Classification of Heart Diseases Patients using Data Mining Techniques

Mukesh Kumar[1], Shankar Shambhu[2], Abha Sharma[3]
Chitkara University School of Engineering and Technology
Chitkara University, Himachal Pradesh, India
mukesh.kumar@chitkarauniversity.edu.in[1], *shankar.shambhu@chitkarauniversity.edu.in*[2], *abha.sharma@chitkarauniversity.edu.in*[3]

***Abstract*—** In healthcare sector, data are enormous and diverse because it contains a data of different types and getting knowledge from these data is crucial. So to get that knowledge, data mining techniques may be utilized to mine knowledge by building models from healthcare dataset. At present, the classification of heart diseases patients has been a demanding research confront for many researchers. For building a classification model for a these patient, we used four different classification algorithms such as NaiveBayes, MultilayerPerceptron, RandomForest and DecisionTable. The intention behind this work is to classify that whether a patient is tested positive or tested negative for heart diseases, based on some diagnostic measurements integrated into the dataset.

***Keywords*—** *Heart disease, Classification, MultilayerPerceptron, RandomForest, DecisionTable, NaiveBayes*

I. INTRODUCTION

As we are aware that one of the main reason of death is heart disease in both men and women worldwide. To prevent the same we need to take early action to increase survival chances. In general we are aware of some basic warning signs of heart attack like discomfort in chest, pain in left arm, short breath, sweating. Few of major risk factors to heart attack n clued high BP, high cholesterol, overweight, poor diet, no physical activity, high stress etc. To prevent heart attack on should control blood pressure, take healthy diet and control cholesterol level. Regular exercise is important to maintain healthy weight. In today's scenario lifestyle is very hectic so stress management is very important and of course keeping a check on alcohol intake.

Researchers all over the world are following different Machine Learning (ML) techniques to classify and predict the reason of heart disease. Data mining helps to formulate and analyse the data to derive data from existing data base [1]. According to American Heart Association (AHA) for the first time in last 50 years, a statistical report has been released showing the death rates caused by cardio vascular diseases in men and women. The age group considered is between 35 to 74 years. These statistics will be helpful to improve and save life of people worldwide.

II. LITERATURE REVIEW

As per Author [1] data mining is a process through which one should be able to understand problem definition. Similarly data collection is very important to identify familiar data. Once the data is gathered it should be checked for correctness and structured in a proper way. All these factors are important to implement data mining in accurate way to produce best results.

Whereas [2], follows K-Nearest Neighbor algorithm to test data. KNN is one of the simplest methods and produces consistent results. Factor considered under this study are age, minimum and maximum pulse rate and blood pressure. The comparison of result was based on previous history of patient.

Another study [3] aided towards early diagnosis of the heart disease and risk factors. The accuracy of results was around 84% using support vector machine and NaiveBayes. Factors under consideration for defining accuracy were chest pain, heart rate and various other attributes. SVM is intended to produce results at a lower rate and is expected to be used for future reference.

Comparison with different classifiers has been done to analyse the best possible method that produces unbiased estimate. To predict chances of heart disease a total of 10 attributes have been considered. All the classifiers were implemented using WEKA tool [4].

III. DATASET DESCRIPTION USED FOR PREDICTION

In our research work, we are taking this dataset from ***https://github.com/renatopp/arff-datasets/blob/master/classification/heart.statlog.arff***. Table-1 below gives us all the detail regarding our dataset taken into consideration. The heart diseases patient's dataset used here contain 270 different records with 14 attributes. The major objective of this work is to classify whether a patient is heart patient or not, based on firm diagnostic measurements integrated in the dataset.

Data available in real world is incomplete especially in the area of a medical sector. So to remove unnecessary and noise in the data, we perform the pre-processing on the data. Pre-processing of data is a very important stage in this research work as it affects classification results of the heart diseases

patients. Initially, unwanted and noisy data is removed from the record and secondly, data mining techniques algorithm is applied to build a classifier model. Classifier Modelling means selecting diverse techniques and applying them to different data dataset of the same type. In this research paper, four different classifications are implemented like NaiveBayes, MultilayerPerceptron, RandomForest and DecisionTable to build the best classifier model for our dataset.

