ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

# A Survey on Predicting the Heart Disease Using Data Mining Methods

[1] Ms. D. Monica Seles, [2] Dr. A. Anitha
[1] PG Scholar, [2] Professor
[1][2] Department of IT, Francis Xavier Engineering College, Tirunelveli.

***Abstract:*-- Due to modern life style, different health diseases are arising day by day. Among that, heart disease makes a human life as very complicated. In Medical field, the prediction of heart disease in early stage has been a challenging one. The main objective of this work is to predict the survival of CHD (Coronary Heart Disease) patients using certain data sets. To overcome ever increasing growth of heart disease, many researchers adopted various kinds of data mining methodologies. Here, the system is designed to discover the condition to find the risk level of patients based on the parameter of symptoms about their health. Several Data Mining Techniques such as Decision Tree (DT), Classification and Regression Tree (CART), Sequential Minimal Optimization (SMO) and Support Vector Machine (SVM) are available to predict heart diseases. The above techniques are employed to predict risk level of patients and provide accuracy as SVM with 84.7%, CART with 85.4%, and SMO with 84.07% and Decision Tree with 89%. Finally, the result showed that DT has a huge potential in predicting risk level of heart disease more accurately.**

## 1. INTRODUCTION

The Healthcare industry has enormous amount of medical data sets which is not mined [3]. This is hidden information which is useful in data analysis to detect the heart disease. Coronary Heart Disease (CHD) is one of the highest flying diseases that increase the rate of morbidity and mortality in this modern world. The heart is an important organ for every human being which helps in the process of blood circulation and makes the human body active. If the heart stops working, the life of living individual will also stop and leads to death. The risk level of patients can be identified using certain datasets involving diabetes, cholesterol, age, family history, obesity, smoking, alcohol consumption, physical inactivity, poor diet and chest pain type. There are many types of heart disease such as coronary artery disease, cardiac arrest, congestive heart failure, Arrhythmia, Peripheral artery disease, Stroke, Congenital heart disease are the main reasons of death for large number of people all over the globe[4].

Data mining is the technique of pulling out knowledgeable information from large amount of data sets those outcomes in anticipating or portraying the information utilizing procedures such as classification, clustering, and association and so on. It plays a vital role in biomedical field in predicting various diseases. Due to large changes in the surroundings, an unhealthy eating habit of person leads to suffer from different types of diseases of same category so physicians may not be able to find right disease. In this, data mining concepts along with use of intelligent algorithms which helps in predicting disease having multiple inputs and provide a way for physicians to tackle with such kind of problems. Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes and Association helps to discover the probability of the co-occurrence of items in a collection. Here, Decision Tree, Artificial Neural Network, Support Vector Machine are available in the prediction process. The Decision tree learning is a method commonly used in data mining; the goal is to create a model that predicts the value of a target variable based on several medical reports. The Neural Network gives the minimized error in prediction of heart disease and SVM are supervised learning models with associated learning algorithms that analyze data used for classification and analysis. This paper proposes a survey on predicting the heart disease using three different algorithms to provide efficient as well as accuracy.

## 2. LITERATURE SURVEY

### *2.1 Decision Tree*

Decision Tree is one of the important classifier algorithms which do not need any parameter or domain knowledge [1].

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

This classification technique can construct the tree recursively which based on certain variables. It can handle high dimensional categorical data and result of this algorithm is easy to read and interpret. Depending on some mathematical considerations, the branches of trees can be constructed and also specify the corresponding threshold value. The objective of this algorithm is to find a variable-threshold pair which maximizes the homogeneity in the result.

This research focuses on health care department which contain rich information and poor knowledge [5]. It can use the dataset from Cleveland Heart disease database of 909records with 15 kinds of attribute. Then the records were partitioned into two parts as training dataset of 455 records and testing set of 454 records. There are three models such as Naive Bayes, Decision Tree and Artificial Neural Network (ANN) used to analysis the patients profile as age, sex, blood pressure, blood sugar which can predict the likelihood of patients.

| Techniques | Accuracy (%) |
|---|---|
| Naïve Bayes | 86.53 |
| **Decision Tree** | **89** |
| ANN | 85.53 |

Compare with other two models, decision tree appear as effective in the prediction of heart disease with accuracy of 89%.

