iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4002

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

Analysis of Heart Disease Prediction System using Data Mining Techniques

Mrs. K. Pushpalatha[1], Dr. S. Gunasekaran[2], Mr. U. Rahulnath[3], Mr. M. Saravanan[4], Mr. B. Sriharri[5], Mr. S. Ajeeth Kumar[6]

[1]*Assistant Professor,* [2]*Professor,* [3, 4, 5, 6]*Department of CSE, Coimbatore Institute of Engineering and Technology , Coimbatore, India*

Abstract: ***Heart Disease is considered as one of the predominant reason of death all through the arena. We are living in an "information age" and Terabytes of data are produced every day. The health care enterprise generates a massive amount of statistics day by day. However, maximum of it is not successfully used. Hence, it is crucial to have a frame work that may effectually recognize the prevalence of heart disease in thousands of samples instantly. Heart disease can be predicted using various factors which incorporates circle of relatives records, cholesterol, diabetes, exercise and so on. The aim of the paper is to discuss the recent research on prediction of heart diseases using data mining techniques. We have investigated and analyzed the heart disease prediction system by using data mining techniques like SVM, KNN, Random tree and K star algorithms. The experimental result shows that the SVM produces optimum performance than the other techniques.***

Keywords: Data mining, Heart Disease, Weka, SVM, Kstar, Random Tree

I. INTRODUCTION

Data mining is defined as a way of extracting necessary and hidden predictive data from the large amount of database [1]. Data mining classifiers are used to find wide application in medical advancement for diverse diagnosis of different data mining techniques are used for prediction and decision making for different kinds of disease like cancer, heart disease, diabetes etc [2]. There are many types of data mining classifier algorithms like SVM (Support Vector Machine), KNN (K Nearest Neighbours), Random forest, K star algorithm can be applied to predict the disease.

Now day's Heart disease is main reason for death in the world. Heart disease is leading cause of death in the world [3]. Some of the major reason for heart diseases are blood pressure, cholesterol, pulse rate. More common type of heart disease that makes the heart works abnormally are congenital heart disease, heart failure, hypertensive heart disease, cardiomyopathy, heart murmurs, rheumatic heart disease, pulmonary stenosis and coronary artery disease. Heart can be attacked by various diseases that leads the heart to not work properly. The World Health Organization (WHO) has estimated that 12 million deaths around the world occurs each year from heart disease. In 2008, 17.3 million people died from heart disease. "Death rate increases by 80% in the world due to heart disease" The World Health Organization (WHO) estimates that by 2030, 23.6 million people worldwide will die from heart disease [4].

When the prediction of heart disease is getting accurate, then the specific patient can take preventive measures so that heart disease as the number one killer in the world can be reduced. The aim of the paper is to study the various data mining techniques using different tools. The algorithms are selected based on the accuracy. The risk level is classified by using SVM, KNN, Random forest and K star algorithms. This paper helps to understand the methodologies in the recent literature for predicting the heart disease using data mining techniques.

The rest of the paper is organised as follows. Section 2 describes the related works in the literature. Section 3 describes various data mining techniques; Section 4 describes the system architecture; Section 5 describes methodologies used for analysis and lastly the results are discussed in section 6.

II. RELATED WORKS

A lot of research works has been carried out in the past to use the data mining techniques for accurate prediction in diagnosing the heart diseases. The factors like age, sex, chest pain, blood pressure, cholesterol, blood sugar, are examined for predicting the heart disease. Salha M. Alzahani and et. al. [5] proposed an experiment on early diagnosis and prediction of heart disease using various data mining techniques. They have concluded that how data mining techniques are useful for the early detection and prediction of heart diseases which may save the human from heart attacks.

