Prediction and Analysis of Heart Disease Using Data Mining Techniques

Anusha N.B[1], Chaitra K[2], Chandana H.M[3], Kiran G[4], Swathi D.V [5]

Anusha N.B, Dept. of Computer Science Engineering, Bellary Institute of Technology and Management, Bellary, India

Chaitra K, Dept. of Computer Science Engineering, Bellary Institute of Technology and Management, Bellary, India

Chandana H.M, Dept. of Computer Science Engineering, Bellary Institute of Technology and Management, Bellary, India

Kiran G, Dept. of Computer Science Engineering, Bellary Institute of Technology and Management, Bellary, India

Swathi D.V, Asst. Professor, Dept of Computer Science Engineering, Bellary Institute of Technology and Management, Bellary, India

---***---

Abstract – Heart is the next major organ comparing to brain which has more priority in human body. It pumps the blood and supplies to all organs of the whole body. Prediction of occurrences of heart diseases in medical field is significant work. Data Analytics is useful for prediction from more information and it helps medical center to predict of various disease. The stored data can be useful for source of predicting the machine learning techniques are used to predict the heart disease such as K-Nearest Neighbor(KNN),Naïve Bayes, Support Vector Machine(SVM),Logistic Regression and weighed KNN.

Keywords*:* Type of Heart Disease, Data mining techniques, Data Mining, Principle Component Analysis(PCA)algorithm.

1. INTRODUCTION

Heart disease is a general name for a variety of diseases, conditions and disorders that affect the heart and the blood vessels. It is most-flying disease of modern world. Symptoms of heart disease vary depending on the specific type of heart disease. The types of heart disease such as Congenital heart disease, Congestive heart failure, Coronary heart disease etc., The cost of treatment for heart disease is very expensive in later stages. The treatment cost is not affordable for everyone therefore people are reluctant to take proper treatment. Prediction and analysis of the heart disease through the normal techniques is time consuming. So by using data mining techniques we can detect the type of heart disease at early stages. Data mining is the process of analyzing large set data and summarizing into useful information. Some of the data mining techniques/algorithms are KNN, SVM, Naïve Bayes, Weighed KNN, Logistic Regression. By using these techniques the prediction can made simple by various characteristic to find out whether the person is having a heart disease or not using data mining techniques. According to recent survey the major reason for deat16h is due to heart attack, one in every four death is due to heart attack. Many technologies have come up effectively to discover the reason for heart attack and predict it before it occurs. In this paper we have made an attempt to analyze those various techniques used to predict the heart disease. It is a world known fact that heart is the most essential organ in human body if that organ gets affected then it also affects the other vital parts of the body. Therefore it is very important for people to go for a heart disease diagnosis periodically. Poor clinical decisions can lead to tragic consequences which are therefore unacceptable. They can achieve the best results by employing appropriate computer-based information and/or decision support systems. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining and machine learning techniques can help remedy this situation. The basic flow for predicting heart disease is as follow, The patient database will be collected, which will majorly contain few attributes/parameters. The few attributed are age, sex, Cp, trestbps, Chol, Fbs, restecg, thalach, exang, oldpeak, slope, Ca, thal.

2. PROBLEM STATEMENT

To predict and analyze the type of heart disease for the persons suffering from heart disease using data mining techniques such as K-Nearest Neighbor (KNN),Naïve Bayes, Support Vector Machine(SVM),Logistic Regression, and Weighed KNN.

3. PROPOSED SYSTEM

In this study we have used an R studio rattle to perform Heart Disease classification of the Cleveland UCI repository. It provides an easy-to-use visual representation of the dataset, working environment and building the predictive analytics. Machine learning process starts from a pre-processing data phase followed by feature selection based on DT entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes.

Table1: Table shows different data mining techniques used in the diagnosis of Heart disease.

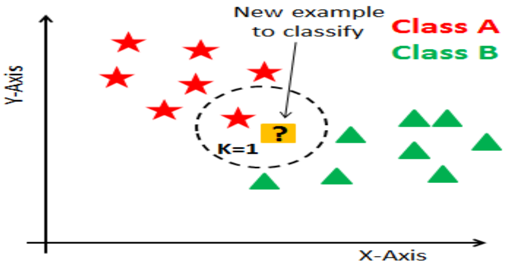
Year	Author	Purpose	Techniques used	Accuracy
2015	Sairabi H.Mujawar et al	Prediction of heart disease using modified K-means and by using Naïve Bayes	Modified K-means algorithm, Naïve Bayes Algorithm	Naive bayes has better accuracy, Heart disease detection = 93%
2016	Ashok kumar Dwivedi et al	Evaluate the performance of data mining techniques for heart disease prediction	Naïve Bayes KNN Logistic Regression	83% 80% 85%
2016	S. Seema et al	Chronic Disease prediction by Mining the data	Naïve Bayes, SVM SVM, naïve bayes	Highest accuracy achieved by naïve bayes, In case of Heart Disease 75.56%. Highest accuracy of 73.58% achieved by SVM in case of Diabetes

4. DATA MINING TECHNIQUES USED :-

K-Nearest Neighbor Algorithm:

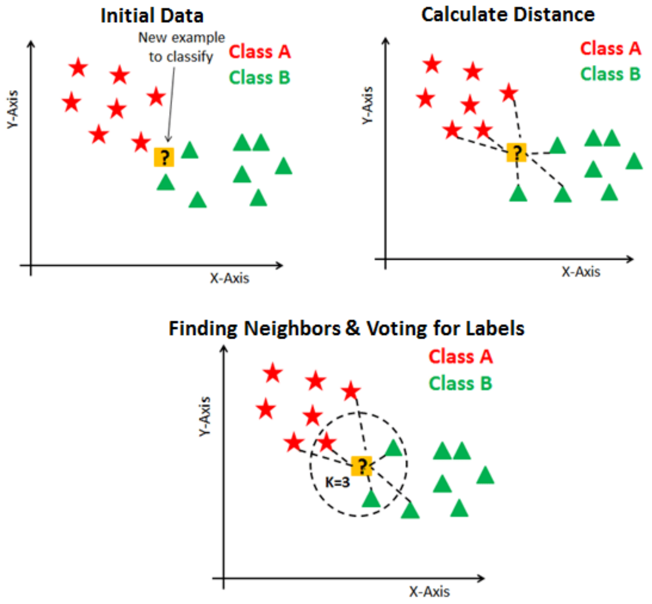
In K-NN, K is the number of nearest neighbors. The number of neighbors is the core deciding factor**.** K is generally an odd number if the number of classes is 2. When K=1, then the algorithm is known as the nearest neighbor algorithm. This is the simplest case.

In the below figure, suppose yellow colored "**?**" let's say P is the point, for which label needs to predict. First, you find the one closest point to P and then the label of the nearest point assigned to P.

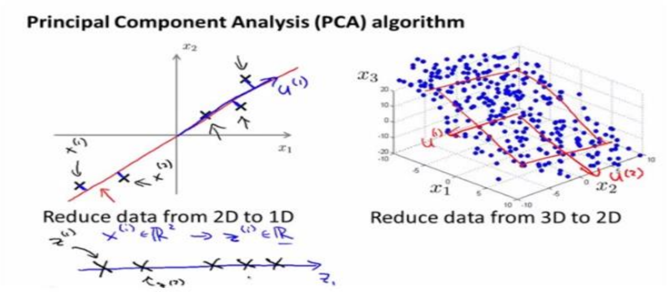


Second, you find the k closest point to P and then classify points by majority vote of its K neighbors. Each object votes for their class and the class with the most votes is taken as the prediction. For finding closest similar points, we find the distance between points using distance measures such as Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance.

The algorithm has the following basic steps:
1.Calculate the distance
2.Find closest neighbors
3.Vote for labels



Principle Component Analysis Algorithm: A principal component can be defined as a linear combination of optimally-weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables.


Principal Component Analysis (PCA) algorithm
Reduce data from 2D to 1D
Reduce data from 3D to 2D

Support Vector Machine Algorithm:

SVM (Support Vector Machine) is a supervised machine learning algorithm that is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyperplane, which acts like a decision boundary between the various classes.

SVM can be used to generate multiple separating hyperplanes such that the data is divided into segments and each segment contains only one kind of data.

Weighed KNN Algorithm:

Weighted KNN is a modified version of KNN. One of the many issues that affect the performance of the KNN algorithm is the choice of the hyperparameter k. If k is too small, the algorithm would be more sensitive to outliers. If k is too large, then the neighborhood may include too many points from other classes.

Naïve Bayes Classifer:

Implementation of Naïve Bayes:

A Naive Bayes' classifier may be a term addressing a simple probabilistic classification supported applying Bayes' theorem. In easy terms, a Naïve Bayes classifier assumes that the presence (or absence) of a specific feature of a category is unrelated to the presence (or absence) of the other feature. As an example, a fruit could also be thought of to be an apple if it's red, round, and regarding 4" in diameter. Even supposing these options rely on the existence of the opposite options, a Naive Bayes' classifier considers all of those properties to independently contribute to the likelihood that this fruit is an apple. Naive Bayes algorithm is based on Bayesian Theorem.

Bayesian theorem Given training data X, posterior probability of a hypothesis H, P(H|X), follows the Bayes theorem P(H|X)=P(X|H)P(H)/P(X)

Algorithm

The Naive Bayes algorithm is based on Bayesian theorem as given by equation

Steps in algorithm are as follows:

1. Each data sample is represented by an n dimensional feature vector, X = (x1, x2….. xn), depicting n measurements made on the sample from n attributes, respectively A1, A2, An.

2. Suppose that there are m classes, C1, C2……Cm. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned if and only if:

P(Ci/X)>P(Cj/X) for all 1< = j< = m and j! = i

Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem,

3. As P(X) is constant for all classes, only P(X|Ci) P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e. P(C1) = P(C2) =…. = P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci) P (Ci).

Logistic Regression:

Logistic regression is a one of the machine learning classification algorithm for analyzing a dataset in which there are one or more independent variables (IVs) that determine an outcome and also categorical dependent variable (DV). Linear regression uses output in continuous numeric whereas logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

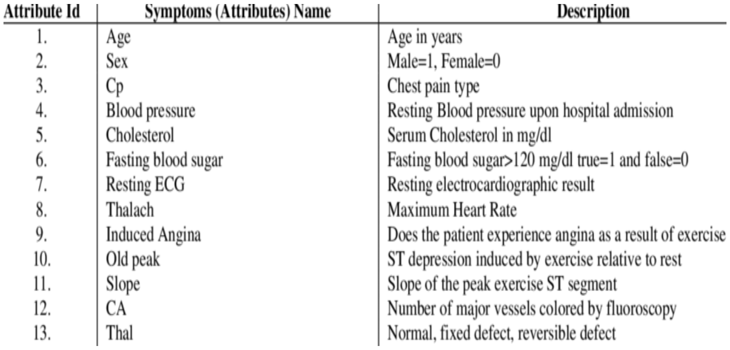
The logistics regression forms three types as below.

1.Binary logistics regression (two possible outcomes in a DV)
2.Multinomial logistics regression (three or more categories in DV without ordering)
3.Ordinal logistics regression (three or more categories in DV with ordering)

5. DATA DESCRIPTION

All the research papers referred in this research work have used 13 input attributes and the same is given in table 2 for prediction of CVD.

Table 2: Description of 13 input attributes

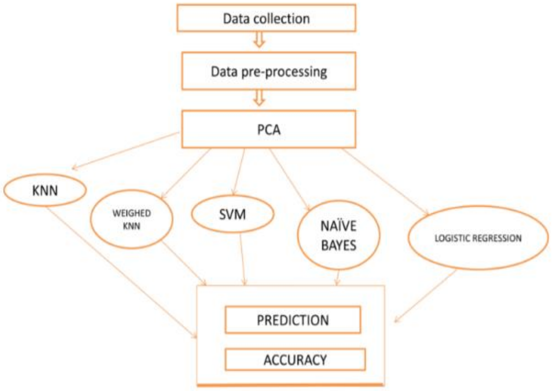


Attribute Id	Symptoms (Attributes) Name	Description
1.	Age	Age in years
2.	Sex	Male=1, Female=0
3.	Cp	Chest pain type
4.	Blood pressure	Resting Blood pressure upon hospital admission
5.	Cholesterol	Serum Cholesterol in mg/dl
6.	Fasting blood sugar	Fasting blood sugar>120 mg/dl true=1 and false=0
7.	Resting ECG	Resting electrocardiographic result
8.	Thalach	Maximum Heart Rate
9.	Induced Angina	Does the patient experience angina as a result of exercise
10.	Old peak	ST depression induced by exercise relative to rest
11.	Slope	Slope of the peak exercise ST segment
12.	CA	Number of major vessels colored by fluoroscopy
13.	Thal	Normal, fixed defect, reversible defect

6. SYSTEM ARCHITECTURE

The phases of the proposed work are:

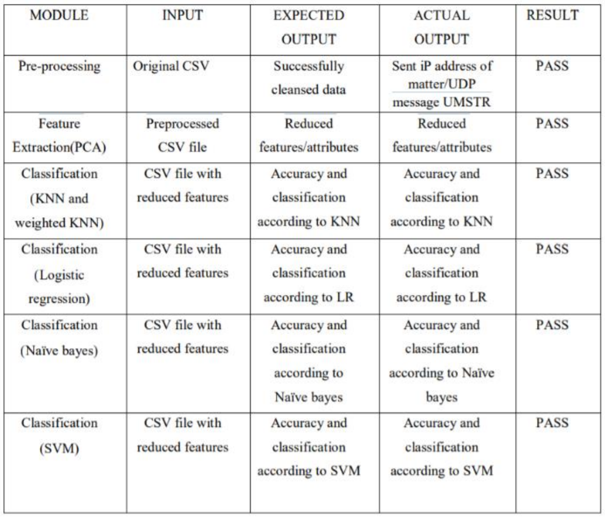
- Collection of raw data.
- Training data using Inception model.
- Testing data.
- Determine accuracy



7. SYSTEM DESIGN

Results and Testing:

Software testing is defined as an activity to check whether the actual results match the expected results and to ensure that the software system is Defect free. It involves execution of a software component or system component to evaluate one or more properties of interest. Software testing also helps to identify errors, gaps or missing requirements in contrary to the actual requirements. It can be either done manually or using automated tools. Some prefer saying Software testing as a White Box and Black Box Testing. In our system we have done manual testing as well as stress testing to check the breakpoint of the network. The manual testing was done using selenium software while stress testing was done manually with the help of hundreds of nodes that were rented from an online server. The first Testing was done in the first module i.e Data Preprocessing which is to ensure that the data set doesn't contain any missing value or unknown value. The original CSV file is taken as input and data cleansing is performed successfully. The second and third testing is done in second module i.e Feature Extraction to reduce the dimensionality of dataset. The preprocessed csv file is taken and pca is successfully applied separately to get the reduced feature dataset. The last four testings are performed for one classifier each i.e KNN, weighted KNN, logistic regression, SVM, naive-bayes to predict and classify the class of heart disease. The csv file with reduced features is taken as input and the accuracy and classification is done.



MODULE	INPUT	EXPECTED OUTPUT	ACTUAL OUTPUT	RESULT
Pre-processing	Original CSV	Successfully cleansed data	Sent iP address of matter/UDP message UMSTR	PASS
Feature Extraction(PCA)	Preprocessed CSV file	Reduced features/attributes	Reduced features/attributes	PASS
Classification (KNN and weighted KNN)	CSV file with reduced features	Accuracy and classification according to KNN	Accuracy and classification according to KNN	PASS
Classification (Logistic regression)	CSV file with reduced features	Accuracy and classification according to LR	Accuracy and classification according to LR	PASS
Classification (Naïve bayes)	CSV file with reduced features	Accuracy and classification according to Naïve bayes	Accuracy and classification according to Naïve bayes	PASS
Classification (SVM)	CSV file with reduced features	Accuracy and classification according to SVM	Accuracy and classification according to SVM	PASS

8. CONCLUSION

The results strongly suggest that machine learning can aid in the diagnosis of cardiac arrhythmias. It helps in the prediction of heart disease in its earliest stage. The earliest prediction of heart disease would help to take precautions at earlier stages. In conclusion, 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 disease. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research or work need to be performed for future. Would like to make use of testing different discretization techniques.

9. REFERENCES

Dr.S.Seema Shedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.

Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016.

Sairabi H.Mujawar, P.R.Devale, "Prediction of Heart Disease using Modified K-means and by using Naïve Bayes", International Journal of Innovative research in Computer and Communication Engineering, vol.3, October 2015, pp.10265-10273.

Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164-168