www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 6 June 2017, Page No. 21623-21631
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14

Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms

Sanjay Kumar Sen
Asst. Professor, Computer Science & Engg.
Orissa Engineering College, Bhubaneswar, Odisha – India.

Abstract

Prediction and diagnosing of heart disease become a challenging factor faced by doctors and hospitals both in India and abroad. In order to reduce the large scale of deaths from heart diseases, a quick and efficient detection technique is to be discovered. Data mining techniques and machine learning algorithms play a very important role in this area. The researchers accelerating their research works to develop a software with the help machine learning algorithm which can help doctors to take decision regarding both prediction and diagnosing of heart disease. The main objective of this research paper is predicting the heart disease of a patient using machine learning algorithms. Comparative study of the various performance of machine learning algorithms is done through graphical representation of the results.

Key words **- detection technique, data mining technique, machine learning technique, machine learning algorithm**.

Introduction

The highest mortality of both India and abroad is due to heart disease. So it is vital time to check this death toll by correctly identifying the disease in initial stage. The matter become a headache for all doctors both in India and abroad. Now a days doctors are adopting many scientific technologies and methodology for both identification and diagnosing not only common disease, but also many fatal diseases. The successful treatment is always attributed by right and accurate diagnosis. Doctors may sometimes fail to take accurate decisions while diagnosing the heart disease of a patient, therefore heart disease prediction systems which use machine learning algorithms assist in such cases to get accurate results.[1]

2.3. Heart Disease

The heart attack occurs when the arteries which supply oxygenated blood to heart does not function due to completely blocked or narrowed.

Various types of heart diseases are[2]
1) Coronary heart disease
2) Cardiomyopathy
3) Cardiovascular disease
4) Ischaemic heart disease
5) Heart failure
6) Hypertensive heart disease

7) Inflammatory heart disease
8) Valvular heart disease

Common risk factors of heart disease include
1) High blood pressure
2) Abnormal blood lipids
3) Use of tobacco
4) Obesity
5) Physical inactivity
6) Diabetes
7) Age
8) Gender
9) Family generation

Data mining is the process of automatically extracting knowledgeable information from huge amounts of data. It has become increasingly important as real life data enormously increasing [3]. Heart disease prediction system can assist medical professionals in predicting state of heart, based on the clinical data of patients fed into the system. There are many tools available which use prediction algorithms but they have some flaws. Most of the tools cannot handle big data. There are many hospitals and healthcare industries which collect huge amounts of patient data which becomes difficult to handle with currently existing systems[1]. Machine learning algorithm plays a vital role in analyzing and deriving hidden knowledge and information from these data sets. It improves accuracy and speed.

Machine Learning is extensively used in diagnosing several diseases like heart [4] and other crucial diseases. Among various algorithms in data modeling, decision tree is known as the most popular due to its simplicity and interpretability [5], [6].Now a days more efficient algorithms such as SVM and artificial neural networks have also become popular [7], [4], [8].

The rest of the paper is organized as follows: Section II provides data description ; Section III algorithm used ; Section IV provided performance comparision. Section V concludes the paper.

WEKA Tool We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. **[9]**WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data preprocessing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

We have applied following five commonly used classifiers for prediction on the basing on their performance. These classifiers are as follows:

Table-3.1: WEKA names of selected classifiers

Generic Name	**WEKA Name**
Bayesian Network	Naïve Bayes (NB)
Support Vector Machine	SMO

C4.5 Decision Tree	J48
K-Nearest Neighbour	1Bk

II.Dataset_Description

We performed computer simulation on one dataset. Dataset is a Heart dataset. The dataset is available in UCI Machine Learning Repository [10]. Dataset contains 303 samples and 14 input features as well as 1 output feature. The features describe financial, personal, and social feature of loan applicants. The output feature is the decision class which has value 1 for Good credit and 2 for Bad credit. The dataset-1 contains 700 instances shown as Good credit while 300 instances as bad credit. The dataset contains features expressed on nominal, ordinal, or interval scales. A list of all those features is given in Table 3.2.

