ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Prediction and Analysis of Heart Disease

Sonali S. Jagtap[1]

M.E. Student, Dept. of CSE, CSMSS College of Engineering, Dr. BAM University, Aurangabad, Maharashtra, India

ABSTRACT: Tremendous amount of people are prone to heart deformity and in today's world lot of people are prone to these diseases that is CVD (Cardiovascular disease) and it is one of the major reasons for heart problem. Consequently we examined and determine first stage it will increase death factor illness and causes death. There is no adequate research refereeing to the tools to discover relationships and trends in data especially in the medical sector. Complex clinical data is driven by technologies such as health care about patients and other hospital resources. Rich collection are used in order to Data mining techniques examine methodology in detail the different perspectives and deriving useful detail. Our project is perspective to design and develop futuristic prediction regarding details of health system for heart diseases based on predictive mining. There are various experiments have been conducted to a specific relationship for the performance of various predictive data mining techniques including Decision tree, Naïve Bayes and k-mean algorithms. In this proposed work, a 14 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree, Naive Bayes and k-mean have been prescribed and their schedule on diagnosis has been compared. Clustering outperforms when compared to Decision tree & Naive Bayes..

KEYWORDS: Data Mining, Decision Tree, Naive bayes, k-mean, Heart

I. INTRODUCTION

Heart disease is nothing but the class of diseases that involve the heart or blood vessels (arteries and veins). Today most countries face high and growing rates of heart disease and it has become a leading cause of debilitation and death worldwide in men and women over age sixty-five and today in many countries heart disease is viewed as a "second epidemic," replacing infectious diseases as the leading cause of death[1]. Most countries face high and increasing rates of heart disease or Cardiovascular Disease. Even though, modern medicine is generating huge amount of data every day, little has been done to use this available data to solve the challenges that face a successful interpretation of heart disease examination results.

Data mining is a knowledge discovery technique to analyze data and encapsulate it into useful information [1]. The current research intends to predict the probability of getting heart disease given patient data set [5]. Predictions and descriptions are principal goals of data mining, in practice [6]. Prediction in data mining involves attributes or variables in the data set to find unknown or future state values of other attributes [7]. Description emphasize on discovering patterns that explains the data to be interpreted by humans [6].

The purpose of predictions in data mining is to help discover trends in patient data in order to improve their health [1]. Due to change in life styles in developing countries, like South Africa, Cardio Vascular Disease (CVD) has become a leading cause of deaths [5]. CVD is projected to be a single largest killer worldwide accounting for all deaths [3]. An endeavor to exploit knowledge, experience and clinical screening of patients to diagnose or recognize heart attacks is regarded as a treasured opportunity [2]. In the health sectors data mining plays an important role to predict diseases [7]. The predictive end of the research is a data mining model.

II. RELATED WORK

Numerous works related to heart disease diagnosis using data mining techniques have motivated this study. A model Intelligent Heart Disease Prediction System (IHDPS) assembled with the aid of data mining techniques like Decision Trees, Naive Bayes and Neural Network was proposed [6], they used a CRISP-DM methodology to assemble

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

the mining models on a dataset obtained from the Cleveland Heart Disease database. The results illustrated the unexpected strength of each of the methodologies in comprehending the objectives of the specified mining objectives. IHDPS was capable of answering queries that the conventional decision support systems were not able to. It facilitated the establishment of vital knowledge, e.g. patterns, relationships amid medical factors connected with heart disease.

The researchers [7] describe explores the utility of various decision tree algorithm in classify that is classification, KDD, J48 and predict the disease [7]. Heart Disease is a fatal disease by its nature. This disease makes a life threatening complexities such as heart attack and death. The importance of Data Mining in the Medical Domain is realized and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied. The observations from the previous work have led to the deployment of the proposed system architecture for this work. Though, various classification techniques are widely used for Disease Prediction, Decision Tree classifier is selected for its simplicity and accuracy. Different attribute selection measures like Information Gain, Gain Ratio, Gini Index and Distance measure can be used.

Another study conducted[8] by experimented on a sample database of patients' records. The Neural Network is tested and trained with 14 input attribute such as Age, Blood Pressure, Angiography's report and the like. The supervised network has been recommended for diagnosis of heart diseases. Training was moved out with the aid of back propagation algorithm. Whenever unknown data was fed by the doctor, the system identified the unknown data from comparisons with the trained data and generated a list of possible diseases that the patient is unsafe to. The success rate for imprecise inputs to recover the desired output is closest to 100%.