### *2.2Classification and Regression Tree*

CART stands for Classification and Regression Tree which involves in two stages such as tree growing and tree pruning [2]. Here, the tree can be grown by selecting among all the possible splits that generate the purer child nodes where purest child node is containing elements of only one class. This paper aimed to develop a classifier to validate the risk level of patients who are suffering from congestive heart failure (CHF).Using standard long term heart rate variability measures (HRV), this classifier separates the low risk patients from high risk patients.

Based on the classification of New York Association, records of patients are labeled as higher or lower risk. The continuous analysis of two public Holter databases, it shows that 12 patients are suffering from mild CHF which labeled as lower risk (NYHA classes I and II) and 34 of them suffering from severe CHF which labeled as higher risk (NYHA classes of III and IV). The patients described under section II with satisfactory signal quality are selected as eligible one. The data can be taken from Congestive Heart Failure RR Interval Database and BIDMC Congestive Heart Failure Database. The records of 250 samples per second were sampled and annotated automatically. There is no significant difference between age limit and gender appeared in both databases.

Here Classification and Regression Tree (CART) is involved to grown up the automatic classifier. Every patient's risk can be classified with the fraction of normal to normal interval (NN) to the total heartbeats interval (RR).According to (NN/RR), above 80% were selected as eligible in order to have satisfactory signal quality. This paper performs the number of classification techniques such as CART, RF (Random Forest) and C4.5 to calculate the Accuracy.

| Classifier | Accuracy (%) |
|---|---|
| **CART** | **85.4** |
| RF | 82.7 |
| C4.5 | 84.6 |

Among the above classifier, CART attains a highest accuracy with 85.4% in the observation of these databases.

### *2.3Support Vector Machine*

Support Vector Machine (SVM) is one of the classification algorithm based on statistical learning theory which used to classify data [7]. SVM classifies both linear and non-linear data and it performs this task by maximizing the margin separating both linear and non-linear data while minimizing the errors appeared in classification method. Here, SVM uses sequential minimal optimization algorithm. SVM involve performing the classification tasks by maximizing the margin of hyper-plane while minimizing the classification errors. The main thing of this paper is to break the use of Artificial Intelligence devices in order and expectation of heart illness. The classification of Coronary Heart Disease (CHD) can be valuable in medical field for finding exact result quickly.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

The dataset of Cleveland Heart Database and Statlog Database are taken from UCI Machine learning dataset repository. It can refer 14 attributes include age, gender, chest pain, Trestbps (resting pulse), cholesterol, blood sugar in fasting, resting ECG value, Thalach-maximum heart rate achieved, Exang-exercise induced angina, old peak rate, slope value, CA, thal and the last attribute is a class. The Statlog database has two classes 1 and 2 to predict attribute as well as Cleveland Database has multiple classes from value of 0 to 4. In these databases, the value 0 and 1 specifies no disease and other shows the presence of heart disease risk.

| Classifier | Accuracy (%) |
|---|---|
| **SVM** | **84.7** |
| ANN | 81.8 |

To analyze the variability of two class and multiclass problems in these dataset, the technique such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) are employed here [6]. The execution of these techniques shows that time taken for analysis and classification in SVM is more viable than ANN.

***2.4Sequential Minimal Optimization***

Sequential Minimal Optimization (SMO) can use to split the training set into smaller sets that can solve the problem of support vector machine analytically [8]. SMO becomes slow while solving the problem which instances are not separate in linear manner. It can work with enormous amount of dataset as it doesn't need any extra storage. This method is only adoptable for linearly separable problems. This paper is performed under test mode of k-fold cross–validation with supplied test sets where k=10. The standard heart disease data sets are taken from UCI Machine Learning repository which has 270 records of patients with heart disease or without heart disease. These records include some 13 parameters such as age, sex, chest pain, serum cholesterol, fasting blood sugar, resting blood sugar, resting electrocardiographic results, maximum heart rate, exercise induced angina, old peak ST depression induced by exercise, slope of the peak exercise, number of major vessels colored by fluoroscopy and thal. The record of 100 patients are also gathered from Enam Medical Diagnosis Centre, Savar, Dhaka, Bangladesh which has 6 attributes such as age, sex, chest pain, high blood pressure, diabetics and maximal heart rate. The methods employed here in classification process are SMO, Bayes Net, MLP (Multilayer Perceptron), K Star, J48.

