International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Prediction of Cardiovascular Disease Using Machine Learning Algorithms with SDNKL

Kuppam Charitha Sri[1], Kanchi Prathyusha[2], Kanteti Sharon Pushpa[3], Kanathala Manasa[4], SK. Khaja Mohiddin[5]

[1,2,3,4] *BTECH, Department of Computer Science Engineering, VVIT, Andhra Pradesh, India.*
[5]*Associate Professor, Dept. of Computer Science Engineering, VVIT, Andhra Pradesh, India.*

Abstract

In human life, healthcare is an unavoidable and important task to be done. Cardiovascular Diseases are a group of diseases that affects heart and blood vessels. The earlier methods of estimating the uncertainty levels of cardiovascular diseases helped in taking decisions to reduce the risk in high-risk patients. The health care business collects a lot of medical data; the machine learning algorithms are utilized in the identification of the patterns for forecasting and controlling for analysis and medication. In the proposed research, techniques like the missing data removal and attributes classification are used in data pre-processing for better prediction process and taking decisions at different stages.

This project proposes a prediction model to predict whether a person has a heart disease or not and to provide awareness or diagnosis on the risk to the patient. This is achieved by comparing the accuracies of different algorithms to the separate results of SVM, KNN, Decision Tree, Naive Bayes classifier and Logistic Regression and uses the algorithm with high accuracy for prediction. Our goal is to enhance the performance of the model by removing unnecessary and insignificant attributes from the dataset and only collecting those that are most informative and useful for the classification task.

Key Words: **Cardiovascular Diseases, Machine Learning Algorithms, Prediction Model, Removal of Irrelevant Attributes**

1. INTRODUCTION

Healthcare means the maintenance or advancement of health through the prevention and diagnosis of people. Nowadays, healthcare is increasing day by day due to lifestyle, hereditary [1]. It generates a lot of data every second. So, data analytics are producing beneficial information from the collected medical data without wasting it. Cardiovascular disease has become the deadliest enemy. A person with cardiovascular disease cannot be cured simply. So, diagnosing patients at the correct time is the toughest work in the medical industry and needs to be diagnosed at initial stages to reduce the risk on the patient in the future. Every human body possesses different numbers for blood pressure, cholesterol, and pulse rate. But the normal values would be, blood pressure is 120/80, cholesterol is 200 mg/dl and pulse rate is 72.

Generally, cardiovascular diseases are diagnosed by cardiologists. Diagnosing this kind of disease is a difficult and important task to be executed correctly and efficiently. If this diagnosis was not executed properly then it may lead to undesired outputs. Therefore, an automatic diagnosing system is beneficial. So combining these machine learning algorithms with medical data sources is useful. This paper suggests different machine learning methods that are useful for forecasting the uncertainty levels of cardiovascular disease for a person depending on the collected attributes.

Machine Learning makes a prediction model that predicts depending on the given input. We all know that a supervised learning technique is used to train the model with the set of known input and output to get correct predictions as the output for new data.

1.1 Types of Cardiovascular Diseases

Cardiovascular disease involves conditions that affect operating your heart like narrowing of the arteries, abnormal heart rhythms, heart failure and cardiopathy.

The major common symptoms that are observed because of a heart attack are as follows:

1.2 Chest pain

It causes discomfort in the chest including a dull ache, crushing or burning feeling, sharp stabbing pain and pain that radiates neck or shoulder. The blood flow reduction by heart blood vessels result in the death of heart muscular tissue cells.

1.3 Nausea, Indigestion, Heartburn and Stomach pain

The above are some of the common symptoms of a heart failure. These symptoms were most common to ladies than men.

1.4 Pain in the arms

The pain frequently starts in the chest area and then moves towards the left arm.

1.5 Feeling Dizzy and Light Headed

It generally leads to loss of consciousness.

1.6 Fatigue

It ends up with the feeling of overtiredness with less energy and a strong desire to sleep.

1.7 Sweating

Sweating over unusual particularly if you aren't exercising or being active might be an early be-careful call of regular heart issues.

2. LITERATURE SURVEY

In the presented paper [1], this paper is a combination of the correlative application and detailed examination of different ML Algorithms in R software which results in an immediate mechanism for the user to use the machine learning algorithms in R software for estimating the cardiovascular diseases. Future enhancement comprises the work of different groups of methods to analyze these algorithms for better performance with more framework settings for these algorithms.