The problem of recognizing constrained association rules for heart disease prediction was studied[9]. The assessed dataset encompassed medical records of people having heart disease with attributes for risk factors, heart perfusion measurements and artery narrowing. Three constraints were introduced to decrease the number of patterns. First one necessitates the attributes to appear on only one side of the rule. The second one segregates attributes into uninteresting groups. The ultimate constraint controls the number of attributes in a rule. Experiments illustrated that the constraints reduced the number of discovered rules remarkably except decreasing the running time. Two groups of rules envisaged the presence or absence of heart disease in four specific heart arteries.

In this paper authors have used four classification algorithms such as J48, Random Forest (RF), Reduce Error Pruning (REP) and Logistic Model Tree (LMT) to classify the "WEATHER NOMINAL" open source Data Set[10]. This paper authors have examined J48, RF, REP and LMT method of classification and observed that RF is having maximum accuracy and minimum error rate. On the basis of accuracy measures, of the classifiers one can easily provide the instruction regarding fault-prone prediction issues of any given data set in the respective situations.

Classifier Algorithms	Instances Correctly Predicted	Instances Incorrectly Predicted	Accuracy in %
J48	6	8	42.85
RF	8	6	57.14
REP	7	7	50.00
LMT	7	7	50.00

Presented a predictive model for the Ischemic Heart Disease (IHD); they applied Back-propagation neural network (BPNN), the Bayesian neural network (BNN), the probabilistic neural network (PNN) and the support vector machine (SVM) to develop classification models for identifying IHD patients on a data obtained from measurements of cardiac magnetic field at 36 locations (6 × 6 matrices) above the torso[11]. The result shows that BPNN and BNN gave the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

highest classification accuracy of 78.43 %, while RBF kernel SVM gave the lowest classification accuracy of 60.78 %. BNN presented the best sensitivity of 96.55 % and RBF kernel SVM displayed the lowest sensitivity of 41.38 %. Both polynomial kernel SVM and RBF kernel SVM presented the minimum and maximum specificity of 45.45 % and 86.36 %, respectively.

After reviewing the above literatures the researcher was motivated to work on a classification model that is sought to predict heart disease generated from University Hospital, Zurich, Switzerland (Switzerland data) database. This project intends to design and develop diagnosis and prediction system for heart diseases based on predictive mining. Number of experiments has been conducted to compare the performance of various predictive data mining techniques including Decision tree, Clustering and Naive Bayes algorithms. In this proposed work, a 13 attribute structured clinical database from UCI Machine Learning Repository has been used as a source data. Decision tree, Clustering and Naive Bayes have been applied and their performance on diagnosis has been compared. Clustering outperforms when compared to Decision tree & Naive Bayes. In this paper the predictive accuracy will be calculated using three algorithm that is Decision Tree, Naive bayes and K-mean. On the basic of this algorithm get the comparison of predictive accuracy and shows which algorithm is better for prediction of heart disease.

III. PATIENT DATASET

The patient data set is compiled from data collected from medical practitioners in University Hospital, Zurich, Switzerland. Only 14 attributes from the database are considered for the predictions required for the heart disease. The following attributes with nominal values are considered: Age, Sex, cp(chest pain), trestbps(resting blood pressure), chol(serum cholestoral), fbs(fasting blood sugar), restecg(resting electrocardiographic), thalach(maximum heart rate), oldpeak, slope, ca, thal, num.

IV. EXPERIMENTATION

Our study was perspective to design and develop futuristic prediction regarding details of health system for heart diseases based on predictive mining. There are different algorithm that is J48, NAIVE BAYES, REPTREE, BAYES NET, SIMPLE CART used in previous work. They show only two classes as present or absent. we apply k-mean algorithm for same to predict and analysis of heart disease with different attribute and different records by increasing classes. we want to study what effect is on results as compare to J48 and Naive bayes.

Clustering is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks. Due to the enormous size of many present day databases, it is often helpful to apply clustering analysis first, to reduce the search space for the downstream algorithms

K-means Algorithm
Use:
For partitioning where each cluster's center is represented by the mean value of the objects in the cluster.
Input:
k: the number of clusters,
Output:
A set of k clusters.
Method:
Step 1: Choose k numbers of clusters to be determined.
Step 2: Choose C k centroids randomly as the initial centers of the clusters.
Step 3: Repeat
    3.1: Assign each object to their closest cluster center using Euclidean distance.
    3.2: Compute new cluster center by calculating mean points.
Step 4: Until

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

4.1: No change in cluster center OR
4.2: No object changes its clusters.

This study used k-mean algorithm to find out the predictive accuracy and find out the centroid distance by using Euclidean distance formula.

Three data mining goals were defined based on exploration of the heart disease dataset and objectives of this research. They were evaluated against the selected model. Results show that the selected model had achieved the stated goals, suggesting that the model could be used in heart disease diagnostic process.

