*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 3 Issue 3 ǁ March. 2015 ǁ PP.74-77*

# Heart Disease Prediction Using Data Mining Techniques

Andrea D’Souza
[1]*(Information Technology, Padre Conçeicao College of Engineering, Goa, India)*

**ABSTRACT**: There are huge amounts of data in the medical industry which is not processed properly and hence cannot be used effectively in making decisions. We can use data mining techniques to mine these patterns and relationships. This research has developed a prototype Heart Disease Prediction using data mining techniques, namely Neural Network, K-Means Clustering and Frequent Item Set Generation. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease to be established. Performance of these techniques is compared through sensitivity, specificity and accuracy. It has been observed that Artificial Neural Networks outperform K Means clustering in all the parameters i.e. Sensitivity, Specificity and Accuracy.

**Keywords**- *Apriori, Heart Disease, Data Mining, K-Means, Neural Network*

## I. INTRODUCTION

Cardiovascular diseases are one of the most common diseases of the modern world.There are certain things that increase a person’s chances of getting cardiovascular disease. Cardiovascular disease (CVD) refers to any condition that affects the heart. Many CVD patients have symptoms such as chest pain (angina) and fatigue, which occur when the heart isn’t receiving adequate oxygen. As per a survey nearly 50 percent of patients, however, have no symptoms until a heart attack occurs. A number of factors have been shown to increase the risk of developing CVD. Some of these are family history of cardiovascular disease, High levels of LDL (bad) cholesterol, Low level of HDL (good) cholesterol, Hypertension, High fat diet, Lack of regular exercise and Obesity. With so many factors to analyze for a diagnosis of cardiovascular disease, physicians generally make a diagnosis by evaluating a patient’s current test results. Previous diagnoses made on other patients with the same results are also examined by physicians. These complex procedures are not easy. Therefore, a physician must be experienced and highly skilled to diagnose cardiovascular disease in a patient.Data mining has been heavily used in the medical field, to include patient diagnosis records to help identify best practices. The difficulties posed by prediction problems have resulted in a variety of problem-solving techniques. For example, data mining methods comprise Artificial Neural Networks and Clustering Techniques (K-Means Clustering). It is difficult, however, to compare the accuracy of the techniques and determine the best one because their performance is data dependent. A few studies have compared data mining and statistical approaches to solve prediction problems. The comparison studies have mainly considered a specific data set.

## II. LITERATURE REVIEW

Up to now, several studies have been reported that have focused on cardiovascular disease diagnosis. These studies have applied different approaches to the given problem and achieved high classification accuracies of 77% or higher.

Robert Detrano’s experimental results showed correct classification accuracy of approximately 77% with logistic regression derived discriminant function.

Colombet et al. evaluated implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database.

Imran Kurt , MevlutTure , A. TurhanKurum compare performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease.

The John Gennari’s CLASSIT conceptual clustering system achieved a 78.9% accuracy on the Cleveland database

## III. METHODOLOGY

This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts for predicting heart diseases accurately. The Artificial Neural Network, K Means Clustering Algorithm and Frequent Item Set generation using Apriori Techniques are used to classify whether a patient suffers from heart disease or not. A study is performed on the techniques to find the most accurate one.

**Data Mining Review**

Data mining techniques analyze data and perform learning to extract hidden patterns and relationships from large databases.

Artificial Neural Network(ANN)the neural network approach is used for analyzing the heart disease dataset. Applying feed forward neural network model and back propagation learning.the heart disease database are trained by the neural network. The input layer contains 13 neurons to represent 13 attributes. There is one output class variable that takes the value 1 or 0 depending on whether a patient suffering from heart disease or not respectively.