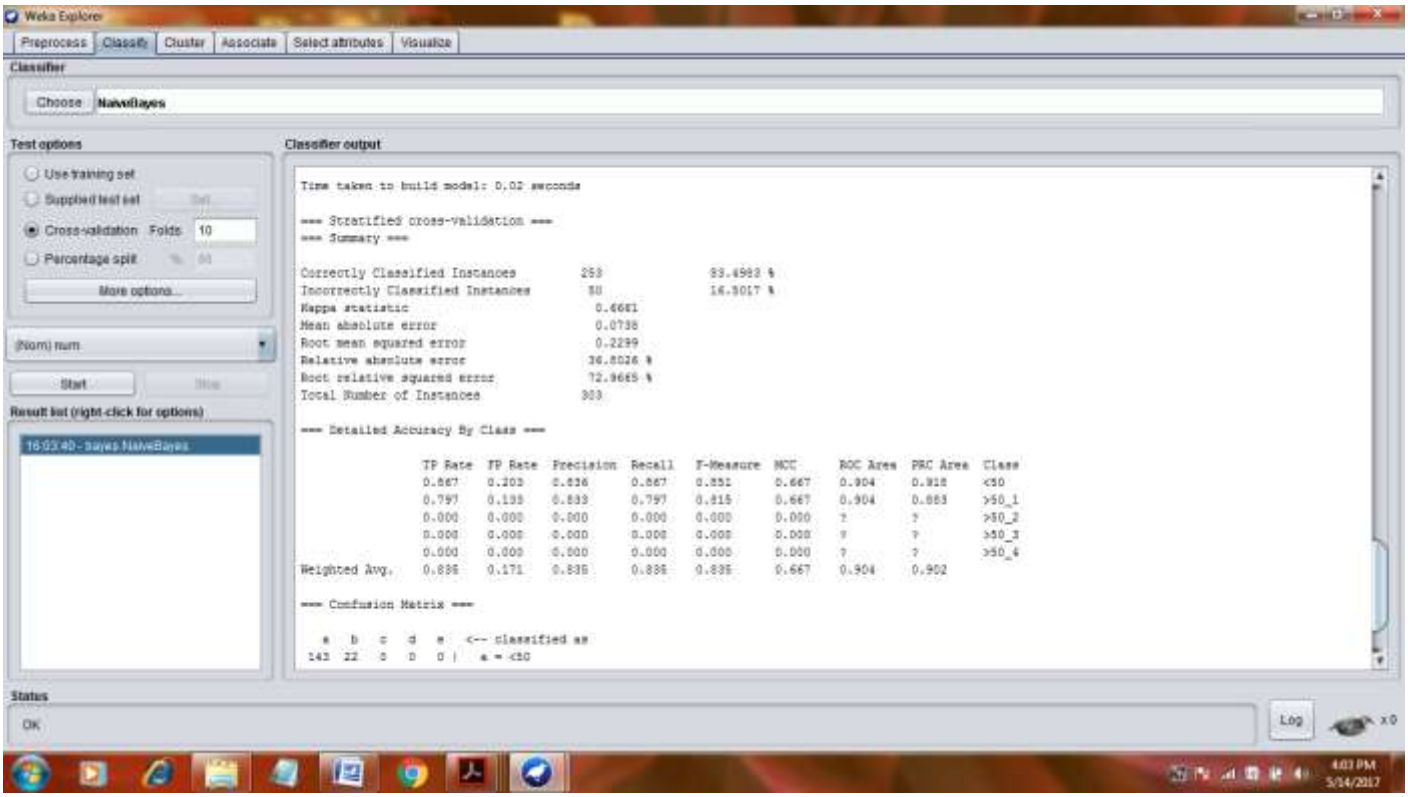
Table 3.2: Features in the Dataset-1

Feature No.	Feature Name
1	age
2	sex
3	cp
4	trestbps
5	choi
6	fbs
7	restesg
8	thalach
9	exang
10	oldpeak
11	slop
12	ca
13	thal
14	num

