INDEX

INTRODUCTION

Heart disease is the leading cause, which has accounted for most number of deaths in an annual year. Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Among the most serious heart diseases is " Heart Attack ", also known as Myocardial Infarction. A heart attack occurs when blood flow to the heart is blocked due to a clogged artery. This blockage prevents oxygen-rich blood from reaching a section of the heart. If not reopened quickly, the section of the heart not receiving blood will begin to die. A heart attack occurs every 20 seconds over the world. An estimated 17.5 million deaths occur due to cardiovascular diseases worldwide. Because of the sudden onset and damage that occurs quickly from a heart attack, one has to act fast. Around half of the heart attack deaths occur before the patient reaches the hospital. Thus, it is imperative to be aware of the upcoming danger . Therefore, a prior prediction of heart attack would be a huge step to save one's life.

Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients. Hence by implementing a heart disease prediction system using Data Mining techniques and performing data mining on various heart disease attributes, one can predict more probabilistically that the patients will be diagnosed with heart disease.

Data mining is the computer based process of extracting useful information from enormous sets of databases. Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature. These data need to be collected in an organized form. This collected data can be then integrated to form an medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the data. Hence, this project uses machine learning techniques to predict the possibility of heart attack by using the patient's records.

LITERATURE REVIEW

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. There are around thirty five research papers that explore the computational methods to predict heart diseases. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies on multiple databases of patients from around the world.

Marjia Sultana, Afrin Haider and Mohammad ShorifUddin have illustrated about how the datasets available for heart disease are generally a raw in nature which is highly redundant and inconsistent. There is a need of pre-processing of these data sets; in this phase high dimensional data set is reduced to low data set.

In year 2006, Carlos Ordonez has studied association rule mining with the train and test concept on a dataset for heart disease prediction. Association rule mining has a disadvantage that it produces extremely large number of rules most of which are medically irrelevant.

In year 2008, Sairabi H. Mujawar predicted heart disease using modified k-means and Naïve Bayes. Diagnosis of heart disease is a complex task and requires great skills. The dataset is obtained from Cleveland Heart Disease Database. The attribute "Disease" with a value '1' indicates the presence of heart disease and a value '0' indicates the absence of heart disease.

In year 2010, Mrudula Gudadhe presented a decision support system for heart disease classification. Support vector machine (SVM) and artificial neural network (ANN) were the two main methods used in this system. A multilayer perceptron neural network (MLPNN) with three layers was employed to develop a decision support system for the diagnosis of heart disease. This multilayer perceptron neural network was trained by back-propagation algorithm which is computationally an efficient method. Results showed that a MLPNN with back-propagation technique can be successfully used for diagnosing heart disease.

In year 2013, S. Vijiyaraniet performed a work, An Efficient Classification Tree Technique for Heart Disease Prediction. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm.The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

In year 2014, K. Sudhakar studied heart disease prediction using data mining. The data generated by the healthcare industry is huge and "information rich". As such, it cannot be

interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analyzed on heart disease database. Classification techniques such as Decision tree, Naïve Bayes and neural network are applied here.

In year 2016, Ashwini Shetty proposed different data mining approaches for predicting heart disease. Their research work analyses the neural network and genetic algorithm to predict heart diseases. The system calculates accuracy using MATLAB. Preprocessing is done using WEKA. The results show that the hybrid system of genetic algorithm and neural network works much better than the performance of neural network alone.

AIM OF THE PROJECT

The main goal is to create a system by applying different data mining Techniques to help medicinal services experts with progressed exactness in the judgement of heart disease.

In this project, we will be using machine learning and data mining technologies to analyze the data of different patients around the world and will try to predict the possibility of occurrence of a heart attack based on various parameters , few of which are cholesterol levels, pulse rate, heart rate, age, gender, blood pressure, smoking history and others with the highest accuracy possible.

Neural network, Naive Bayes, Decision Tree etc. are some techniques used in the diagnosis of heart disease.Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. Hence, we will be using decision tree technique in our system.

METHODOLOGY USED

The following objectives are set for this heart prediction system.

• The prediction system should not assume any prior knowledge about the patient records it is comparing.

• The chosen system must be scalable to run against large database with thousands of data

Records set with medical attributes will be obtained. With the help of the dataset, the patterns significant to the heart attack diagnosis are extracted.

Decision Tree (one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes) approach will be used here. It will be able to distinguish the dominant attributes and provides different labels of LIKELY PRESENCE for heart disease. The general approach took after for Decision Tree classification for satisfying the objective is:

Training => Algorithm => Model => Testing => Evaluation

Also , a software will be used which contains apparatuses for information pre-processing, classification, regression, clustering, association rules, and visualization.

The working of this system is described in a step by step:

1. Dataset collection which contains patient details.
2. Attributes selection process selects the useful attributes for the prediction of heart disease.
3. After identifying the available data resources, they are further selected, cleaned, made into the desired form.
4. Different classification techniques as stated will be applied on preprocessed data to predict the accuracy of heart disease.

SELECTED HEART DISEASE ATTRIBUTES

Name	Type
Age	Continuous
Sex	Discrete
Cp	Discrete
Trestbps	Continuous
cholestral(mg/dl)	Continuos
Thalach(max heart rate achieved)	Continuos

RESOURCES

Software

OS : Microsoft Windows 8 or above

Matlab 2013a / Matlab 2015a

Spark SQL

Spark MLLib

Spark Apache

HDFS

Hardware

Processor: Intel Core i3 CPU or above

Memory : 4GB Ram (For fast processing of MATLAB)

SIGNIFICANCE OF THE PROJECT

There has recently been significant effort to alleviate doctors' workload and improve the overall efficiency of the health care system with the help of machine learning.

Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. However using data mining technique can reduce the number of test that are required. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique.

Learning of the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease. Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world’s biggest killer demands higher accuracy. Our project seeks to improve diagnosis accuracy to improve health outcomes.

TIMELINE

OCT 2017-NOV 2017: ANALYSIS
DEC 2017 : DESIGN
JAN 2018-FEB 2018 : CODING
MARCH 2018: TESTING
By the end of April 2018 , the project will be ready.

CONCLUSION

Heart attack is crucial health problem in human society. We introduced the heart disease prediction system using decision tree technique for the prediction of heart disease. The results will guide providers, healthcare organizations, nurses, and other treatment providers in using new

Effective and tailored medical treatment can be developed using these technologies.
As identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more
complex models to increase the accuracy of predicting the early onset of heart attack.
In future , our aim is to carry forward the work of temporal medical dataset , where dataset varies with time and retraining of dataset is required. Also, one can perform additional experiments with more dataset and algorithms to improve the accuracy and to build a model that can predict specific heart disease types.