) 3(a), 3(b) and 3(c) of Random Forest model before optimization

3) Naïve Bayes

The confusion matrices of Naïve base model are shown in the Figure4.



(a) (b)

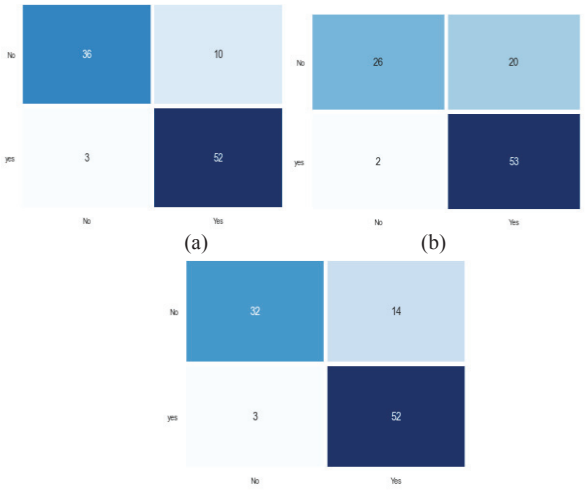


(c)

Fig. 4. CM (Confusion matrix) 4(a), 4(b) and 4(c) of Naïve Base model before optimization

4) SVM

The confusion matrices of SVM model are shown in Figure5.



(a) (b)

(c)

Fig. 5. CM (Confusion matrix) 5(a), 5(b) and 5(c) of SVM model before optimization

TABLE I. PERFORMANCE OF MODELS USING 3-FOLD CROSS VALIDATION BEFORE OPTIMIZATION ALL MEASURES ARE IN %

Method	Avg. Precision	Avg. Recall	Avg. F1 Score	Avg. Accuracy
1. KNN	86.3	85.3	85.3	85.8
2. RFC	87.3	87.3	87.3	86.8
3. NB	85.3	88.3	86.3	87.9
4. SVM	82.3	80.3	81.3	81.3

From the Table 1, we can see that performance of Naïve base is highest for all the measures.

C. Feature Optimization using genetic algorithm

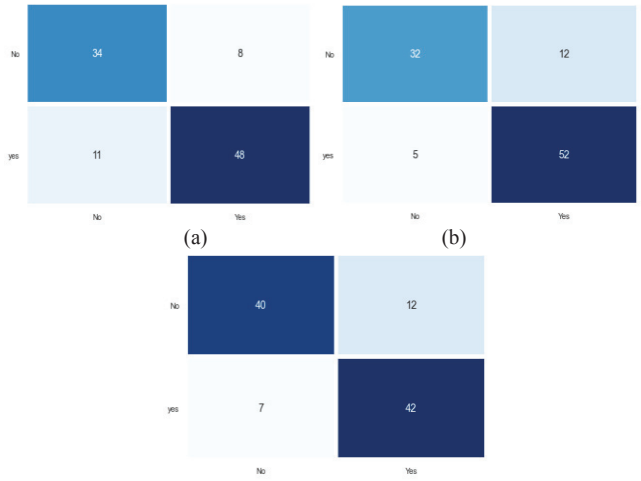
The genetic algorithm, created by John Holland and De Jong. It is an adaptive heuristic search algorithm, it is based on Charles Darwin's theory of natural evolution and inspired by genetics and natural selection. In the natural selection fittest entities are selected from the population. The fitness of the child is dependent on the fitness of the parent. Since descendants inherit the features of the parents and included in for the next generation. This process is repeated until a population of fittest is found. Optimization techniques are widely used to improve the performance of machine learning algorithms [16].

In the proposed study we have set our population size to 303. Which contains our all candidate of the dataset. Subsequently, a fitness function is defined which take candidate solution to input and yield an output which is matched to the problem for fitness consideration. Further, in the selection step we can select fit chromosomes. Then crossover step is reproduction of genes are performed. Finally, mutation is performed in which random tweak in the chromosome, which also promotes the idea of diversity in the population. In our research on heart disease prediction analysis we used this optimization algorithm in which mutation rate is 0.10, number of parents(n_parents) =100, features(n_feat)=30, generation(n_gen)=15.