III. Algorithm used

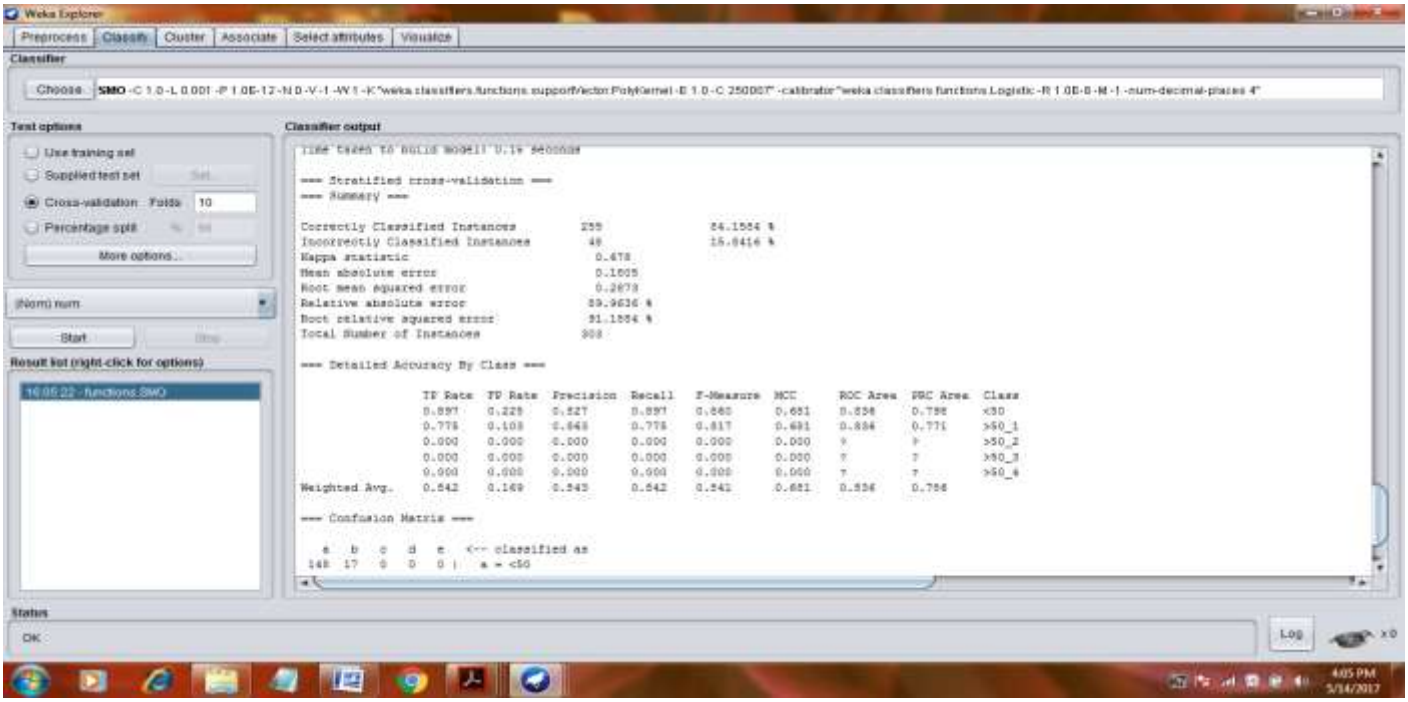
Naïve base classifier :- This classifier is a powerful probabilistic representation, and its use for classification has received considerable attention. This classifier learns from training data the conditional probability of each attribute Ai given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of A1…..An and then predicting the class with the highest posterior probability. The goal of classification is to correctly predict the value of a designated discrete class variable given a vector of predictors or attributes. In particular, the Naive Bayes classifier is a Bayesian network where the class has no parents and each attribute has the class as its sole parent. Although the naive Bayesian (NB) algorithm is simple, it is very effective in many real world datasets because it can give better predictive accuracy than well known well known methods like C4.5 and BP [11],[12] and is extremely efficient in that it learns in a linear fashion using ensemble mechanisms, such as bagging and boosting, to combine classifier predictions [13]. However, when attributes are redundant and not normally distributed, the predictive accuracy is reduced [14].

