Prediction of Heart Disease using Decision Tree Classification Algorithms

S.Spino[#1], Dr.M.Mohamed Sathik[#2],Dr.S.Shajun Nisha[#3]

1M.Phil Research Scholar, PG and Research Department of Computer Science, SadakathullahAppa College, Tirunelveli, Tamilnadu, India

2Principal & Research Coordinator, Research Department of Computer Science, SadakathullahAppa College, Tirunelveli, Tamilnadu, India

3Assistant Professor and head, PG and Research Department of Computer Science, SadakathullahAppaCollege,Tirunelveli, Tamilnadu, India

Abstract

Cardiovascular Sickness is a major reason of mortality in the present living style. The Cardiovascular has three vital function such as transport of nutrients, oxygen and hormones to cell all over the organs and removal of metabolic wastes. Heart Disease refers to various types of state that can affect heart function. Data Mining Techniques has been accomplished in healthcare domain. Data Mining is used to discover intriguing and useful patterns from huge data sets. Data mining techniques which are applicable to medical data Include association rule mining for Classification. Data mining Classification technique forecast the heart disease risk level of each person based on attributes such as age, gender, Blood pressure, cholesterol, pulse rate.This Paper focuses around the prediction of heart disease accuracy value using the Decision Tree classification technique.

Keywords - *Data mining, Heart Disease, Decision Tree, Random Forest, J48.*

I. INTRODUCTION

Data mining tool have been generated for compelling inquiry of Medical data, so as to help clinicians in get better their termination for the treatment purpose. In Cardiovascular disease research, data mining approach have played out a massive role. The effectively existing medical data is an obvious and extraordinary methodology in the investigation of heart related infection characterization to uncover the disguised medicinal data. Decision Tree models are frequently used in data mining to inspect data and induce the tree and its rules that will be used to make prediction. The heart is a vitally important organ of our body. The heart functions as a pump in the circulatory system to deliver ceaseless flow of blood throughout the body. This ceaseless movement consists of the system circulation to and from the body and the pulmonary circulation to and from the Lungs .If the operation of a heart is not proper, it will impact other parts of a human. Cardiovascular diseases are one of the most noteworthy airborne infections of the Modern world. There are number of risk factors which increment the Heart disease such as High pressure, Corpulence, Cholesterol, Smoking, Being physical inactive, Eating an unhealthy foods .The aim of this paper is to employ and analysed the data mining Classification technique to predict the heart disease risk level in a patient through extraction of interesting patterns from the dataset using vital parameters. The aim of this paper is to employ and analysed the data mining Classification technique to predict the heart disease risk level in a patient through extraction of interesting patterns from the Heart Disease data set of the UCI Learning Repository. The inquiries are conducted with WEKA tool and the Decision Tree algorithms applied on the heart dataset.

A. Literature Survey

Data mining techniques and functions are used to recognize the level of risk factors which assist the patients to take precautions in advance to save their life [4]. Data mining algorithms like Decision trees (J48), Bayesian classifiers, Multilayer perceptron, simple logistic and Ensemble techniques are employed to resolute the heart sickness. Several types of studies have been done to centre on prediction of heart disease. Different types of data mining techniques are used for identify and reached different accuracy level for different methods. Different classification techniques used for forecast the risk level of every one based on age, gender, Blood pressure, cholesterol. Naive Bayes, KNN, Decision Tree Algorithm, Neural Network is used to classify the patient risk level.Data mining technique in discover of heart disease, this research centre on using three classification techniques called Decision Tree ,Naive Bayes and KNN, the datasets are obtained from the UCI Learning Repository ,it consist of 14 attributes . The dataset examined using the KNN algorithm makes a precision of 100%.The decision tree technique of the Naive Bayes assess to 92.0792% of correctly classified instance .The Analogy of these classification algorithms solves to bring forth the KNN algorithm as a foremost method of predicting the Heart diseases[4].

B. *Motivation and Justification*

Decision Tree provides a logical method of decision making because it distinctly lay out the issue so that all the view point can be challenged and it also allow to analyse overall feasible outcomes of a decision. Programmatically Decision Tree can be used to allocate monetary and time or additional values to feasible results so that decision can be automated.