In the given paper [2], by examining the test results, it is concluded that J48 tree technique to be a classifier to predict heart problems because it contains more accuracy and less time to build and also observed that applying reduced error pruning to J48 results in higher performance. In summary, as identified through the literature review, we believe only a marginal success is attained in the design of a predictive model for patients having heart issues and the necessity of combinational and more complex models to increase the accuracy of analyzing the early onset of heart issues.

In the published paper [3], prediction of heart problems with distinct Decision Tree methods using classification. Heart disease is a mortal disease by its nature. This disease is a threat to life such as heart issues and may cause death. Data mining plays an important role in medical field and relevant actions are taken for prediction of disease. Many Classification Algorithms used for Disease Prediction, Decision Tree is taken because of its simplicity and accuracy.

In the taken paper [4], ML methods and Data Mining methods are selected for prediction of Heart Disease and diagnosing it. The disadvantage mainly depends on the applications related to classification techniques for prediction of heart disease, apart from this taking many data cleaning and pruning techniques that make a dataset suitable for mining. By analyzing the correct classification techniques will lead to the implementation of prediction systems that give more accuracy.

In the published paper [5], cardiovascular problems Prediction using ML Algorithms, Predicting heart problems uses ML Algorithm provides prediction results for users. I this method Random Forest Algorithm was selected for its efficiency and accuracy and to find out prediction of heart problem percentage by knowing the correlation details between heart disease and other diseases. For the best performance and to increase accuracy new algorithms are used.

In the following paper [6], cardiovascular problems predicting using ML Algorithms, We introduced a heart disease prediction system with different classifier techniques. The techniques are Naive Bayes and decision tree classification methods. We selected Decision Tree classifier for its performance, accuracy is more compared to others. Naive Bayes accuracy is more in some cases and small in other cases.

3. DATASETS AND DESCRIPTION

3.1 DATA SOURCE

A sufficient people data is collected and maintained in healthcare databases. Cardiovascular disease refers to the conditions which involved in narrow or blocked blood vessels which lead to heart attack or stroke. Other than the above conditions which affect heart muscle, valves or rhythm are also considered to get the disease. Records will get from the Cleveland, Hungarian, Switzerland, Long Beach VA database (UCI machine Learning Repository). Health care datasets includes medical data, various measurements, and specific patterns of population related to the disease. The records are classified into training dataset and test dataset. Some of 920 records along with 76 attributes related to the medical were obtained. The following table (Table 2) shows the list of attributes on which the system is working.

3.2 ANALYSIS OF DATA

It is defined as the process of cleaning, transforming, filling missing values and modeling the data to give us helpful information for healthcare decision making. The purpose of this is to pre-process the data and to get useful information by data and taking the decisions based upon the data analysis.

3.3 OPERATING ENVIRONMENT

The python provides Matplotlib, an amazing visualization library in that language for 2D plots of arrays. This is a multi-platform data visualization library which was on NumPy arrays and it is designed to figure with the broader SciPy stack. This library helps to understand trends, patterns, and perform statistical and graphical computations mainly for analyzing of data. With all those packages python provides users can make quick analysis of data used in prediction system for given application. PYTHON provides open source software which makes the best compatibility in UNIX and Windows for prediction results, PYTHON offers a better outcome compared to other languages. Heart diseases may come in many different ways, there is some typical arrangement to identify some key factors which decides whether somebody will, at some time later fall into the risk of those occurring diseases. The following are the basic characteristics that are to be checked to know the risk of occurrence of problems of diseases in future.

No	Name	Description
1	Age	Age in Years
2	Sex	1=male, 0=female
3	Cp	Chest pain type(1 = typical angina, 2 =atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
4	trestbps	Resting blood sugar(in mm Hg on admission to hospital)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl(1= true, 0=false)
7	restecg	Resting electrocardiographic results(0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricularhypertrophy)
8	thalach	Maximum heart rate
9	exang	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	Slope of the peak exercise ST segment (1=upsloping, 2=flat, 3= downsloping)
12	Ca	Number of major vessels colored by fluoroscopy
13	thal	3= normal, 6=fixed defect, 7= reversible defect
14	Num	Class(0=healthy, 1=have heart disease)

