INTERNATIONAL JOURNAL OF ENHANCED RESEARCH IN SCIENCE TECHNOLOGY & ENGINEERING
VOL. 2 ISSUE 3, MARCH.-2013 ISSN NO: 2319-7463

Heart Disease Prediction System using Naive Bayes

Dhanashree S. Medhekar[1], Mayur P. Bote[2], Shruti D. Deshmukh[3]
[1]dhanashreemedhekar@gmail.com, [2]mayur468@gmail.com, [3]deshshruti88@gmail.com

`

Abstract: As large amount of data is generated in medical organisations (hospitals,medical centers) but as this data is not properly used. There is a wealth of hidden information present in the datasets. This unused data can be converted into useful data. For this purpose we can use different data mining techniques. This paper presents a classifier approach for detection of heart disease and shows how Naive Bayes can be used for classification purpose. In our system, we will categories medical data into five categories namely no,low, average,high and very high.Also, if unknown sample comes then the system will predict the class label of that sample. Hence two basic functions namely classification (training) and prediction (testing) will be performed. Accuracy of the system is depends on algorithm and database used.

Keywords: data mining, heart disease, Naive Bayes.

I. INTRODUCTION

1.1 Data mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.2 Basic terms related to data mining:

1.2.1 Classification

Classification is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be “sunny”, “rainy” or “cloudy”.

1.2.2 Supervised learning:

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier. The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

1.2.3 Unsupervised learning:

In machine learning, unsupervised learning refers to the problem of trying to find hidden structure in unlabeled data. Since the examples given to the learner are unlabeled, there is no error or reward signal to evaluate a potential solution. This distinguishes unsupervised learning from supervised learning .

1.2.4 Prediction:

Models continuous-valued functions, that is predicts unknown or missing values.

1.3 Data Source:

Description of Cleveland Dataset:
This dataset contains information concerning heart disease diagnosis . The data was collected from the cleveland clinic foundation, and it is available at UCI Repository . Six instances containing missing values have been deleted from the original dataset .

Format :
A data frame with 303 observations on the following 14 parameters :

P1 - Age
P2 - Gender
P3 – CP (chest pain)
P4 - trestbps : resting blood pressure
P5 – cholesterol
P6 – fbs: fasting blood sugar>120 ? yes=1,no = 0
P7 – restecg: resting electrocardiographic results 0,1,2
P8 – thalach : maximum heart rate achieved
P9 – exang : exercise induced angina (1= yes ; 0= no)
P10 – oldpeak = ST depression induced by exercise relative to rest
P11 – slope : the slope of the peak exercise ST segment
P12 – ca: no. of major vessels (0 to 3) colored flurosopy
P13 – thal :3 =normal ,6=fixed defect ,7= reversable defect
P14 – diagnosis of heart disease