The first data mining goal set for this study was given patients result for each attribute, classify patients into two categories; those who are diagnosed with heart disease and those who are free from the disease. The selected model built with J48 algorithm was able to answer this question by predicting 100% of the cases correctly. As the classification accuracy value is high, cardiologists could rely on it for assisting heart disease diagnosis. The second data mining goal for this study was given patients result for each attribute, the selected model built with Naive Bayes algorithm was able to answer this question by predicting 98% of the cases correctly. The third data mining goal for this study was given patients result for each attribute, the selected model built with K-Mean algorithm was able to answer this question by predicting 100% of the cases correctly.

V. RESULTS AND ANALYSIS

No. of Records in Training dataset	No. of records in Testing dataset	No. of classified instances	Number of Classes	Accuracy (%)
270	10	10	0 10(100%)	100%
270	20	20	0 4 (20%) 1 10 (50%) 2 6 (30%)	100%
270	30	30	0 11 (37%) 1 10 (33%) 2 9 (30%)	100%
270	57	57	0 18 (32%) 1 12 (21%) 2 18 (32%) 3 9 (16%)	100%
270	75	75	0 25 (33%) 1 18 (24%) 2 24 (32%) 3 8 (11%)	100%
270	84	84	0 14 (17%) 1 43 (51%) 2 13 (15%) 3 14 (17%)	100%
270	100	100	0 20 (20%) 1 34 (34%) 2 16 (16%) 3 10 (10%) 4 20 (20%)	100%
270	120	120	0 16 (13%) 1 36 (30%) 2 24 (20%) 3 20 (17%) 4 24 (20%)	100%

Table 1: Predictive performance of the classifiers K-Mean with different records

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

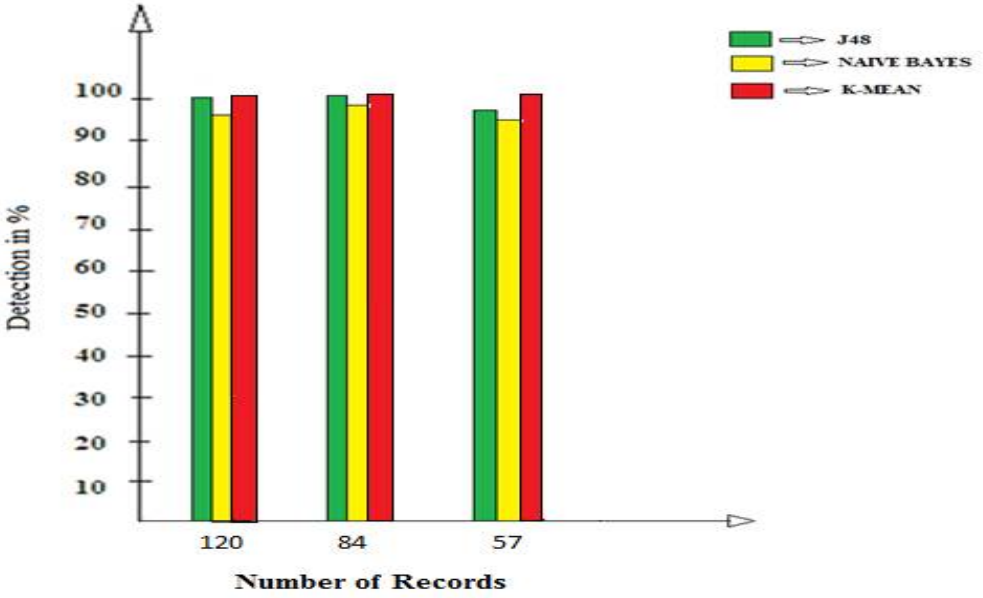

Figure 1: Comparison between J48, Naive Bayes and K-mean with different records.