Table 1:Attributes Before Reduction

S.No	Attribute Name	Description
1	Age	Age of the person in years
2	Cp	Chest pain type [1-Typical Angina 2-Atypical Type Angina 3-Non-angina pain 4-Asymptomatic)
3	Trestbps	Resting Blood Pressure in mm Hg
4	Chol	Serum cholesterol in mg/dl
5	Restecg	Resting Electrocardiographic Results(0=normal, 1=having ST-T wave abnormality, 2=left ventricular hypertrophy)
6	Thalach	Maximum Heart Rate Achieved
7	Exang	Exercise Induced Angina
8	Old Peak	ST depression induced by exercise relative to rest
9	Slope	Slope of the Peak Exercise ST segment
10	Ca	Number of major vessels colored by fluoroscopy
11	Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect
13	Num	Class Attribute

Table 2: Attributes After Reduction

4. PROPOSED SYSTEM

The below figure (figure 1) shows functioning of the system is described step by step.

Step 1: The dataset contains the details of the patients.

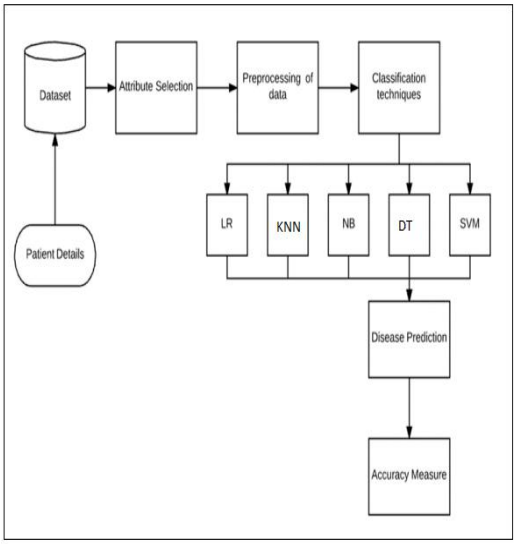

Figure 1: proposed system

Step 2: Attribute selection takes the attributes which are useful for the prediction of the heart disease.

Step 3: After identifying the data from the available resources, they are further selected for processing which includes data cleaning, removal of noise i.e. missing data.

Step 4: Different classification algorithms are performed on the preprocessed data for finding a chance of getting heart disease.

Step 5: It also finds the accuracy of the algorithms and compares the accuracy among all the algorithms.

5. ALGORITHMS USED

A. Logistic Regression

Logistic regression belongs is a statistics that is used by machine learning. It is mainly useful to classify binary problems. Logistic regression seems to be the best regression analysis to be used when the variables are dichotomous. If there exists greater than two discrete outcomes for a class then we use Multiclass logistic regression. The Representation of Logistic Function:

$$P = (y = 1|X) = \frac{1}{1 + e^{-wa}}$$

Here e is Euler's number and a is an input we give for the function.

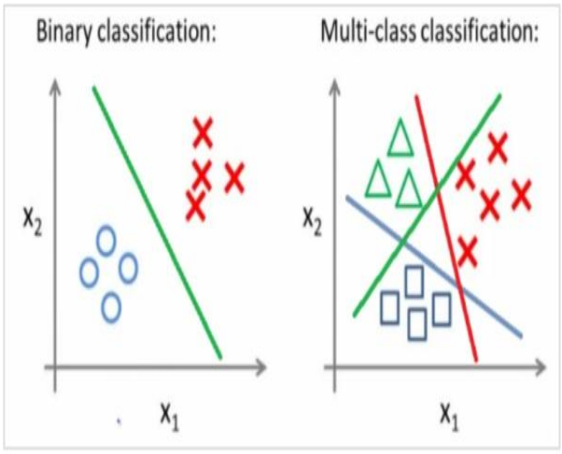

Figure 2: classification with Logistic Regression

Roc curve for the logistic regression:

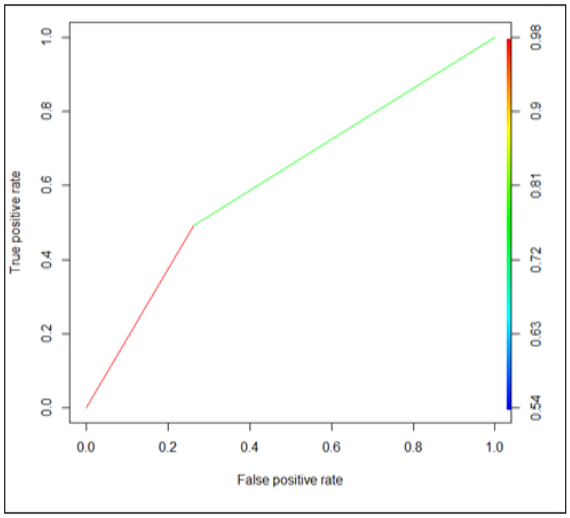

Figure 3: Roc curve

B. *Naive Bayes*

Naive Bayes classification model does the classification process based on probability. With Bayes theorem, given B has happened, the probability for occurring A can be found. Here, A is said to be hypothesis and B is the evidence. The assumption is that independent predictors are taken. That is if one specific feature has no affect on the other. Hence it is called as Naïve. The probability of A with respect to B is given as follows.

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

C. *K-Nearest Neighbors*

Implementation of KNN is simple and easy. KNN comes under the group of supervised algorithms.

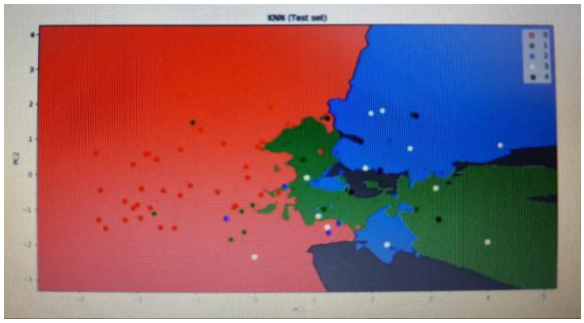

Figure 4: k-Nearest Neighbors

By using this, problems based on classification as well as regression can be solved. KNN algorithms make use of data and classify a new set of data points basing on similarity measures. Data points are classified by considering a maximum vote to its neighbors and the data is assigned for class having nearest neighbors. The accuracy of the algorithm might be increased by increasing the k-value that is increasing of the nearest neighbor's number. Based on the data the best k-value can be fixed.

D. *SVM*

It solves regression as well as classification problems. It comes beneath the class of supervised algorithms. Moreover it is often used for solving classification problems. In SVM, plotting of data is done as a single point in an n-dimensional sample space having each feature the definite value of a coordinate. We then perform classification with a coordinate that separates the two classes perfectly. For each attribute this method plots hyper plane as a coordinate in the dataset.

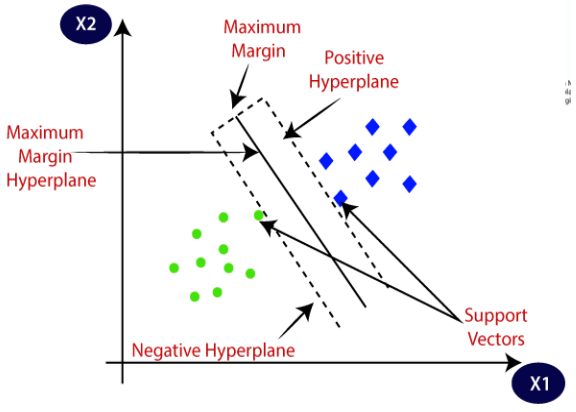

Figure 5: SVM

E. *Decision Tree*

Decision Tree comes under the category of supervised algorithm where continuous splitting of data is done based on particular parameter. This is understood clearly with two items called decision nodes where the

decision of splitting occurs and leaves which are the final outcomes of tree. Decision trees can be found in two types: classification trees and another type are regression trees.

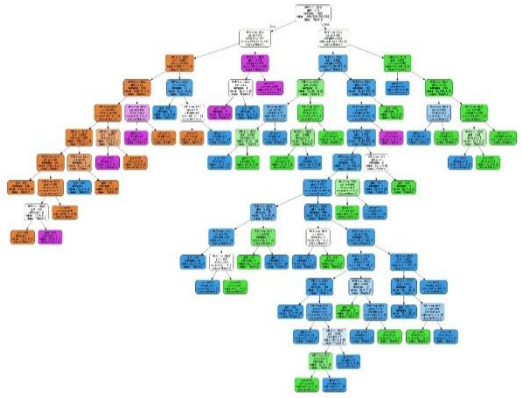

Figure 6: Decision tree

Classification trees are expressed as Yes or No type trees. Regression trees are expressed as continuous types. Iterative Dichotomiser3 (ID3) algorithm is chosen to be the best algorithm to implement Decision tree technique.