The confusion matrices of each model after features optimization

1) KNN

The confusion matrices of KNN is shown in Figure6



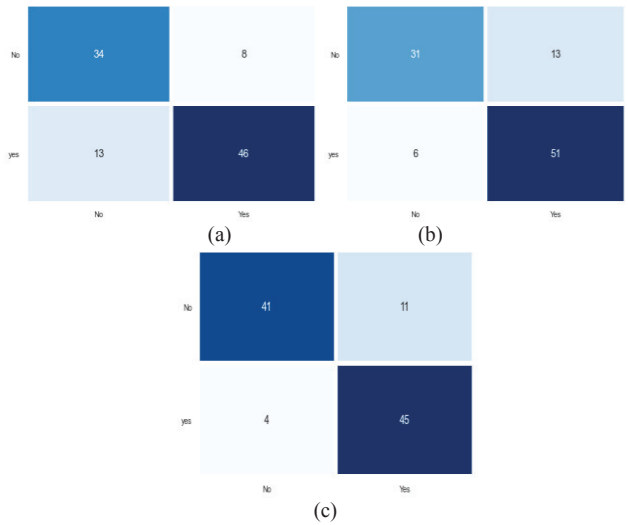
(a) (b)

(c)

Fig. 6. CM (Confusion matrix) 6(a), 6(b) and 6(c) of KNN model after optimization

2) *Random Forest*

The confusion matrices of Random forest are shown in Figure 7



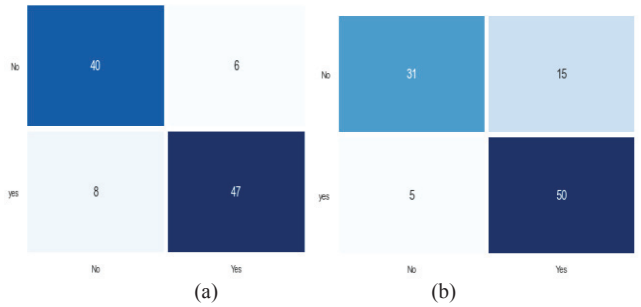
(a) (b)

(c)

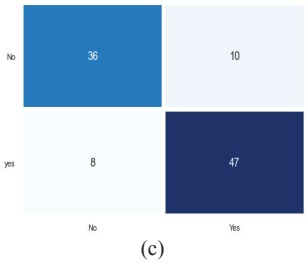
Fig. 7. CM (Confusion matrix) 7(a), 7(b) and 7(c) of Random Forest model after optimization

3) *Naïve Bayes*

The confusion matrices of Naïve Base is shown in the figure 8



(a) (b)

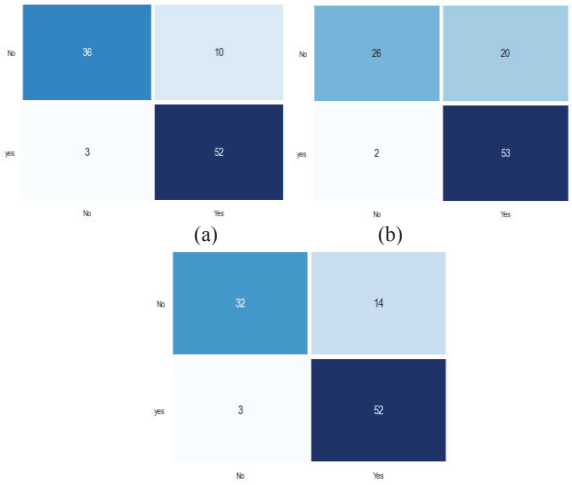


(c)

Fig. 8. CM 8(a), 8(b) and 8(c) of Naive Base model

4) *SVM*

The confusion matrices of SVM is shown in figure 9



(a) (b)

(c)

Fig. 9. CM (Confusion matrix) 9(a), 9(b) and 9(c) of SVM

TABLE II. PERFORMANCE OF MODELS USING 3-FOLD CROSS VALIDATION AFTER FEATURE OPTIMIZATION, ALL MEASURES ARE IN %

Method	Avg. Precision	Avg. Recall	Avg. F1 Score	Avg. Accuracy
1. KNN	85.3	85.3	85.3	81.8
2. RFC	88.3	87.3	88.3	86.8
3. NB	95.3	96.3	96.3	96.0
4. SVM	85.3	82.3	83.3	83.4

After applying genetic optimization on different machine learning algorithms we got 96% accuracy with Naive Bayes algorithm which is 87% before applying the genetic optimization technique. Through an analysis of the confusion matrices, we recognize that how the classification algorithms emphasized the balanced training dataset. Poor performance is also observed in the identification of prediction of diabetes. The algorithm which gave us the worst performance is SVM 81.31%. All classification algorithms have yielded positive results but have the ability to progress further.