II. METHODOLOGY

In order to fetch out experimentations and implementations WEKA was used as the Data mining tool. WEKA is a high quality of data mining tool for the users to classify the accuracy on the basis of dataset by applying different algorithm approaches and compared in the field of bioinformatics. In this paper have used these data mining techniques to predict the heart disease through classification of different algorithms accuracy.

A. *Dataset*

The Heart Disease data set from the UCI Learning Repository is used for this study. The Heart disease data set is split into Training data and Testing Data.

B. *Decision Tree*

Decision Tree models are frequently used in data mining to inspect data and induce the tree and its rules that will be used to make prediction. Decision tree is a classifier in the form of tree structure .A decision tree can be used to classify an example by begin at root of the tree and going through it until a leaf node is reached, which produce the classification of the instance.

1.J48

Decision tree J48 is the implementation of ID3 algorithm developed by the WEKA .WEKA is used as Data mining tool that contributes various algorithms to be applied on data set. The J48 on that dataset would permit to find the target variable of a new data set record and it used to implement Univariate Decision Tree approach.

2.Random Forest

Random Forest is a group of Decision Tree it can be used for classification tasks and that it's uncomplicated to view the relative consequence it assigns to the input features. Random Forest is a predictive modelling tool and it builds multiple Decision trees and consolidate them together to get a more valid and fixed prediction.

C. *OUTLINE OF THE PROPOSED WORK*

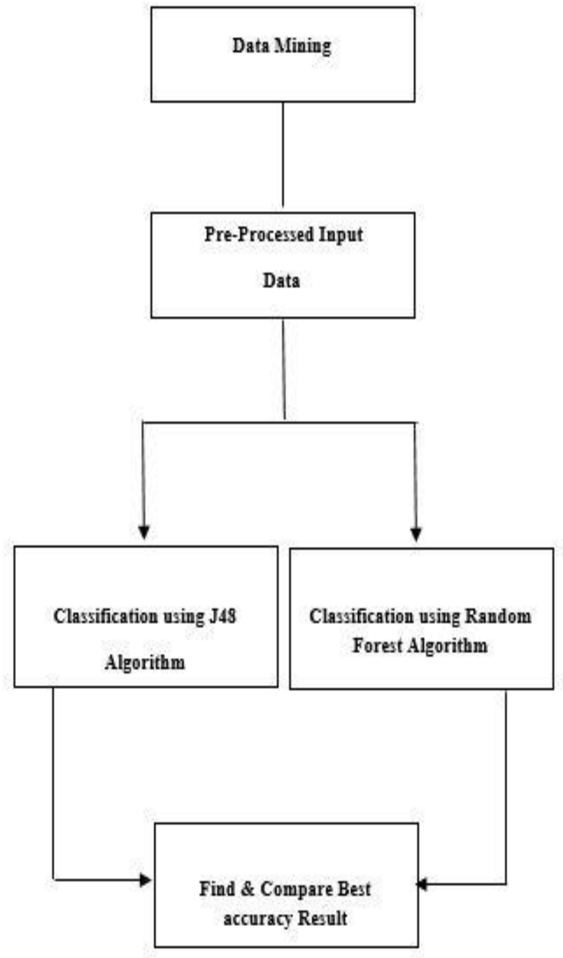
Data Mining
Pre-Processed Input Data
Classification using J48 Algorithm
Classification using Random Forest Algorithm
Find & Compare Best accuracy Result

Fig1.Outline of the proposed work

III. EXPERIMENTAL RESULT

A. *Performance Metrics*

1. True Positive Rate

True Positive also called the sensitivity it used to computes the quantity of actual positives and it also mention to sensitivity or recall.TP Rate is used to assess the rate of actual positive which are absolutely identified.

$$\text{True Positive Rate} = \frac{\mathbf{TP}}{\mathbf{TP+FN}}$$

2. False Positive Rate

A False Positive is an error in some estimation process in which a condition tested for this inaccurately found to have been detected .FP Rate positives contrast with false negatives ,which are results indicating that some condition tested for is absent.