Table 1: Dataset description used to implement classification algorithm

S. No.	Attribute	Description of Attribute	Type
1	age	Age of the patients	Real
2	sex	Gender of the patients	Binary
3	CP	Chest Pain	Nominal
4	restbp	Resting Blood Pressure	Real
5	Cholestoral	Serum Cholestoral	Real
6	fbs	Fasting Blood Sugar	Binary
7	restec	Resting Electrocardiographic	Nominal
8	hr	Max Heart Rate	Real
9	exercise	Exercise induced Angina	Binary
10	oldpeak	ST depression induced by exercise relative to rest	Real
11	slope	Slope of the peak exercise ST segment	Ordered
12	vessels	major vessels (0-3) colored by flourosopy	Real
13	thal	Thal	Nominal
14	Result	Present/absent	

IV. EXPERIMENTAL SETUP

The dataset is simulated and analyzed in WEKA toolkit. WEKA is freely available software with already compiled algorithms of machine learning to perform data mining tasks. Data mining helps in finding precious information concealed in enormous amounts of data. For this purpose, we have WEKA toolkit which has the collection of complied algorithms of machine learning for data mining purpose. Here, we have used K-cross validation test where value of K is 10. We used this validation test option to minimize process distortion and recover process efficiency. The four different classifiers NaiveBayes (NB), MultilayerPerceptron (MP), RandomForest (RF) and DecisionTable(DT) were simulated on WEKA toolkit. The results of simulation are showing that the considered technique gives good results in the literature as compared to other similar methods, taking into consideration. Figure 1 show the result of dataset uploaded into the WEKA tool kit.