K- Means Clustering is one of the widely recognized clustering tools that are applied in a variety of scientific and industrial applications. K-means groups the data in accordance with their characteristic values into K distinct clusters. Data categorized into the same cluster have identical feature values. K, the positive integer denoting the number of clusters, needs to be provided inadvance.Since theCleveland heart database contains different types of variables: symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio the dissimilarity between the objects is calculated by

$$d(i,j)=\frac{\sum_{f=1}^{p}\delta_{ij}^{(f)}d_{ij}^{(f)}}{\sum_{f=1}^{p}\delta_{ij}^{(f)}} \quad (1)$$

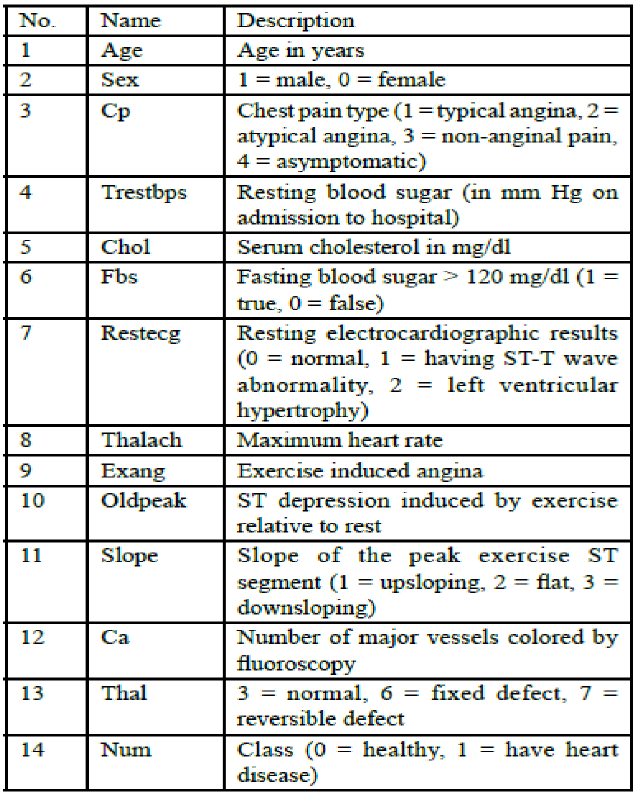
f is binary or nominal: $d_{ij}(f) = 0$ if $x_{if} = x_{jf}$ , or else $d_{ij}(f) = 1$
f is interval-based: use the normalized distance
f is ordinal or ratio-scaled: compute ranks $r_{if}$ and treat $z_{if}$ as interval-scaled

Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions. Other algorithms are designed for finding association rules in data having no transactions.As is common in association rule mining, given a set of item sets, the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

**Data Sources**

A total of 297 records with 14 medical attributes (factors) were obtained from the Cleveland Heart Disease database. Table 1 lists the attributes. The records were split into two datasets: training dataset (200 records) and testing dataset (97 records). To avoid bias, the records for each set were selected randomly. The attribute “Num” was identified as the predictable attribute with value “1” for patients with heart disease and value “0” for patients with no heart disease.

Table 1: Attributes of Cardiovascular dataset



| No. | Name | Description |
|---|---|---|
| 1 | Age | Age in years |
| 2 | Sex | 1 = male, 0 = female |
| 3 | Cp | Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic) |
| 4 | Trestbps | Resting blood sugar (in mm Hg on admission to hospital) |
| 5 | Chol | Serum cholesterol in mg/dl |
| 6 | Fbs | Fasting blood sugar > 120 mg/dl (1 = true, 0 = false) |
| 7 | Restecg | Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy) |
| 8 | Thalach | Maximum heart rate |
| 9 | Exang | Exercise induced angina |
| 10 | Oldpeak | ST depression induced by exercise relative to rest |
| 11 | Slope | Slope of the peak exercise ST segment (1 = upsloping, 2 = flat, 3 = downsloping) |
| 12 | Ca | Number of major vessels colored by fluoroscopy |
| 13 | Thal | 3 = normal, 6 = fixed defect, 7 = reversible defect |
| 14 | Num | Class (0 = healthy, 1 = have heart disease) |

**Results and Findings**

Initially dataset had 14 attributes and 303 records. During data pre processing missing value records were identified and deleted from dataset. After deleting records with missing values we were left with 297 records. . On these 297 records data mining techniques Artificial Neural Networks (ANNs), K Means Clustering and Frequent Item Set Generation using Apriori Algorithm were applied.

A distinguished confusion matrix was obtained to calculate sensitivity, specificity and accuracy. Confusion matrix is a matrix representation of the classification results. Table 2 shows confusion matrix.