$$\text{False Positive Rate}=\frac{\mathbf{TP}}{\mathbf{FP+TN}}$$

3. Precision

Precision is an essential model estimation metrics it refers to the percentage of outcomes which are applicable.

$$\text{Precision}=\frac{\mathbf{TP}}{\mathbf{TP+FP}}$$

4. Recall

Recall is an essential model estimation metrics it refers to the percentage of overall predicted applicable outcomes rightly classified by algorithm.

$$\text{Recall }=\frac{\mathbf{TP}}{\mathbf{TP+FN}}$$

5. F-Measure

The F-Measure analyse Precision and Recall of the test to calculate the rate. Precision (P) is the amount of True Positive outcome and the Recall(R) is the amount of true positive outcomes should have been return.

$$\text{F-Measure} = 2\left(\frac{\boldsymbol{Precision*Recall}}{\boldsymbol{Precision+Recall}}\right)$$

6. The Mathews Correlation Coefficient

The Mathews Correlation Coefficient (MCC) is used to assess of the binary classification .It returns -1 to +1, +1 specify a refine prediction, 0 specify no superior than random foretell and -1 indicates complete dissent between prediction and examination.

$$\text{MCC}\frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

7. Kappa Statistics

Kappa Statistics is a grade benchmark, largely used to compare the detected accuracy value with the excepted accuracy value, this statistical evaluation can be accomplished on both single classifier and multiple classifier among themselves.

$$\text{k} = \frac{P_O - P_e}{1 - P_e}$$

8. Mean absolute error (MAE)

The MAE is essentially used to ascertain the mean error magnitude of the forecast sets, except its direction. MAE is a linear outcome which means all the independent difference are weighted equivalent in the average.

$$\text{MAE} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} |\mathbf{y_i}-\mathbf{x_i}|}{\mathbf{n}}$$

9. RMSE

RMSE is a quadratic scoring rule ,it is utilized to assess the errors average magnitude ,the related equation for root mean squared error is give in a twosome of reference, either it is to indicate the formula in words or to express the unassociativity between the forecast and observed values

$$\text{RMSE}= \sqrt{\frac{\sum_{\mathbf{i=1}}^{\mathbf{n}}(\hat{\mathbf{y}}-\mathbf{y_i})^{\mathbf{2}}}{\mathbf{n}}}$$

10. Relative Absolute Error

The relative absolute error is identical the relative squared error in the sense that it is also comparatively to uncomplicated predictor, which is just the average of the authentic values. In this case, though, the error is just the inclusive absolute error instead of the total squared error.

$$\text{RAE}=\frac{\sum_{j-1}^{n} |\boldsymbol{P}_{(i)}-\boldsymbol{a}_i|}{\sum_{i=1}^{n} |\bar{\boldsymbol{a}}-\boldsymbol{a}_i|}$$

B. Performance Evaluation

TABLE I

Dataset Classification Based On Its Properties

	J48
Correctly Classified Instances	495
Incorrectly Classified Instance	2
Kappa Statistic	0.9927
Mean absolute error	0.0027
Root Mean squared error	0.0518
Relative absolute error	0.7307%
Root relative squared error	0.7307%
Loss	0.4024%
Accuracy	99.5976%

TABLE II
Dataset Classification Based On Its

	Random Forest
Correctly Classified Instances	496
Incorrectly Classified Instance	1
Kappa Statistic	0.9963
Mean absolute error	0.0226
Root Mean squared error	0.0544
Relative absolute error	6.156%
Root relative squared error	12.7038%
Loss	0.2012%
Accuracy	99.7988%