II. RELATED WORK

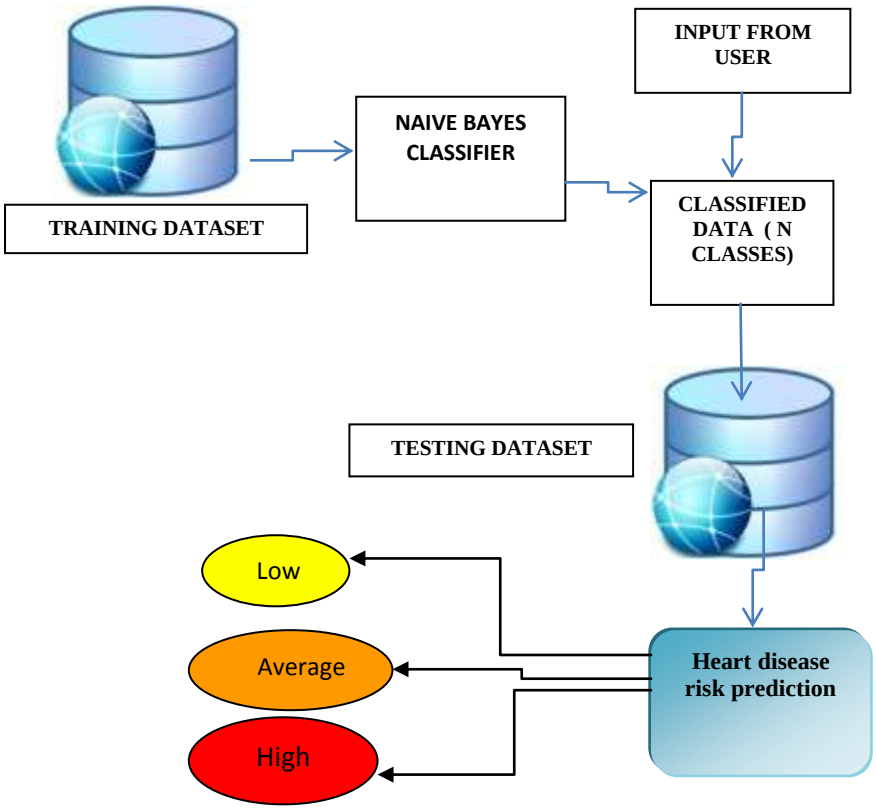

Fig.1 System Architecture

As shown in Fig 2.1,the training dataset is given as input to the classifier.This classified data is further used for testing purpose. We have used algorithm Naive Bayes. Mainly system will work in two phases:

1)Training phase
2)Testing phase

2.1.1 Training Phase:

Classification assumes labeled data: we know how many classes there are and we have examples for each class (labeled data).

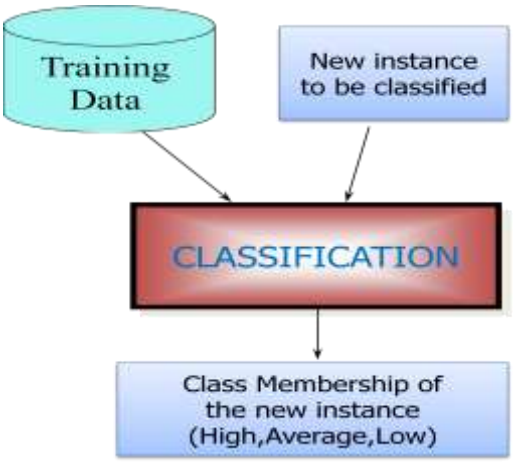

Fig.2 Classification

Classification is supervised. Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

2.1.2 Testing Phase:

Testing phase involves the prediction of unknown data sample.

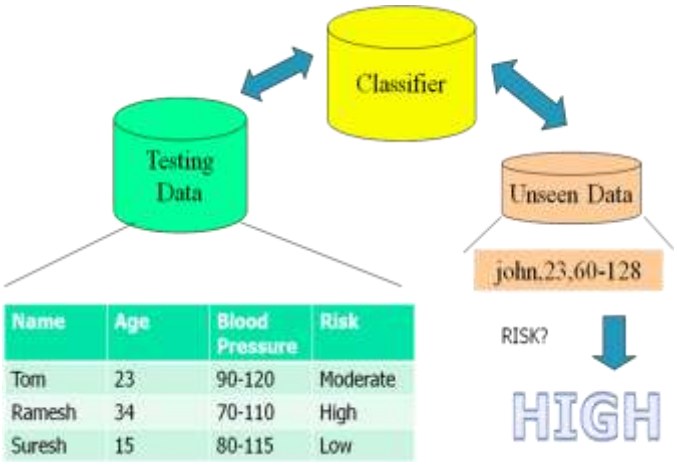

Fig.3 Prediction

Models continuous-valued functions, i.e., predicts unknown or missing values. In testing we check those data that does not come under the dataset we have considered. After the prediction, we will get the class labels.

III. TECHNIQUES USED

3.1. Naive Bayes:

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

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

$$P(H|X)=P(X|H)P(H)/P(X) \quad (1.1)$$

Algorithm:

The Naive Bayes algorithm is based on Bayesian theorem as given by equation (1.1)

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). Note that the class prior probabilities may be estimated by P(Ci) = si/s , where Si is the number of training samples of class Ci, and s is the total number of training samples. on X. That is, the naive probability assigns an unknown sample X to the class Ci [2

IV. RESULTS AND ANALYSIS

Results and analysis is done on Cleveland dataset.Results are shown in the form of pie charts,bar charts. Table 1 shows the accuracy obtained by changing the number of instances in the training dataset.