TABLE III. COMPARISON WITH AVAILABLE METHODS

Method	[1]	[2]	[3]	[4]	Proposed models
Random forest	93.40%	88.7%	98,92%	78.53%	86%
SVM	76.57%	91.11%	96%	70.59%	83.4%
Naïve Bayes	78.88%	82.7%	_	_	96%
KNN	_	85.55%	_	_	81.8%

IV. CONCLUSION

In this reasecrh an automated system for heart disease prediction has been developed using machine learning and feature optimization technique to assist a doctor. The UCI dataset contain 14 attributes. These attributes have been used to train and classify SVM, KNN, Naïve Bayes and Random forest. Among these models Naïve base achieved highest accuracy of 87.9% using 3-fold cross validation. The achieved accuracy is satisfactory but real time heart disease prediction system should be more efficient. Therefore, a feature optimization technique GA has been applied. For the GA mutation rate is 0.10, number of parents(n_parents) =100, features(n_feat)=30,generation(n_gen)=15. After optimization of features again these models are trained. A 3-fold cross validation is applied to the dataset so that bias free performance can be measured. The Naïve base model achieved accuracy of 96%. This result is quite satisfactory for the heart disease prediction. In the future reasecrh other feature optimization technique can be applied to further improve the performance of machine learning algorithms.

REFERENCES:

[1] Zriqat IA, Altamimi AM, Azzeh M. A comparative study for predicting heart diseases using data mining classification methods. arXiv preprint arXiv:1704.02799. 2017 Apr 10.

[2] Muhammad Y, Tahir M, Hayat M, Chong KT. Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific reports. 2020 Nov 12;10(1):1-7.

[3] Alarsan FI, Younes M. Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. Journal of Big Data. 2019 Dec;6(1):1-5.

[4] Saxena K, Sharma R. Efficient heart disease prediction system. Procedia Computer Science. 2016 Jan 1;85:962-9.

[5] Pattekari SA, Parveen A. Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences. 2012 Jun 12;3(3):290-4.

[6] Sowmiya C, Sumithra DP. A Comparative Study of heart disease prediction using Data Mining Techniques. International Journal of Scientific & Engineering Research. 2016 Dec;7(12).

[7] Venkatalakshmi B, Shivsankar MV. Heart disease diagnosis using predictive data mining. International Journal of Innovative Research in Science, Engineering and Technology. 2014 Mar;3(3):1873-7.

[8] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. Heart disease prediction using machine learning algorithms. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012072). IOP Publishing(2021).

[9] Chaurasia V, Pal S. Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology. 2013;1:208-17.

[10] Chitra R, Seenivasagam V. Review of heart disease prediction system using data mining and hybrid intelligent techniques. ICTACT journal on soft computing. 2013 Jul;3(04):605-09.

[11] Medhekar DS, Bote MP, Deshmukh SD. Heart disease prediction system using naive Bayes. Int. J. Enhanced Res. Sci. Technol. Eng. 2013 Mar;2(3).

[12] Learning M. Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology. 2017;10(7):2137-59.

[13] Soni J, Ansari U, Sharma D, Soni S. Intelligent and effective heart disease prediction system using weighted associative classifiers. International Journal on Computer Science and Engineering. 2011 Jun;3(6):2385-92.

[14] Patel J., Upadhyay T., Patel S,, Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.

[15] https://archive.ics.uci.edu/ml/datasets/heart+disease

[16] Kaur H, Wasan SK. Empirical study on applications of data mining techniques in healthcare. Journal of Computer science. 2006 Feb 1;2(2):194-200.