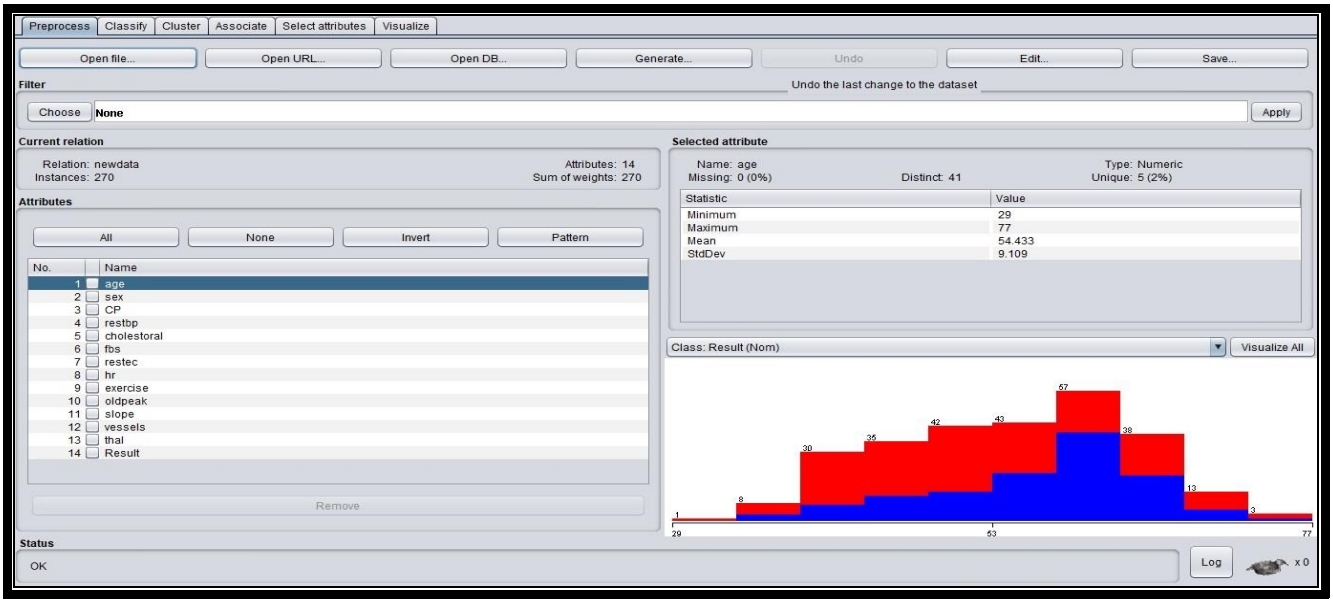

Figure 1: Environmental Setup of Weka Tool for Implementation

V. RESULT AND DISCUSSION

Here to develop the classification model for patients of heart we are implemented and analyzed four classification algorithms NaiveBayes, MultilayerPerceptron, RandomForest and DecisionTable. In this research paper, for prepare training dataset to prepare our model and for testing dataset to test the developed model we used 10-fold cross validation. First, we check our dataset for baseline accuracy with ZeroR classification algorithm. The baseline accuracy for our dataset is 55.55 %, which implies that we need to do a lot of work to improve the accuracy of our dataset. Table 2 shows the experimental result of different classification algorithms. We have conceded some implementation to estimate the accuracy and effectiveness of multiple classification algorithms for classifying dataset of heart patients.

Table 2: It shows the performance of different Classifiers used for implementation

Evaluation Criteria for Classification Algorithm	**Classification Algorithms used**			
	NB	**MP**	**RF**	**DT**
Timing to build model (in Sec)	0 Sec	0.43 Sec	0.05 Sec	0.03 Sec
Correctly classified instances	75	69	72	75
Incorrectly classified instances	11	12	14	14
Accuracy (%)	87.20%	85.18 %	83.72 %	84.26 %

Here we show that the NaiveBayes classification algorithm has performed well with more accuracy as compared to other algorithm used. In the WEKA tool, a classified instance whose percentage has accurately classified is called accuracy of the classifying model. Other evaluation criteria for classification algorithm implementation are Kappa Statistic (KS), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) & Root relative squared Error (RRSE). In table 3, we illustrate simulation result for a different algorithm with their evaluation criteria.

Table 3: Training and Simulation error of each classifier used in the implementation

Evaluation Criteria	**Classification Algorithms used**			
	NB	**MP**	**RF**	**DT**
Kappa statistic(KS)	0.745	0.745	0.676	0.6868
Mean absolute error(MAE)	0.1705	0.1705	0.2691	0.2656
Root mean squared error (RMSE)	0.3387	0.3387	0.345	0.3643
Relative absolute error (RAE)	33.64 %	33.64 %	53.09 %	52.55 %
Root relative squared error(RRSE)	65.65 %	65.65 %	66.88 %	70.82 %

To choose the algorithms for soaring performance, different algorithms are implemented and evaluated with respect to some evaluation criterion on selected dataset. The classification algorithm which achieves the utmost performance in provisions of soaring specificity and sensitivity value is measured by the finest algorithm. From table 4, it is clear that the NaiveBayes classification algorithm achieves the maximum value.

The efficiency of the machine learning classifier can be assessed with numerous measures. The estimate of these measures basically depends on the contingency table which is further obtained from the classification algorithm implemented. Table 4; contain the value of the contingency table of a particular heart patient dataset.

Table 4: Comparison of accuracy measures of each classifier used in an implementation

Classifier	TP	FP	Precision	Recall	Class
NaiveBayes (68 percentage split)	0.826	0.075	0.927	0.826	present
	0.925	0.174	0.822	0.925	absent
MultilayerPreceptron (70 percentage split)	0.826	0.075	0.927	0.826	Present
	0.925	0.174	0.822	0.925	Absent
RandomForest (68 percentage split)	0.783	0.100	0.900	0.783	Present
	0.900	0.217	0.783	0.900	Absent
DecisionTable (68 percentage split)	0.787	0.095	0.902	0.787	Present
	0.905	0.213	0.792	0.905	Absent

Any classification algorithm performance is extremely depending on the nature of dataset that is used for training. In WEKA tool, confusion matrices which are generated after simulation of classification algorithm are useful to evaluate different classifiers. In the confusion matrix, columns represent the predicted classification classes, and rows represent the actual classes.