Table 1 shows the Predictive performance of the classifiers K-Mean with different records. We apply k-mean algorithm for same to predict and analysis of heart disease with different attribute and different records by increasing classes in this table 1, training records is fixed i.e. 270 records and we apply different records of testing dataset. After applied testing records we get number of classified instances mean correctly classified record with number of classes. If number of records of training dataset was 10 that time get only 1 class with 100% accuracy. If we increase testing records we get increased number of classes with 100% accuracy. Figure 1 shows the comparisons between J48, Naive Bayes and K-Mean with different records. This figure shows the highest predictive accuracy of J48, Naive Bayes and K-Mean of fixed training dataset that is 270 with 120, 84, 57 records of training dataset. A general observation on the dataset with regards to accuracy is the dimensionality of the class attribute. This means, the smaller the dimension or attribute values for the class variable; the higher the accuracy of the model. This was observed from the 'classified' and the 'clustered' dataset. The classified dataset has a class with two attribute values (i.e. Present and Absent), thus; having a model with the highest accuracy to be 100%. This accuracy is equal to compared to the clustered dataset which has four clusters (i.e. cluster1, cluster2, cluster3 and cluster4) as values for the class attribute, and the accuracy obtained from using the clustered dataset to build a model was across all algorithms used under different experiment. To prove this further, the dataset was clustered into four clusters and the same test specifications which yielded 100% accuracy on the classification experiment was used on the clustering experiment on the four clustered outcomes; but the highest accuracy was 100%. This means that the clustering experiment achieved 100% accuracy compared to the classification experiment outcome. All three models performed well in predicting heart disease cases. The most effective model to predict patients with heart disease appears to be a K-Mean and J48 classifier implemented on selected attributes with a classification accuracy of 100% and performs better in predicting heart disease with classification accuracy of 100%.

VI. CONCLUSION AND FUTURE WORK

The findings of this study revealed all the models built from Decision Tree classifier, Naive Bayes classifier and K-mean have high classification accuracy and are generally comparable in predicting heart disease cases. However, comparison that is based on True Positive Rate suggests that the J48 model and k-mean performs better in predicting heart disease with classification accuracy of 100%. The performances of the models were evaluated using the standard

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

metrics of accuracy, precision. Cross Validation was adopted for randomly sampling the training and test data samples. All three models performed well in predicting heart disease cases. J48 and Naive Bayes showed only two classes as present or absent. We applied k-mean algorithm for same to predict and analysis of heart disease with different attribute and different records by increasing classes. Decision Support in Heart Disease Prediction System is developed using J48, Naive Bayes Classification and Clustering. The system extracts hidden knowledge from a historical heart disease database. This model could answer complex queries, each with its own strength with ease of model interpretation and an easy access to detailed information and accuracy. The system is expandable in the sense that more number of records or attributes can be incorporated and new significant rules can be generated using underlying Data Mining technique. Presently the system has been using 14 attributes of medical diagnosis. It can also incorporate other data mining techniques and additional attributes for prediction. This study showed that data mining techniques can be used efficiently to model and predict heart disease cases. The outcome of this study can be used as an assistant tool by cardiologists to help them to make more consistent diagnosis of heart disease.

As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types.

REFERENCES

1. A. Aziz, N. Ismail, and F. Ahmad, "Mining Students' Academic Performance", Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.
2. S. M. Kamalapur, S. Reddy, "Women Health in India: An Analysis", International Research Journal of Social Sciences, Vol.2 (10), 11-15, October 2013.
3. N. A. Sundar, P. P. Latha, and M. R. Chandra, "Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Data Base", International Journal of Engineering Science & Advanced Technology, vol. 2, no. 3, pp. 470– 478, 2012.
4. Patil B.P, Dr. Y. S. and Kumaraswamy, "Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction", IJCSNS International Journal of Computer Science and Network Security, Vol. 9 No. 2, February 2009.
5. J. Liu, Y. T. HSU, and C. L. Hung, "Development of Evolutionary Data Mining Algorithms and their Applications to Cardiac Disease Diagnosis", in WCCI IEEE World Congress on Computational Intelligence, pp. 10–15, 2012.
6. Ms. Ishtake S.H, Prof. Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research, Vol. 1, Issue 3, April 2013.
7. K. Thenmozhi and P. Deepika, " Heart Disease Prediction Using Classification with Different Decision Tree Techniques", International Journal of Engineering Research and General Science, Vol. 2, issue 6, October-November, 2014.
8. Guru, N., Anil D., Navin, R., "Decision Support System For Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. 1, January-June 2007.
9. Chaitrali S. D., Sulabha S. A., "Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques", International Journal of Computer Application, Vol. 47, No.10, June 2012.
10. D. L. Gupta, A. K. Malviya, Satyendra Singh, "Performance Analysis of Classification Tree Learning Algorithms", International Journal of Computer Applications, Vol. 55, No.6, October 2012.
11. Kangwanariyakul Y., Chanin N., Tanawut T., Thanakorn N., "Data Mining of Magneto cardiograms for Prediction of Ischemic Heart Disease", EXCLI Journal, Vol. 33, No.9, 2000.
12. Hemlata S., Sharma S., Gondhalakar S., "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering, Vol. 1, Issue 3, 2012.

BIOGRAPHY

Sonali S Jagtap is a ME Student in CSE, CSMSS College of Engineering Aurangabad, Dr. BAM University, Maharashtra.