DOI: 10.18535/ijecs/v6i6.14

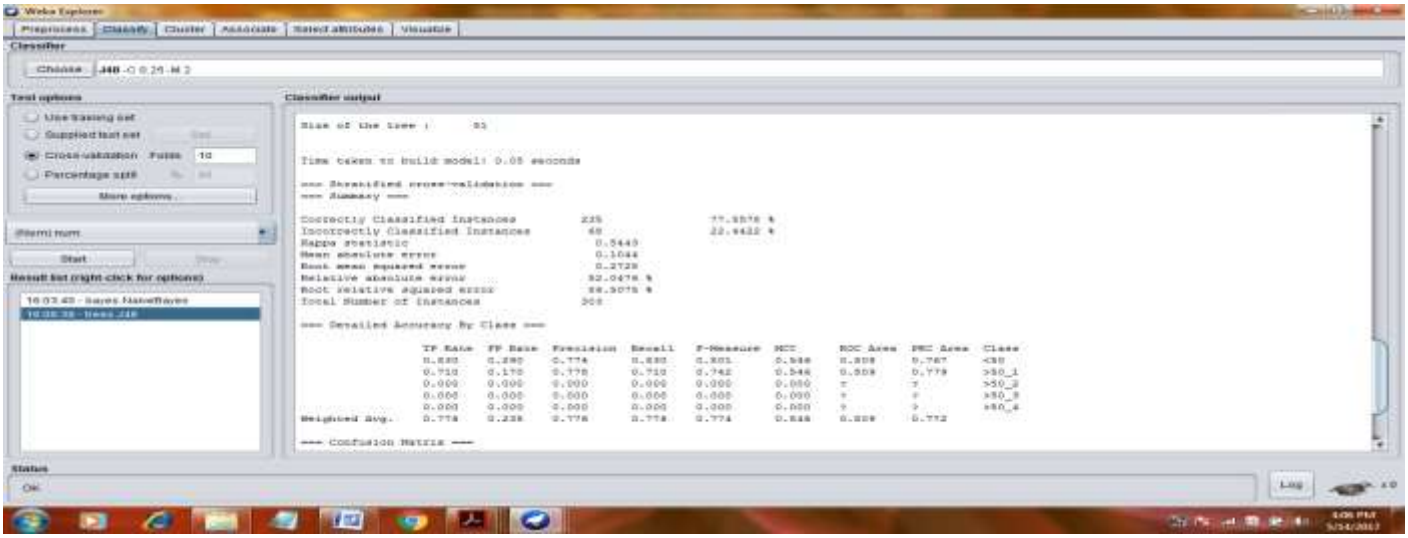


Support Vector Machine: Support vector machines exist in different forms, linear and non-linear. A support vector machine is a supervised classifier. What is usual in this context, two different datasets are involved with SVM, training and a test set. In the ideal situation the classes are linearly separable. In such situation a line can be found, which splits the two classes perfectly. However not only one line splits the dataset perfectly, but a whole bunch of lines do. From these lines the best is selected as the "separating line". The best line is found by maximizing the distance to the nearest points of both classes in the training set. The maximization of this distance can be converted to an equivalent minimization problem, which is easier to solve. The data points on the maximal margin lines are called the support vectors. Most often datasets are not nicely distributed such that the classes can be separated by a line or higher order function. Real datasets contain random errors or noise which creates a less clean dataset. Although it is possible to create a model that perfectly separates the data, it is not desirable, because such models are over-fitting on the training data. Overfitting is caused by incorporating the random errors or noise in the model. Therefore the model is not generic, and makes significantly more errors on other datasets. Creating simpler models keeps the model from over-fitting. The complexity of the model has to be balanced between fitting on the training data and being generic. This can be achieved by allowing models which can make errors. A SVM can make some errors to avoid over-fitting. It tries to minimize the number of errors that will be made. Support vector machines classifiers are applied in many applications. They are very popular in recent research. This popularity is due to the good overall empirical performance. Comparing the naive Bayes and the SVM classifier, the SVM has been applied the most[15].