Table 5: Confusion Matrix of each classifier used in an implementation

Classifier	TP	FP	Class
NaiveBayes	38	8	Present
	3	37	Absent
MultilayerPreceptron	38	8	Present
	3	37	Absent
RandomForest	36	10	Present
	4	36	Absent
DecisionTable	37	10	Present
	4	38	Absent

Based on the above Table 2, we noticeably see that the maximum accuracy is 87.20% for NaiveBayes and the minimum accuracy is 83.72% for RandomForest classification algorithm. These entire algorithm discussed above are tested using 10-cross validation. In our work, out of 270 different records, approximately 68% of data are used as training data and rest of the data is taken as test data.

Table 5. Shows use the confusion matrix of each classifier used for implementation here. In NaiveBayes algorithm, out of 86 testing data 75 is tested correctly and 11 are tested incorrectly classified. But in case of RandomForest out of 86 testing data, 72 are correctly classified and 14 are incorrectly classified.

VI. CONCLUSTION

In this paper, we are considered the dataset of heart diseases patients which is further collected at National Institute of Heart and Digestive and Kidney Diseases. The dataset has 270 instances with fourteen different attributes. We are simulated NaiveBayes, MultilayerPerceptron, RandomForest and DecisionTable classification algorithm and found that the NaiveBayes have the maximum accuracy (87.20%) for classifying the heart patients whether they are positive or

negative. In the future, the results can be utilized to create a control plan for Heart patients because Heart patients are normally not identified until a later stage of the disease or the development of complications.

References

[1] K. Sudhakar, Dr. M. Manimekalai, "Study of Heart Disease Prediction using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, pp.1157-60, January 2014.

[2] R. Chitra, V. Seenivasagam, " Review of heart disease prediction system using data mining and hybrid intelligent techniques", ICTACT JOURNAL ON SOFT COMPUTING, July 2013, volume: 03, issue: 04 pp.605-09.

[3] C. Ordonez, "Association rule discovery with the train and test approach for heart disease prediction," IEEE transactions on Information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society, vol. 10, no. 2,pp.334–43,Apr.2006

[4] H. Chen, S. Y. Huang, PS Hong, CH Cheng, EJ Lin, Heart Disease Prediction System, Computing in cardiology, pp. 557-560 2011.

[5] Vikas Chaurasia, and Saurabh Pal, Data Mining Approach to Detect Heart Dieses, International Journal of Advanced Computer Science and Information Technology, Vol. 2, No. 4, pp. 56- 66, 2013.

[6] Abhishek Taneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Journal of Computer Science and technology, Vol. 6, No. 4, 2013.

[7] K. S. Kavitha, K. V. Ramakrishnan, Manoj Kumar Singh, International Journal of Computer Science, vol. 7, issue 5, pp. 272-283, 2010.

[8] K. Usha Rani, Analysis of Heart Disease Dataset Using Neural Network approach International Journal of Data Mining & Knowledge Management Process, Vol. 1, No. 5, September 2011.

[9] V. Sugumaran, V. Muralidharan and K.I. Ramachandran, Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing, Mechanical Systems and Signal Processing, Volume 21, Issue 2, Pages 930-942, February 2007.

[10] Srinivas, K.," Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[11] Yanwei Xing, "Combination Data Mining Methods with New Medical Data to Predicting Outcome of Coronary Heart Disease", IEEE Transactions on Convergence Information Technology, pp(868 – 872), 21-23 Nov. 2007

[12] S. K. Yadav and Pal S., "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology (WCSIT), 2(2), 51-56, 2012.