The upper left cell denotes the number of samples classifies as true while they were true (i.e., TP), and the lower right cell denotes the number of samples classified as false while they were actually false (i.e., TN). The other two cells (lower left cell and upper right cell) denote the number of samples misclassified. Specifically the upper right cell denoting the number of samples classified as false while they actually were true (i.e., FN), and the lower left cell denoting the number of samples classified as true while they actually were false (i.e., FP).
Below formulae were used to calculate sensitivity, specificity and accuracy:
Sensitivity = TP / (TP + FN)
Specificity = TN / (TN + FP)
Accuracy = (TP + TN) / (TP + FP + TN + FN)

Table 2: Confusion matrix

| | Classified as healthy | Classified as not healthy |
|---|---|---|
| Actual healthy | TP | FN |
| Actual not healthy | FP | TN |

The configuration used for the neural network architecture is
Error tolerance = 0.1
Learning rate = 0.09Maximum number of cycles = 20Number of layers = 3
Number of Neurons in input layer = 13
Number of Neurons in hidden layer = 9
Number of Neurons in output layer = 1

The configuration of the K Means Clustering algorithm is K=2(heart disease or no heart disease)

Table 3: Sensitivity, Specificity and Accuracy of ANN and K Means

| | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| ANN | 68% | 83.33% | 79.38% |
| K Means Clustering | 54.0145% | 71.25% | 63.299% |

The Apriori algorithm was analysed for different values of minimum support count. The most frequent itemsets generated are given below:
Case 1: minimum support count=15
{age>40 AND age<60, male, chest pain type 4, trestBps> 90, chol>200 AND chol<400, fbs NOT > 120, restecg=probable, thalach< 150 , oldPeak=1 ,diagnosis= 1} Support: 15
Case 2: minimum support count= 10
{ male, chest pain type 4, trestBps> 90, chol>200 AND chol<400, fbs NOT > 120, restecg=probable, thalach< 150 , oldPeak=1 , exang<1.5, diagnosis= 1} Support: 10

## IV. CONCLUSION

Through this work, three techniques to classify patients as having heart disease or not are compared. The techniques used are Artificial Neural Network, K Means Clustering and Frequent ItemSet Generation Using Apriori Algorithm. The techniques were compared on the basis of Sensitivity, Specificity and Accuracy.The results show that Artificial Neural Networks outperform K Means Clustering in all the parameters i.e. Sensitivity, Specificity and Accuracy. The frequent itemset patterns that The limitations of this project are that data sets are not available. The available data sets are not updated and do not contain parameters that are found to be relevant through latest research.The proposed work can be improved and expanded for building a Heart disease prediction system. Real and complete data from Health care organizations and agencies needs to be collected.

## REFERENCES

**Journal Papers:**

[1] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "*International application of a new probability algorithm for the diagnosis of coronary artery disease*". *American Journal of Cardiology,Vol. 64, No. 3, 1987, pp. 304-310.*

[2] Colombet, I.; Ruelland, A.; Chatellier, G.; Gueyffier, F. (2000). "*Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression*"*Proceedings of AMIA Symp 2000, p 156-160.*

[3] Kurt, I.; Ture, M.; Turhan, A., "*Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease*". *Journal of Expert Systems with Application, Vol. 3, 2008, pp. 366-374.*

[4] Gennari, J., "*Models of incremental concept* formation". *Journal of Artificial Intelligence, Vol. 1, 1989, pp. 11-61.*

[5] SellappanPalaniappan, RafiahAwang, "*Intelligent Heart Disease Prediction System Using Data Mining Techniques*", *IJCSNS International Journal of Computer Science and Network Security, Vol.8 No.8, August 2008*

[6] AnchanaKhemphila, VeeraBoonjing ,*Comparing performances of logistic regression*, decision trees , and neural networks for classifying heart disease patients

[7] ShantakumarB.Patil, Y.S.Kumaraswamy, "*Intelligent and Effective Heart At-tack Prediction System Using Data Mining and Artificial Neural Network*" *European Journal of Scientific Research ISSN 1450-216X Vol.31 No.4 (2009), pp.642-656*

[8] N.Deepika* et al. " *Association rule for classification of Heart attack patients*" (IJAEST) International Journal Of Advanced Engineering Sciences And Technologies. Vol No. 11, Issue No. 2, 253 – 257

**Books:**

[1] Simon Haykin, *Neural Networks: A Comprehensive Foundation* (China machine press, Beijing, 2004).
.