World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.218

HEART DISEASE PREDICTION SYSTEM USING NAÏVE BAYES DATA MINING TECHNIQUE

Amruta Powar*[1] and Prof. Dr. Vijay Ghorpade[2]

[1]Department of Computer Science, D. Y. Patil College of Engg. Kolhapur, Maharashtra, India.

[2]Department of Computer Science, Bharati Vidyapeeth‘s College of Engg. Kolhapur, Maharashtra, India.

Article Received on 27/04/2018 Article Revised on 17/05/2018 Article Accepted on 06/06/2018

***Corresponding Author**
Amruta Powar
Department of Computer Science, D. Y. Patil College of Engg. Kolhapur, Maharashtra, India.

ABSTRACT

According to WHO maximum death in worldwide are happened due to heart disease. So, advanced data mining techniques can be used to take out hidden patterns from healthcare industry, so that it becomes easy to predict heart disease. In this paper we develop heart disease prediction system using naïve bayes data mining technique, which will help in predicting heart disease so that diagnosing it can take less medical tests and provide effective treatments.

KEYWORDS: Data mining, Heart Disease, Naïve Bayes.

1. INTRODUCTION

Data mining deals with finding the relationships and global patterns from large databases which are unseen among large amounts of data. These days healthcare industry has large amounts of patient data but these data are not mined in order to give some hidden information, and thus to make effective decisions, advanced data mining techniques can be used. Using medical profiles such as age, gender, chest pain type, fasting blood sugar, resting blood pressure, resting electrographic results, cholesterol, maximum heart rate achieved, exercise induced angina, old peak, slope, number of vessels colored, defect type, obesity and smoking we can predict the probability of patients getting a heart disease.

Quality of service is the serious challenge that the healthcare industry has face to. Quality of service deals with diagnosing disease correctly & provides efficient treatments to patients. Heavy loss can happen because of poor diagnosis. Diagnosis is an important task that must be executed correctly and efficiently. The diagnosis is always based on doctor's experience and knowledge. This leads to unwanted results and excessive medical costs of treatment provided to patients. Therefore an automatic medical diagnosis system needs to be designed that can take benefit of heart disease database which is publicly available.

In this paper, we propose 3 steps to predict the heart disease.

2. MATERIALS AND METHODS

2.1 Data sources

In this paper, we use the heart disease data from machine learning repository of UCI.[1]

2.2 Feature description

15 clinical features have been recorded for each instance. Table 1 shows the 15 attributes and their description

Table 1: Clinical features and their description.

Sr. No.	Attribute	Description
1	Age	Age in years
2	Sex	Male (1) or female (0)
3	Cp (chest pain type)	1 - typical type 1, 2 - typical type angina, 3 - non-angina pain, 4 - asymptomatic
4	Trestbps (in mm hg)	Resting blood pressure
5	Chol (in mm/dl)	Serum cholesterol
6	Restecg (Resting electrographic results)	0 - normal, 1- having ST_T wave abnormal, 2 - left ventricular hypertrophy
7	(Fbs) Fasting blood sugar	1 ≥ 120 mg/dl, 0 ≤ 120 mg/dl
8	Thalach	Maximum heart rate achieved
9	Exang (Exercise induced angina)	0- no, 1 - yes
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope of the peak exercise ST segment	1 - unsloping, 2 - flat, 3 - downsloping
12	Ca (Number of major vessels colored by floursopy)	0-3 value
13	Thal (Defect type)	3- normal, 6- fixed defect, 7- reversible defect
14	Obes (obesity)	1 - yes, 0 - no
15	Smoke (Smoking)	past, 2- current, 3- never

2.3 Data pre-processing

Online dataset has every value of each attribute are separated by comma. First these records are arranged in the row-column format. But initially dataset has some fields, in which some value in the records was missing. These were identified and replaced with most appropriate values using mean code method. After cleaning, we use min max normalization method. It transforms a value which fits in the range 0 to 1 by using following formulae.

$$value = \frac{\mathrm{x} - xmin}{xmax - xmin}$$

Where x is actual value of an attribute, xmin is the lower limit of that attribute and xmax is the upper limit of that attribute. For labeling we have to define output parameter. Here num is the output parameter which ranges from 0 to 4. Here 0 is for healthy person, 1 is for 30% possibility for heart disease, 2 is for 50% possibility, 3 is for 80% possibility and 4 are for those persons who have higher possibility for heart disease.

2.4 Clustering

Partitioning of data set into subsets so that the data in each of the subset share a general feature is known as Clustering. Numerous methods are available for clustering. In proposed system K-Means clustering algorithm is used. K means is a partitional clustering approach where each cluster is associated with a centroid (center point). Each point is assigned to the cluster with the closest centroid. K-means groups the data in accordance with their characteristic values into K distinct clusters where K is the positive integer denoting the number of clusters, needs to be provided in advance. Here, Data is categorized into the same cluster having identical feature values.

2.5 Naïve Bayes

An advantage of the naïve bayes is that it requires a small amount of training data to estimate the parameters. Naive bayes is used to compute posterior probabilities given observations. For example, a patient may be observed to have certain symptoms. Bayes theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. In simple terms, a naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

Generally all machine learning algorithms need to be trained for supervised learning tasks like prediction. Here training means to train them on particular inputs in such a way that, if later on we may test them for unknown inputs (which they have never seen before) for which they may predict based on their learning. According to Naive bayes algorithm first we have to convert the data set into a frequency table. Create a frequency table for all the features against the different classes. Likelihood table is created by finding the probabilities.

Naïve Bayes Testing Phase will be used to compute posterior probabilities. For example, a patient may be observed to have certain symptoms. Bayes' theorem is used to compute the probability that a proposed diagnosis is correct, given that observation. Naïve Bayes technique recognizes the characteristics of patients with heart disease. It shows the possibility of each 15 input attribute for the predictable state.

Following steps are used to compute probability.

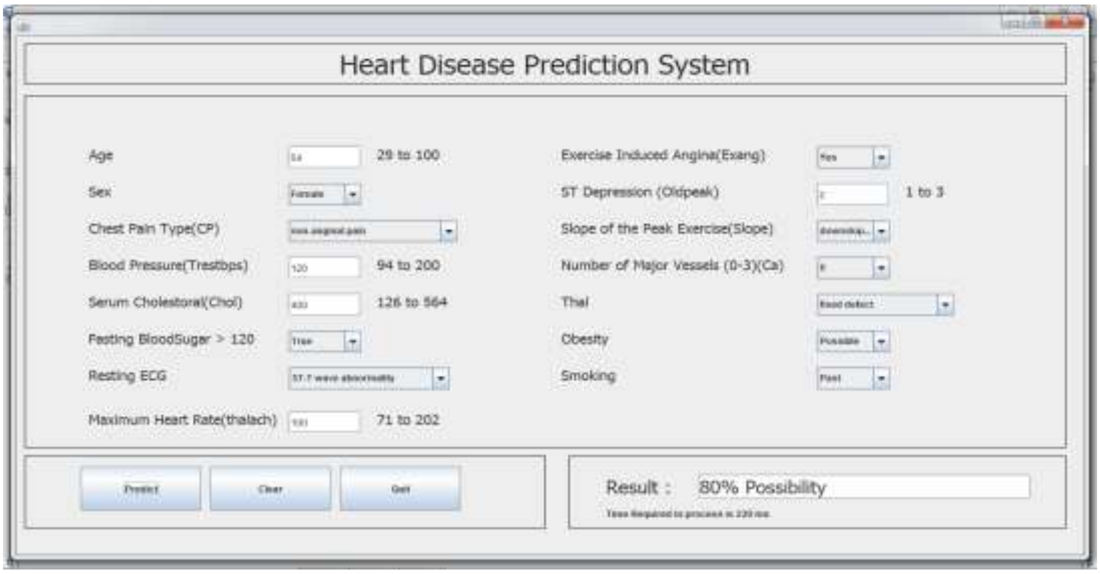
Step 1: First convert the normalized data set into a frequency table.

Step 2: By finding the probabilities create likelihood table.

Step 3: now, use naive following bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

Where P (c|x) is the posterior probability of class c given predictor x, P(c) is the prior probability of class, P (x|c) is the likelihood which is the probability of predictor given class, P(x) is the prior probability of predictor. In this way this module will calculate probability of patient getting a heart disease. That means we get prediction output.



Heart Disease Prediction System

Age	[illegible]	29 to 100	Exercise Induced Angina(Exang)	Yes	
Sex	Female		ST Depression (Oldpeak)	[illegible]	1 to 3
Chest Pain Type(CP)	non anginal pain		Slope of the Peak Exercise(Slope)	downslop...	
Blood Pressure(Trestbps)	120	94 to 200	Number of Major Vessels (0-3)(Ca)	[illegible]	
Serum Cholestoral(Chol)	[illegible]	126 to 564	Thal	fixed defect	
Fasting BloodSugar > 120	True		Obesity	Possible	
Resting ECG	ST-T wave abnormality		Smoking	Past	
Maximum Heart Rate(thalach)	[illegible]	71 to 202			

Result : 80% Possibility

Time Required to process is [illegible] ms

Figure 1: Snapshot of accepting input of 15 attributes from client.

Figure 1 shows snapshot of accepting input that is list of 15 attributes from client. The main functions of the Heart Disease Prediction System (HDPS) interface shown as Figure 2 include:

1. **Input clinical data section:** users input 15 pieces of clinical data.
2. **Predict button:** users click the button to get the result.
3. **Clear button:** users click the button to clear the previous input.
4. **Quit button:** users click the button to leave this interface.
5. **Result:** this text box shows the prediction result of the provided clinical data. It also shows the time (in milliseconds) required to complete the process.

2.6 Analysis of proposed System

Following parameters are used to analyze the proposed system:

1. **Rand Index:** It is a measure of the similarity between two data clustering.
2. **Davies–Bouldin index (DBI):** It is a metric for evaluating Clustering algorithm. It is used to check how well the clustering has been done.
3. **Accuracy:** It is used to check how accurate the result is.

3. RESULTS AND DISCUSSION

In proposed system we have used 15 attributes. We have also used maximum no. of records for training and remaining no. of records for testing. The testing procedure evaluates the proposed system model in terms of sensitivity, specificity and accuracy. We take 50 records for testing. And we calculate following table.

No. of Records	Sensitivity	Specificity
10	0.66	0.66
20	0.88	0.92
30	0.86	0.88
40	0.82	0.88
50	0.91	0.93

A confusion matrix is obtained to calculate the accuracy of classification shown in following table:

	Predicted Yes	Predicted No
Actual No	27	2
Actual Yes	0	21

From above discussion we can say that our system is 82% accurate which is the average accuracy. And when the dataset is increased then accuracy of the system is also get increased.

4. CONCLUSION

By using naïve bayes data mining technique, Heart disease prediction system is developed. The system extracts hidden knowledge from a heart disease database from UCI repository. This is the effective model to predict patients with heart disease. This system can be further expanded. For example, it can use other or extra medical attributes than 15 attributes used in this system. This system can also be developed using other data mining techniques.

REFERENCES

1. Blake, C. L., Mertz, C. J.: UCI Machine Learning Databases" http://mlearn.ics.uci.edu/databases/heartdisease.
2. Sellappan Palaniappan, Rafiah Awang Intelligent Heart Disease Prediction System Using Data Mining Techniques" IEEE Transactions on Knowledge and Data Engineering, 2014; 1- 14.
3. Chaitrali S. Dangare Sulabha S. Apte "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888), 2012; 47(10).
4. Shantakumar B. Patil, Y. S. Kumaraswamy "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network" European Journal of Scientific Research ISSN 1450-216X, 2009; 31: 642-656.
5. Shadab Adam Pattekari and Asma Parveen "Prediction System For Heart Disease Using Naive Bayes" International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624, 2012; 3(3): 290-294.
6. G. Subbalakshmi, K. Ramesh, M. Chinna Rao "Decision Support in Heart Disease Prediction System using Naive Bayes" Indian Journal of Computer Science and Engineering (IJCSE)ISSN : 0976-5166, 2011; 2.
7. Sivagowry S, Dr. Durairaj. M and Persia. A "An Empirical Study on applying Data Mining Techniques for the Analysis and Prediction of Heart Disease" IEEE conference paper 978-1-4673-5788, 2013.
8. An efficient k-means clustering Algorithm: Analysis and implementation by Tapas kanungo, David M. Mount, Nathan s. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.