C. *Predictor Comparision*

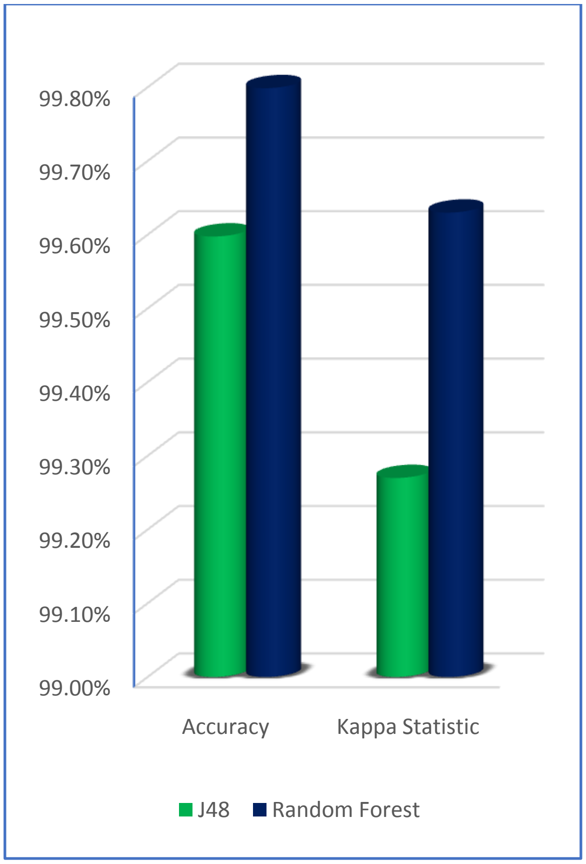

Fig.2 Accuracy of Classification Algorithm(J48 AND Random Forest)

IV. CONCLUSION

In this proposed work have discussed some of functional work techniques that can be used for forecast heart disease classification and the precision of classification techniques is assess found on the selected classifier algorithm. A major challenge in data mining area is to build exact and computationally efficient classifiers for Medical applications. The performance of Random forest shows high level compare with J48 classifier.

ACKNOWLEDGEMENT

I would like to thank to Dr.M.MohamedSathik for providing his valuable inputs to improve this paper.I convey my fruitful thanks to Dr.S.ShajunNisha for their valuable suggestions and support for necessary resources to accomplish this research.

REFERENCES

[1] Theresa Princy.R and J.Thomas, “Human Heart Disease Prediction System using Data Mining Techiques,” International Conference on Circuit ,Power and Computing Technologies[ICCPCT], 2016.

[2] Ritika Chandha and Shubhankar Mayank, “Prediction of heart disease using datamining techniques,” Springer, 2016.

[3] Sujata Joshi and Mydhili k.Nair, “Prediction of Heart Disease Using Classification Based Data Mining Techniques,” Springer, vol. 2, 2015.

[4] K.Gomathi and Dr.Shanmugapriyaa, “Heart Disease Prediction Using Data Mining Classification,” International Journal for Research in Applied Science & Engineering Technology(IJRASET), vol. 4, no. II, February 2016.

[5] Rishabha Saxena, Aakriti Johri, Vikas Deep and Purushottam Sharma, “Heart Disease Presdiction System Using CHC-TSS Evolutionary,KNN,and Decision Tree Classification Algorithm,” Springer, 2019.

[6] Purushottam, Pro.(Dr)Kanak Saxena and Richa Sharma, “Efficiant Heart Disease PredictionSystem,” Elsevier, 2016.

[7] Ajad Patel, Sonali Gandhi, Swetha Shetty and Prof.Bhanu Tekwani, “Heart Disease Prediction Using Data Mining,” International Research Journal of Engineering and Technology(IRJET), vol. 04, no. 01, January 2017.

[8] A.Kathija, Dr.S .Shajun Nisha and Dr.M.Mohamad Sathik, “CLASSIFICATION OF BREAST CANCER DATA USING C4.5 CLASSIFIER ALGORITHM,” International Journal of Recent Engineering Research and Development (IJRERD), vol. 02, no. 02.

[9] K.Gomathi and Dr. D. Shanmuga Priyaa, “Multi Disease Prediction using Data Mining,” International Journal of System and Software Engineering, vol. 4, no. 2, 06 December 2016.

[10] Chitra and V.Seennivasagam, “Review of heart disease prediction system using datamining and hybrid itelligent techniques,” International Journal of Computer Application, vol. 47, 2012.