Sarath Babu and et. al. [6] discussed the Heart disease diagnosis system using data mining techniques.They have used three different classification techniques such as K-means classification, Decision tree and MAFIA(A Maximal Frequent Item set) algorithm for early detection of heart disease and its diagnosis correctly on time and providing treatment with affordable cost.
M. Akhil jabbar et al. [7] presented an effective classification technique based genetic algorithm is used for heart disease prediction. The scope of using genetic algorithm to predict the disease from large dataset is that the actual size of data is to get best attribute set. There is certain limitation in the prediction of heart disease using data mining approach. By reducing the set of attributes we can make it less complex and better accurate. Eman AbuKhousa, Piers Campbell [8] have implemented the Naïve Bayes technique to build the predictive model. RaghunathNambiar, and AdhiraajSethi et al [9] evaluated the data mining techniques such as CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and Decision Table (DT). Different classifiers and research are conducted to find the best classifier for calculating the patients Diagnosis. Among the entire algorithm CART gives more accuracy.

III. DATAMINING

Data mining is simply explained as the "knowledge discovery in databases" process, or KDD. Data mining is the process of extracting hidden pattern in the huge amount of data sets. Data mining is used in various fields like telecommunication industry, Retail industry, biological data analysis, and Intrusion detection etc., Use of data mining techniques in the medical field causes high impact on diagnosing, predicting and understanding of healthcare data. The patient's informations are maintained in the database. With the help of the data mining techniques, cardiologist can get the valuable information from the datasets [10].

A. Classification

Classification is a supervised technique which assigns items in the collection to target category or classes. Classification is one popular technique under data mining techniques. The aim of classification technique is used to predict the target class for each case in the data. Classes and groups are classified from data or objects[11]. Classification is the process of learning a function that data objects to a subset of a given class set. This task take X as input component and takes Y as the outcome of X, i.e.

$$C(X) \in Y$$

Where, X represents feature vector, Y represents response taking values in the set C, C(X) defines the values in the set C.

B. Clustering

Clustering is the process of grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. It is the process of dividing or partitioning a set of data or objects into a set of meaningful sub-classes. Thus formed sub-groups are called clusters. Each and every near object is neighborhood object. For example, this process is used to obtain the group having the same risk factor.

C. Association

Association is a data mining function that discovers the probability of the co-occurrence of items in a collection. The relationships between the co-occurring items are expressed as association rules. This is used to describe the relationship of one item on other items in the same operation. Here, this is used to express the relationship of attributed used for analysis.

D. Feature Extraction

Feature selection is the process of reducing the inputs data for processing and for analyzing or finding the most meaningful inputs from the given sections. It can be also called as variable selection, attribute selection and variable subset selection in this process of selecting a subset of relevant features for use in model construction.

IV. ANALYSIS OF HEART DISEASE PREDICTION SYSTEM

The Prediction of the heart disease is carried out using the data mining techniques is illustrated in Figure 1 [12]. The system is evaluated with four different data mining techniques namely SVM, K-star, Random Tree and KNN. Here, we have collected datasets from the data repository. It contains 274 records of 15 attributes shown in (Table 1) where some records where some are for non-heart disease patients and some are for patients with heart disease.
Values are entered in the WEKA data mining software as training and testing. Then we choose our proposed classifier one by one and note down classifier output for each[13].