| Classifier | Accuracy in collected data set (%) | Accuracy in standard data set (%) |
|---|---|---|
| **SMO** | **89** | **84.0741** |
| K star | 75 | 75.1852 |
| J48 | 86 | 76.6667 |
| Bayes Net | 87 | 81.1111 |
| MLP | 86 | 77.4074 |

The out of all such classifier, SMO obtain high accuracy in both collected and standard data set with percentage of 89 and 84.07 respectively.

***2.5 Heart Disease Prediction System***

The main contribution of this paper is to help non-specialized doctors to take correct decision about heart disease risk level [9][10]. It contain heart disease database with screening clinical information of heart disease and data can be gathered from Long Beach and Cleveland clinic foundation. The input attributes are age, sex, chest pain, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, slope of the peak exercise ST segment, number of major vessels colour by fluoroscopy and thal.

| Techniques | Accuracy (%) |
|---|---|
| SVM | 70.59 |
| C4.5 | 73.53 |
| MLP | 74.85 |
| **Efficient Heart Disease Prediction** | **86.7** |

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

From analyzing of above algorithm, Efficient Heart Disease Prediction shows highest accuracy as 86.7%.

### 3.SUMMARY OF LITERATURE SURVEY

| Technique | Accuracy (%) |
|---|---|
| Decision Tree | 89 |
| CART | 85.4 |
| SVM | 84.7 |
| SMO | 84.07 |
| Efficient Heart Disease Prediction | 86.7 |

### 4. CONCLUSION

In this paper, a survey can be taken from the year of decade paper to analyze the heart disease prediction. The different data mining techniques are used here to find the accuracy of risk level in heart disease. These methodologies can use number of dataset or attributes from various databases analyzed by the author. The future work focuses on dataset with small number of attributes and predicts high the accuracy level using some other data mining algorithms.

### REFERENCES

[1] Jyoti Soni, Ujma Ansari and Dipesh Sharma, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Application, ISSN: 0975-8887, Vol 17, No.8, March 2011.

[2] Paolo Melillo, Nicola De Luca, Marcello Bracale and Leandro Pecchia, "Classification tree for Risk Assessment in Patients Suffering From Congestive Heart Failure via Long-Term Heart Rate Variability", IEEE Journal of Biomedical and Health Informatics, Vol 17, No.3, pp.727, May 2013.

[3] Hlaudi Daniel Masethe, Mosima Anna Masethe, "Prediction of Heart Disease usin Classification Algorithm", Proceedings of the World Congress on Engineering and Computer Sciences , Vol II, pp.22-24, October 2014.

[4] Sana Bharti, Shaliendra Narayan Singh, "Analytical Study of heart Disease Prediction Comparing With Different Algorithms", International Conference on Computing, Communication and Automation (ICCCA2015), ISBN: 978-1-4799-88990-8/15, IEEE, 2015.

[5] Purushottam, Kanak Saxena, Richa Sharma , "Efficient Heart Disease Prediction System using Decision Tree", International Conference on Computing, Communication and Automation (ICCCA2015), ISBN: 978-1-4799-88990-8/15, IEEE, 2015.

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

[7] Radhimeenakshi. S, "Classification and Prediction of heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network", 2016 International Conference on Computing for Sustainable Global Development (INDIA Com), 2016.

[8] Marjia Sultana, Afrin Haider and Mohammed Shorif Uddin, "Analysis of Data Mining Techniques for Heart Disease Prediction", iCEEiCT2016, 2016.

[9] Purushottam, Kanak Saxena, Richa Sharma, "Efficient Heart Disease Prediction System",Procedia Computer Science, 2016.

[10] Salma Banu, Suma Swamy, "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey" 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016.