DOI: 10.18535/ijecs/v6i6.14



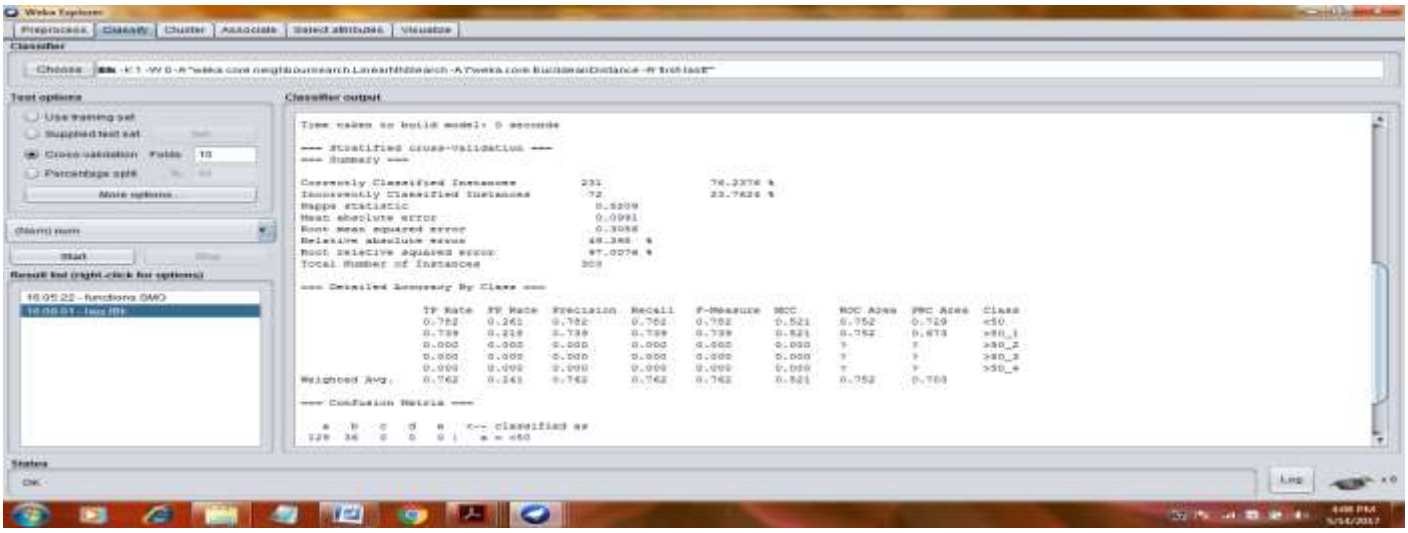
Decision Tree: A decision tree partitions the input space of a dataset into mutually exclusive regions, each of which is assigned a label, a value or an action to characterize its data points. The decision tree mechanism is transparent and we can follow a tree structure easily to see how the decision is made. A decision tree is a tree structure consisting of internal and external nodes connected by branches. An internal node is a decision making unit that evaluates a decision function to determine which child node to visit next. The external node, on the other hand, has no child nodes and is associated with a label or value that [3].