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

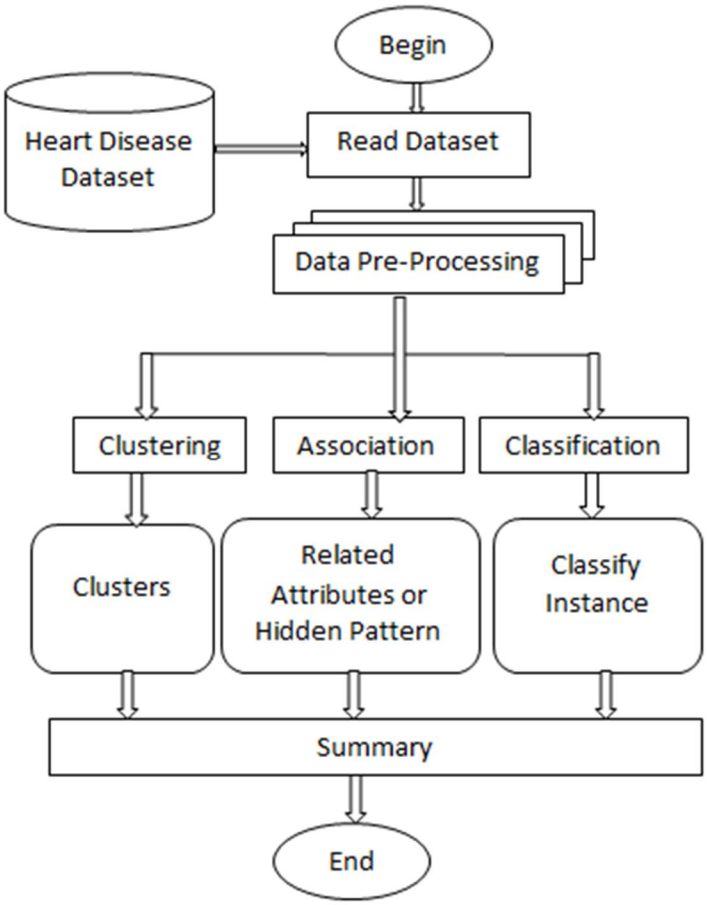


Fig. 1. System Architecture

V. IMPLEMENTATION

A. Support Vector Machine

A Support Vector Machine (SVM) [14] is a supervised learning classifier characterized by a separating hyperplane. The hyperplane is a line that partitions a plane in two sections where each class lay in either side. There are 2 sorts of SVM classifiers:

1) Linear SVM Classifier
2) Non-Linear SVM Classifier

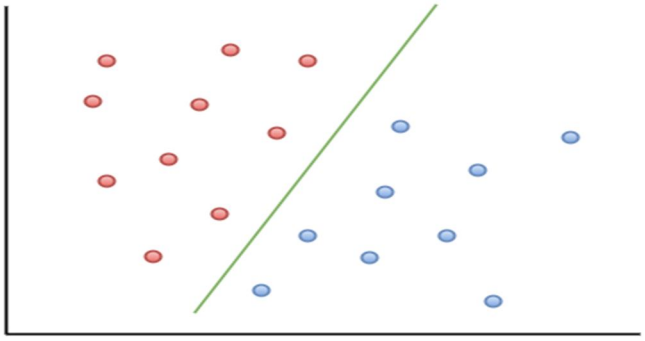


Fig. 2. Support Vectors Machine

SVMs are powerful when the quantity of highlights are very vast. Since the SVM algorithm works locally on numeric values. It handle a z-score standardization on numeric characteristics. Fig 2. Shows support vectors and hyperplane. The data lying inside

support vectors are used as the base data for machine model. SVM algorithm is not sensitive to other data. The objective of SVM is to find the best data boundary with the largest possible distance from all classes.

B. Kstar

Kstar [14] is a lazy learner , an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that it uses an entropy-based distance function. It uses information theory to calculate the distance between two instances with K star function denoted as

$$K*\left(\frac{C}{a}\right) = -log\ \mathrm{P}*\left(\frac{C}{a}\right)$$

Where P (c/a) denotes the summation probability of an example of being in category C.

C. Random Tree

Random Tree [16] is a supervised Classifier; it is a concert learning algorithm that generates many individual learners. It employs a gear idea to yield a random set of data for constructing a decision tree. In standard tree each node is divide using the best split among all variables. It can deal with both classification and regression problems. The classifier gets the input feature vector and classifies it with every tree in the forest. In regression, the classifier reply is the average of responses over all the trees. It can be described by the following equation.

$$Randome\ tree = \frac{1}{k}\sum_{k=1}^{k} Kth\ tree\ response$$

D. K Nearest Neighbour

K Nearest Neighbor[17] (also known as Collaborative Filtering or detail-based Learning) is a convenient data mining technique that grant you to use your past data instances, with known output values, to predict an unknown output value of a new data instances. It is a non-parametric method used for both classification and regression. The Euclidean distance method is used to calculate the distance between samples. For example if a is the first sample denoted by (a1, a2, a3….an) and b is the second sample which is denoted by (b1, b2…bn), the distance is calculated by using the following formula.

$$d = \sqrt{(a1-b1)^2 + (a2-b2)^2 + \cdots (an-bn)^2}$$

VI. DISCUSSONS AND RESULTS

In order to predict the probability of patients having heart disease and people who have no heart disease, a confusion matrix was created, where A indicate patients with heart disease, and B indicate patients with no heart disease

	A	B
A	(TP)	(FN)
B	(FP)	(TN)

TP - Represents the people with Heart Disease.
TN - Represents the people who doesn't have Heart Disease.
A confusion matrix contains report about real and predicted classifications fixed by an allotted system. The data in the matrix are check to know the performance of such systems.
The confusion matrix contains the following four entries:
TP (true positive): The number of records classified as true while they were actually true.
FP (false positive): The number of records classified as true while they were actually false.