6. ACCURACY MODULE

By using above algorithms the accuracy is predicted. This module considers the highest accuracy resulted from all the above algorithms that predict the maximum cases of getting a cardiovascular disease. Here every algorithm generates a distinct accuracy for the attributes taken as they might be the reason for the cardiovascular disease. Accuracy for the model can be calculated with:

JACCARD INDEX: This is known as statistic that is helpful for finding the similarities between the sets. The similarity among the sample sets is emphasized by this measurement. This is formatically termed as the ratio of intersection size and union size of the sample sets. The representation of this Jaccard index:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

F1 SCORE: This is referred as balanced F-measure or F-score. The best score is at 1 and the worst score is at 0. The comparative contribution of recall and precision are same for F1 score. F1 score is given as:

F1 = 2 * (precision * recall) / (precision + recall)

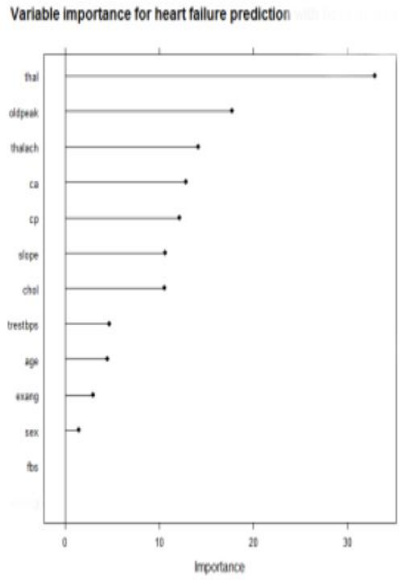

Figure 7: Attributes Importance

6. DATA VISUALIZATION

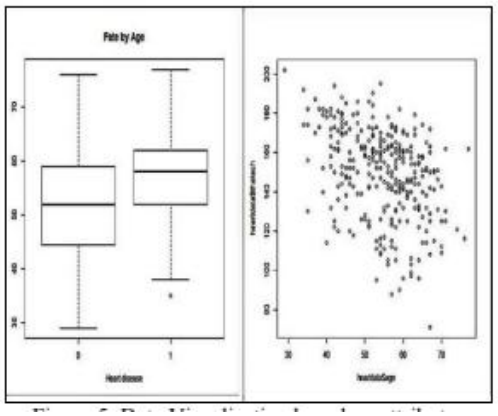

Figure 8: Data visualization based on attributes

The way of representing of data and information as graphs is termed as Data visualization. To understand outliers, trends and data patterns, the tools in data visualization provide elements like charts and so on for clear visualization.

7. RESULTS AND DISSCUSSION

The outputs and the accuracies generated by each algorithm are reviewed and the outputs are displayed. In this model, the algorithm which has the highest accuracy gives accurate result. The following tables show the accuracies generated by each algorithm as follows:

	Algorithm	Jaccard-score	F1-score
0	Decision Tree	0.533333	0.498633
1	SVM	0.500000	0.333333
2	Logistic Regression	0.533333	0.436752
3	KNN	0.466667	0.341463
4	Naive bayes	0.533333	0.520448

Table 3: Accuracies before attributes reduction

	Algorithm	Jaccard-score	F1-score
0	Decision Tree	0.533333	0.507576
1	SVM	0.500000	0.333333
2	Logistic Regression	0.533333	0.436752
3	KNN	0.466667	0.341463
4	Naive bayes	0.600000	0.547100

Table 4: Accuracies after attributes reduction

The output of this model would be a number from 0 to 4 which indicates the risk level of the patient. 0 indicates Normal, 1 indicates abnormal, 2 indicates low risk, 3 indicates risk and 4 indicates high risk.

```
from sklearn import model_selection
X = data[['age','cp','trestbps','chol','restecg','thalach','exang','oldpeak','slope','ca','tha
#X = np.array(data.drop(['class'], 1))
y = np.array(data['class'])
```

```
# initialize list of lists
data_test = [[63.0,4.0,124.0,197.0,0.0,136.0,1.0,0.0,2.0,0.0,3.0]]
```

```
#Naive bayes
gnb_pred = gnb.predict(df_test)
print(gnb_pred)
```

```
[0]
```

8. CONCLUSION

This paper contributes the application and analysis of different machine learning techniques in the R software for forecasting the cardiovascular diseases. This is non ethical study aims to use available machine learning algorithms in R software. Future work includes any different machine learning algorithms which can enhance the performance with even more parameter settings for these algorithms.

9. REFERENCES

[1]. Dinesh Kumar G, Arumugaraj K, Santosh Kumar, Mareeswari V "Prediction of cardiovascular Disease using machine Learning Algorithms", 2018.

[2]. Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel "Heart disease prediction using Machine learning and Data Mining Technique" Volume 7- Number1 Sept 2015 March 2016.

[3]. G. Parthiban, S.K.Srivasta "Applying Machine learning methods in Diagnosing Heart disease for Diabetic Patients" International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 3– No.7, August 2012

[4]. Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. International Journal of Engineering Research & General Science, Vol 2(6), pp 6-11, Oct 2014.

[5]. Igor Kononenko" Machine learning for medical diagnosis: history, state of art& perspective" Elsevier - Artificial intelligence in Medicine, Volume23, Aug 2001.

[6]. Gregory F. Cooper, Constantin F. Aliferis", Richard Ambrosino, John Aronisb, Bruce G. Buchanan, Richard Caruana', Michael J. Fine, Clark Glymour", Geoffrey Gordon", Barbara H. Hanusad, Janine E. Janoskyf, Christopher Meek", Tom Mitchell", Thomas Richardson", Peter Spirtes" An evaluation of machine learning methods for predicting pneumonia mortality"- Elsevier Feb 1997.

[7]. Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease comparing with different algorithms": Computing, Communication& Automation (ICCCA), 2015InternationalConference.

[8]. B.Dhomse Kanchan, M.Mahale Kishore "Study of Machine learning algorithms for special disease prediction using principal of component analysis" Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016 International Conference.

[9]. Matjaz Kuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, JureFettich" Analyzing and improving the diagnosis of ischemic heart disease with machine learning" Elsevier - Artificial intelligence in Medicine, Volume23, May 1999.

[10]. Geert Meyfroidt, Fabian Guiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"-Best Practice & Research Clinical Anesthesiology, Elsevier Volume 23 (1) – Mar 1, 2009.

[11]. Gregory F.Cooper, Constantin F.Aliferis, Richard Ambrosino" An evaluation of Machine learning methods for predicting pneumonia mortality"-Elsevier, 1997.

[12]. Sanjay Kumar Sen" Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms"- International Journal of Engineering And Computer Science ISSN: 2319-7242Volume6Issue 6 June 2017.

[13]. Abhishek Taneja" Heart Disease Prediction System Using Data Mining Techniques"-Vol.6, No(4) December 2013.

[14]. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee" Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review"- Advances in Computational Sciences and Technology ISSN 0973-6107, Volume10, Number7 (2017).

[15]. Beant Kaur, Williamjeet Singh" Review on Hear Disease Prediction System using Data Mining Techniques"- International Journal on Recent and Innovation Trends in Computing and Communication Volume: 2 Issue: 10, October 2014.Transl. J. Magn. Japan, vol. 2, pp. 740- 741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[16]. Sonam Nikhar, A.M. Karandikar" Prediction of Heart Disease Using Machine Learning Algorithms"- Vol-2 Issue-6, June 2016.

[17]. S. U. Ghumbre and A. A. Ghatol, "Heart Disease Diagnosis Using Machine Learning Algorithm," Advances in Intelligent and Soft Computing Proceedings of the International Conference on Information Systems Design and Intelligent Applications Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India 7 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, pp. 217–225, 2012.

[18]. Meherwar Fatima, Maruf Pasha" Survey of Machine Learning Algorithms for Disease Diagnostic"- Journal of Intelligent Learning System and applications, 2017.

[19]. Younus Ahmad Malla, Mohammad Ubaidullah Bokari" A Machine Learning Approach for Early Prediction of Breast Cancer"- International Journal of Engineering and Computer Science, Volume6, Issue5, May 2017.

[20]. B. D. C. N. Prasad, P. E. S. N. Krishna Prasad, and Y. Sagar, "A Comparative Study of Machine Learning Algorithms as Expert Systems in Medical Diagnosis (Asthma)," Advances in Computer Science and Information Technology Communications in Computer and Information Science, pp. 570–576, 2011. [20]. Heart Disease Forecasting System using K-Mean Clustering Algorithm with PSO and other Data Mining Methods Shilna S1, Navya EK2 ISSN(P): 2349-3968, ISSN (O): 2349-3976. Volume III, Issue IV, April-2016.