K-Nearest Neighbour: This classifier is considered as a statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest-neighbour classifier finds the closest training-point to the unknown

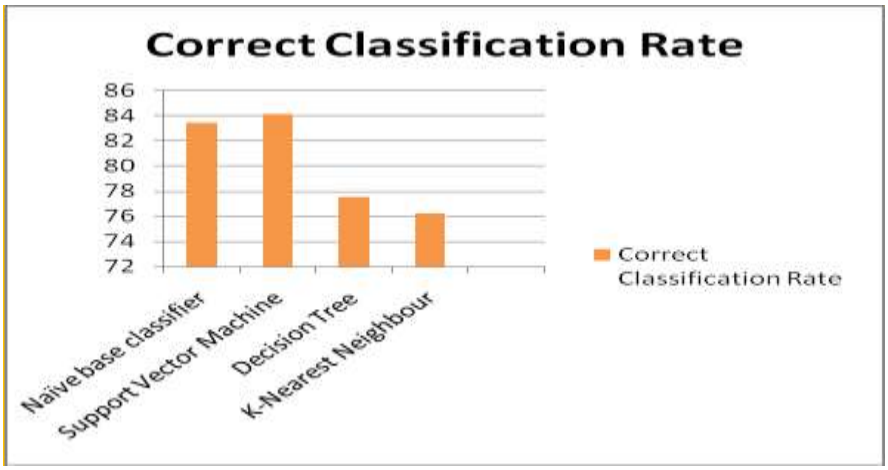
DOI: 10.18535/ijecs/v6i6.14

point and predicts the category of that training point according to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance[15].



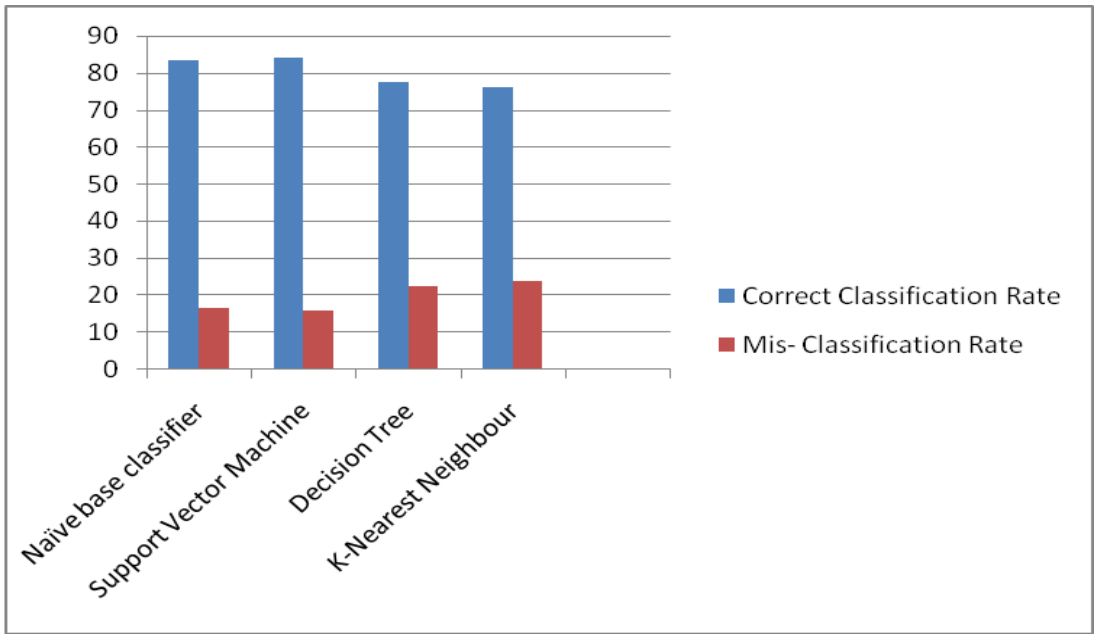
IV. PERFORMANCE COMPARISONS

Algorithm classification	Correct Classification Rate	Mis- Classification Rate
Naïve base classifier	83.4983	16.5017
Support Vector Machine	84.1584	15.8416
Decision Tree	77.5578	22.4422
K-Nearest Neighbour	76.2376	23.7624