FN (false negative): The number of records classified as false while they were actually true.
TN (true negative): The number of records classified as false while they were actually false

Table 1. Description of 15 used attributes

S.no	Parameters	Parameters description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Male or female	1=male 0=female
3.	Chestbps	Resting blood pressure	Continuous value in mmHg
4.	Cp	Chest pain type	1=typical tape 1 2=typical type angina 3=non-angina pain 4=asymptomatic
5.	Chol	Serum cholesterol	Continuous value in mm/dL
6.	Fbs	Fasting blood sugar	1>=120 mg/dL 0<=120 mg/dL
7.	Restecg	Resting electrographic results	0=normal 1=having ST-T wave abnormal 2=left ventricular hypertrophy
8.	Thalach	Maximum heart rate Achieved	Continuous value
9.	Old peak	ST depression induced by exercise relative to rest	Continuous value
10.	Exang	Exercise induced angina	0=no 1=yes
11.	Ca	Number of major vessels colored by fluoroscopy	0-3 value
12.	Slope	Slope of the peak exercise ST segment	1=unsloping 2=flat 3=downsloping
13.	Thal	Defect type	3=normal 6=fixed 7=reversible defect
14.	Obes	Obesity	1=yes 0=no
15.	num	Diagnosis of heart disease	0%<50% 1%>50%

To analyse the confusion matrix we consider the following performance parameter:
Accuracy (ACC) = Σ True positive +Σ True negative

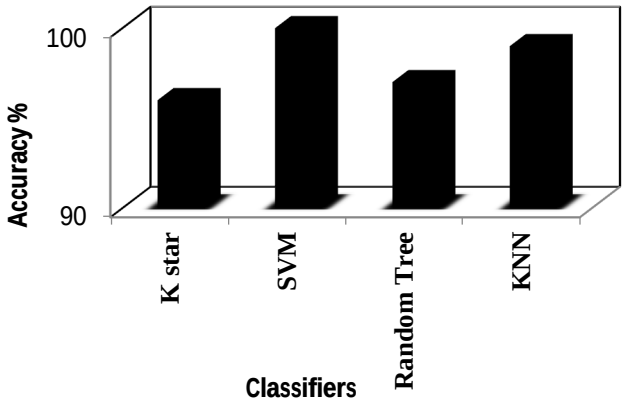
Σ Total population

Table. 2 shows the accuracy parameters of the inspected four classifiers namely (SVM, Kstar, Random Tree and K nearest neighbour).

Table. 2. Shows Accuracy parameters of Classifiers

Parameters	K-star	SVM	Random Tree	KNN
TP Rate	0.993	1.000	0.986	0.986
FP Rate	0.006	0.000	0.003	0.011
Precision	0.993	1.000	0.996	0.989
Recall	0.993	1.000	0.996	0.989
F-Measure	0.993	1.000	0.996	0.989
MCC	0.985	1.000	0.993	0.978
ROC Area	0.999	1.000	0.987	0.994
PRC Area	0.999	1.000	0.995	1.000

Fig. 3. Comparative study of accuracy measures for our data set

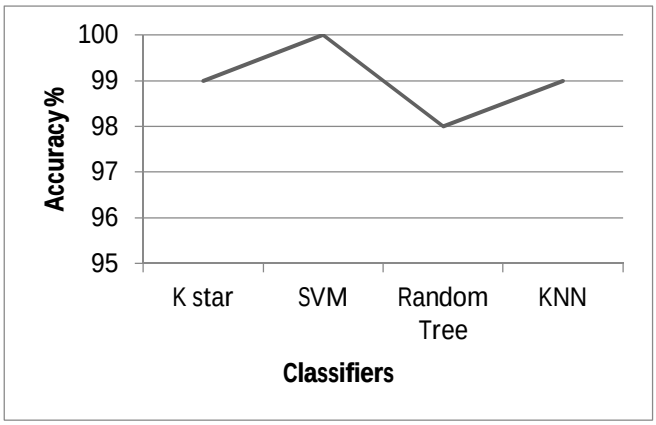


From Table 2 and Fig. 3, the basic overall accuracy of SVM is better than the other three algorithms namely Kstar, Random Tree, KNN.