Table.1 Accuracy(%)

Number Of records in Traning dataset	Number of records in Testing dataset	Number of Correctly classified instances	Number of Incorrectly classified instances	Accuracy (%)
303	276	245	31	88.76
303	240	215	25	89.58
303	290	258	32	88.96

Fig.4 shows the classified data in the form of Pie chart.0,1,2,3,4 represents the posssibility of heart disease.
0:No, 1:Low, 2:Average, 3:High 4:Very high

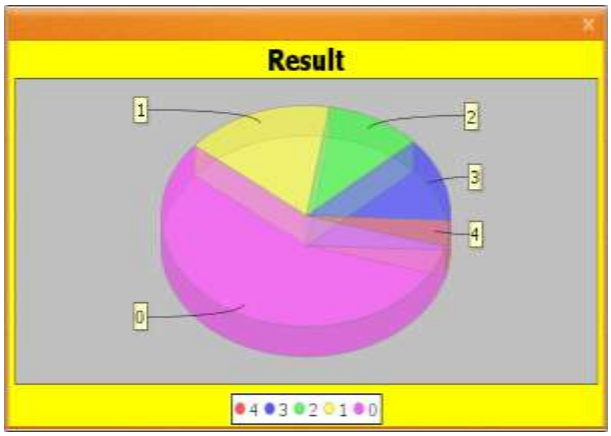

Fig.4 Classified data

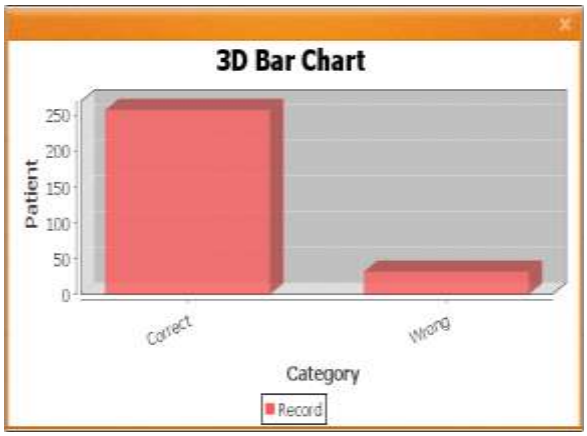

Fig.5 Prediction

Fig.5 shows the correctly and wrongly classified records in the form of bar chart.

V. CONCLUSIONS AND FUTURE WORK

This system classifies the given data into differerent categories and also predicts the risk of the heart disease if unknown sample is given as an input.The system can be served as training tool for medical students.Also,it will be helping hand for doctors.As we have developed generalised system,in future we can use this system for analysis of different datasets by only changing the name of dataset file which is given for training module.

REFERENCES

[1]. Mai Shouman, Tim Turner, Rob Stocker, "Using data mining techniques in heart disease diagnosis and treatment", Japan-Egypt Conference on Electronics, Communications and Computers 978-1-4673-0483-2 c_2012 IEEE.
[2]. N. Aaditya Sunder, P. PushpaLatha, "Performance analysis of classification data mining techniques over heart disease database" Inernational Journal Of Engineering Science and Advance Technology"-vol-2 issue-3,470-478,May-June 2012.
[3]. Han, J., Kamber, M.: "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2006.
[4]. IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August, 2008.
[5]. SellappanPalaniappan, RafiahAwang, Intelligent Heart Disease Prediction System Using Data Mining Techniques, ©2008 IEEE.
[6]. ShantakumarB.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 Vol.31 No.4 (2009), pp.642-656 © EuroJournals Publishing, Inc. 2009.
[7]. R. Bhuvaneswari and K. Kalaiselvi, Naive Bayesian Classification Approach in Healthcare Applications International Journal of Computer Science and Telecommunications, [Volume 3, Issue 1, January 2012].
[8]. Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.
[9]. [9]Data mining concepts and techniques, second edition, Han Kamber.