DOI: 10.18535/ijecs/v6i6.14

Graph for Correct Classification VS. Misclassification rate



V. Conclusion

In this paper, we carried out an experiment to find the predictive performance of different classifiers. We select four popular classifiers considering their qualitative performance for the experiment. We also choose one dataset from heart available at UCI machine learning repository. Naïve base classifier is the best in performance. In order to compare the classification performance of four machine learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results in table 1, it can be concluded that Naïve base classifier is the best as compared to Support Vector Machine, Decision Tree and K-Nearest Neighbour.

After analysing the quantitative data generated from the computer simulations, Moreover their performance is closely competitive showing slight difference. So, more experiments on several other datasets need to be considered to draw a more general conclusion on the comparative performance of the classifiers.

DOI: 10.18535/ijecs/v6i6.14

Reference

[1] Prerana T H M1, Shivaprakash N C2 , Swetha N3 ”Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes,Introduction to PAC Algorithm, Comparison of Algorithms and HDPS” International Journal of Science and Engineering Volume 3, Number 2 – 2015 PP: 90-99 ©IJSE Available at www.ijse.org ISSN: 2347-2200

[2] B.L Deekshatulua Priti Chandra “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm “ M.Akhil jabbar* International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[3] Michael W.Berry et.al,”Lecture notes in data mining”,*World Scientific*(2006)

[4] S. Shilaskar and A. Ghatol, “Feature selection for medical diagnosis : Evaluation for cardiovascular diseases,” Expert Syst. Appl., vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[5] C.-L. Chang and C.-H. Chen, “Applying decision tree and neural network to increase quality of dermatologic diagnosis,” Expert Syst. Appl., vol. 36, no. 2, Part 2, pp. 4035–4041, Mar. 2009.

[6] A. T. Azar and S. M. El-Metwally, “Decision tree classifiers for automated medical diagnosis,” Neural Comput. Appl., vol. 23, no. 7–8, pp. 2387–2403, Dec. 2013. [10] Y. C. T. Bo Jin, “Support vector machines with genetic fuzzy feature transformation for biomedical data classification.,” Inf Sci, vol. 177, no. 2, pp. 476–489, 2007.

[7] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Syst. Appl., vol. 41, no. 9, pp. 4434–4463, Jul. 2014.

[8] A. E. Hassanien and T. Kim, “Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks,” J. Appl. Log., vol. 10, no. 4, pp. 277–284, Dec. 2012.

[9] Sanjay Kumar Sen 1, Dr. Sujata Dash 21Asst. Professor, Orissa Engineering College, Bhubaneswar, Odisha – India.

[10]. UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine-learningdatabases/ statlog/german/

[11] Domingos P and Pazzani M. “Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier”, *in Proceedings of the 13th Conference on Machine Learning*, Bari, Italy, pp105-112, 1996.

[12] Elkan C. “Naive Bayesian Learning, Technical Report CS97-557”, Department of Computer Science and Engineering, University of California, San Diego, USA, 1997.

[13] B.L Deekshatulua Priti Chandra “Reader, PG Dept. Of Computer Application North Orissa University, Baripada, Odisha – India. “Empirical Evaluation of Classifiers’ Performance Using Data Mining Algorithm” ***International Journal of C omputer Trends and Technology (IJCTT) – Volume 21 Number 3 – Mar 2015*** ISSN: 2231-2803 http://www.ijcttjournal.org Page 146

[14] Elkan C. Magical Thinking in Data Mining: Lessons From CoIL Challenge 2000, Department of Computer Science and Engineering, University of California, San Diego, USA, 2001.

[15] Witten I and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java*, Morgan Kauffman Publishers, California, USA, 1999.