Accuracy does not always gives true performance measure of the investigated algorithm. For this reason the classifier performance is usually expressed in terms of Receiver Operating Characteristic (ROC) curve [10].

To evaluate the performance of each classifier we have calculated the area under ROC curve (AUC) with their respective classifier.

Fig. 4. ROC curves for the given data mining techniques



In the above Fig. 4, it demonstrates that performance of SVM under ROC curve is better when comparing with other three algorithms.

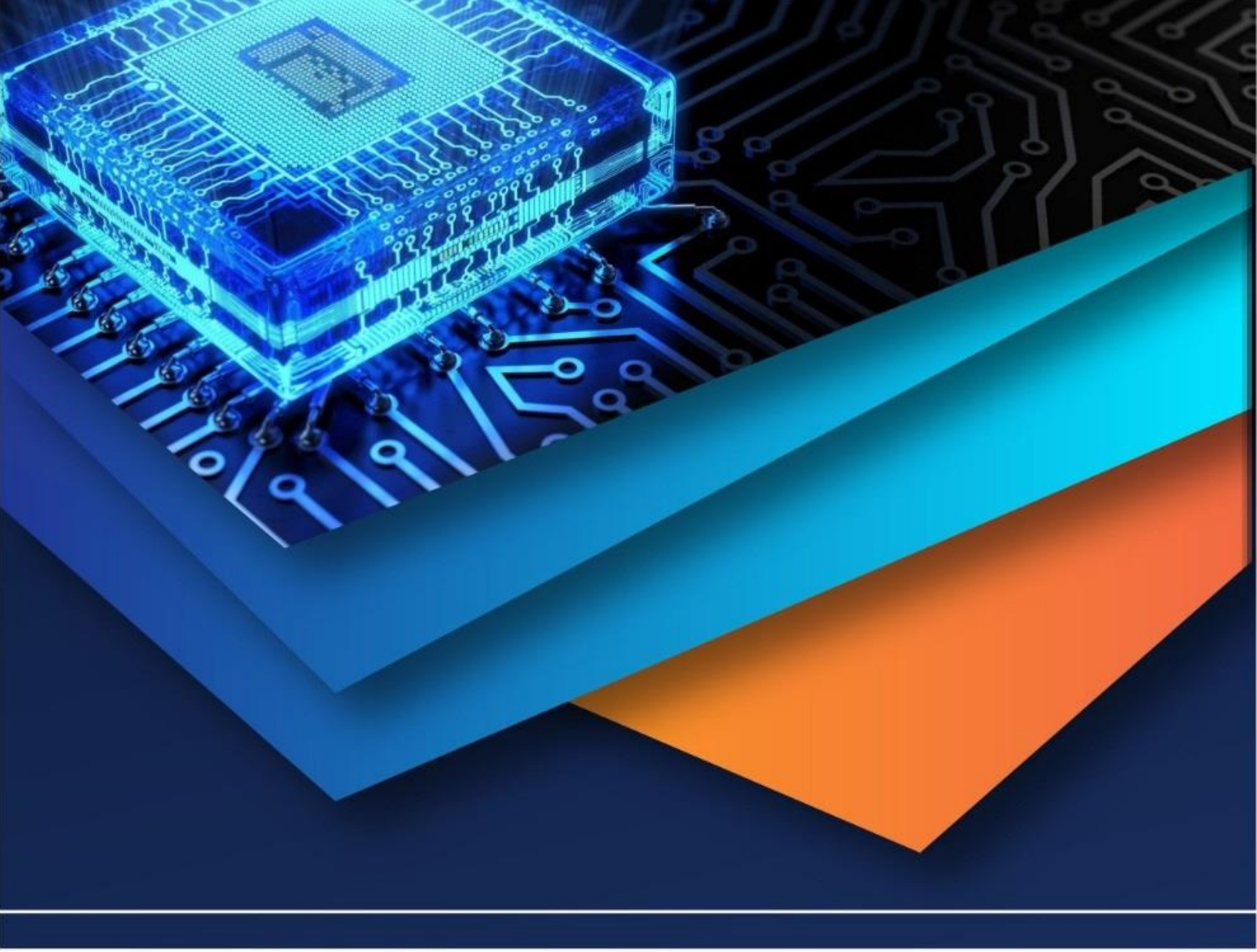
VII. CONCLUSION

The overall objective of our work is to predict more accurately the presence of heart disease in human body. Heart disease is one of the most dangerous disease that causes millions of death in the World. This paper involves prediction of heart disease based upon the given attributes. The heart disease becomes a multitude throughout the world. It is a difficult task to predict, which requires higher knowledge for prediction. Data Mining plays an important role in decision making which extracts hidden information.Four data mining classification techniques were applied namely SVM, Kstar, Random Forest and KNN. From results it has been seen that SVM provides better accurate results as comparing with other three techniques.This paper uses data mining techniques to find out the accurate prediction for heart disease. Heart Disease can be caused by several factors such as blood pressure, cholesterol, age, gender, blood sugar, chest pain type etc. It has proven that classification based techniques contribute high effectiveness and obtain high accuracy compare than the previous method.

REFERENCES

[1] Micheline Kaber and Jian Pei Jiawei Han, Data Mining Concepts and Techniques, 3rd ed., 2012.
[2] Mohammed Abdul Khaled, Sateesh Kumar Pradhan and G.N Dash, “A survey of data mining techniques on medical techniques on medical data for finding locally frequent disease”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, pp. 149-153, August 2013.
[3] C.Sowmiya and Dr. P. Sumitra,”Analytical Study of Heart Disease Diagnosis Using Classification Techniques”, 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT].
[4] Miss. Chaitrali S. Dangare, Dr. Mrs. Sulabha S. Apte. (2012). A Data Mining Approach For Prediction of Heart Disease Using Neural Networks. International Journal of Computer Engineering & Technology (IJCET), 30-40.
[5] Salha M. Alzahani, Afnan Althopity, Ashwag Alghamdi, Boushra Alshehri, and Suheer Aljuaid, “An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction”, Lecture Notes on Information Theory Vol. 2, No.4, December 2014, PP 310-315.
[6] Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M, “Heart Disease Diagnosis Using Data Mining Technique”, International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017.International Conference on Electronics, Communication and Aerospace Technology, ICECA 2017.
[7] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu “ Heart Disease Prediction System using Associative Classification and Genetic Algorithm”, ICECIT, 2012.
[8] Swati Shilaskaret al, Feature selection for medical diagnosis: Evaluation for cardiovascular diseases”, Journal of Expert System with Application, Vol.40, pp 4146-4153, 2013
[9] RaghunathNambiar, AdhiraajSethi, RuchieBhardwaj, Rajesh Vargheese,” A Look at Challenges and Opportunities of Big Data Analytics in Healthcare”, 2013 IEEE International Conference on Big Data.
[10] Helma C, Gottmann E, Kramer S (2000) Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research 9: 329-358.
[11] Marjia Sultana*, Afrin Haider and Mohammad Shorif Uddin,” Analysis of data Mining Techniques for Heart Disease Prediction”, ICECIT, 2016.
[12] Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale, “Heart Disease Prediction using data mining techniques”, 2017 International Conference on Intelligent Computing and Control (I2C2).
[13] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1):1–37, 2008.
[14] J.G. Cleary and L.E. Trigg. K*: An instance based learner using an entropic distance measure. In In Proceedings of the 12th International Conference on Machine Learning, pages 108–114, 1995.
[15] Wikipedia contributors, ―Random tree, ‖ Wikipedia, The Free Encyclopedia. Wikimedia Foundation, 13-Jul- 2014.
[16] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on, 13(1):21–27, 1967.

[17] Theresa Princy. R, J. Thomas,"Human Heart Disease Prediction System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY