Heart Disease Diagnosis Using Naive Bayes Classifiers

**Abstract:** Heart disease is a major cause of morbidity and mortality in modern society. Medical diagnosis is extremely important but complicated task that should be performed accurately and efficiently. The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. There is a huge data available within the healthcare systems. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not “mined” to discover hidden information for effective decision making. Using medical profile such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease.

**L. Pandeeswari**  
M.Phil (Computer Science), Research Scholar,  
Vivekanandha College for Women, Unjanai, Tiruchengode.  
India.

**K.Rajeswari**  
M.Sc., M.Phil.  
Assistant Professor in Computer Science  
Vivekanandha College for Women, Unjanai, Tiruchengode.  
India.

**ISSN 2319-9725**
1. **Introduction:**

Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. “Data Mining is a non-trivial extraction of implicit, previously unknown and potential useful information about data”. In short, it is a process of analyzing data from different perspective and gathering the knowledge from it. The discovered knowledge can be used for different applications for example healthcare industry. Nowadays healthcare industry generates large amount of data about patients, disease diagnosis etc. Data mining provides a set of techniques to discover hidden patterns from data. A major challenge facing Healthcare industry is quality of service. Quality of service implies diagnosing disease correctly & provides effective treatments to patients. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. Fayyad defines data mining as “a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database”. Giudici defines it as “a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database”. Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k-means clustering is unsupervised).

2. **Heart Disease:**

The heart is important organ of human body part. It is nothing more than a pump, which pumps blood through the body. If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it. A number of factors have been shown that increases the risk of Heart disease:

   a. Family history
   b. Smoking
   c. Poor diet
d. High blood pressure  
e. High blood cholesterol  
f. Obesity  
g. Physical inactivity  
h. Hyper tension

Factors like these are used to analyze the Heart disease. In many cases, diagnosis is generally based on patient’s current test results & doctor’s experience. Thus the diagnosis is a complex task that requires much experience & high skill.

3. Literature Survey:

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

An Intelligent Heart Disease Prediction System (IHDPS) is developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al. Each method has its own strength to get appropriate results. To build this system hidden patterns and relationship between them is used. It is web-based, user friendly & expandable. To develop the multi-parametric feature with linear and nonlinear characteristics of HRV (Heart Rate Variability) a novel technique was proposed by Heon Gyu Lee et al. To achieve this, they have used several classifiers e.g. Bayesian Classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree) and SVM (Support Vector Machine).

The prediction of Heart disease, Blood Pressure and Sugar with the aid of neural networks was proposed by Niti Guru et al. The dataset contains records with 13 attributes in each record. The supervised networks i.e. Neural Network with back propagation algorithm is used for training and testing of data.

The problem of identifying constrained association rules for heart disease prediction was studied by Carlos Ordonez. The resultant dataset contains records of patients having heart disease. Three constraints were introduced to decrease the number of patterns. They are as follows:
a. The attributes have to appear on only one side of the rule.
b. Separate the attributes into groups. i.e. uninteresting groups.
c. In a rule, there should be limited number of attributes.

The result of this is two groups of rules, the presence or absence of heart disease.

Franck Le Duff et al. builds a decision tree with database of patient for a medical problem.

Latha Parthiban et al. Projected an approach on basis of coactive neuro-fuzzy inference system (CANFIS) for prediction of heart disease. The CANFIS model uses neural network capabilities with the fuzzy logic and genetic algorithm.

Kiyong Noh et al. Uses a classification method for the extraction of multiparametric features by assessing HRV (Heart Rate Variability) from ECG, data pre-processing and heart disease pattern. The dataset consisting of 670 peoples, distributed into two groups, namely normal people and patients with heart disease, were employed to carry out the experiment for the associatve classifier.

4. Data Set:

In our work, six attributes have been reduced to four attributes which are employed for heart disease prediction. The data of various patients is entered in the proposed system and the diagnosed results generated by the system corresponding to patients have been saved in the database. The resultant data set thus obtained is used by the classification model for calculating the efficiency of the proposed system. Attributes have been converted to categorical form for more clarity. Moreover, training set method is used as the test mode.

4.1. Input Attributes:

a. Type - Chest Pain Type
b. Rbp - Resting blood pressure
c. Eia - Exercise induced angina
d. Oldpk - Old peak
e. Vsl - No. of vessels colored
f. 6.Thal -Maximum heart rate achieved
4.2. Reduced Input Attributes:

a. CPTYPE - Chest Pain Type  
b. BP - Blood Pressure  
c. CA - No. of vessels colored  
d. TMT – Treadmill Test

5. Algorithms:

5.1. Naive Bayes Classifier Technique:

It is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, it can outperform more sophisticated classification methods.

5.2. Naive Bayes:

a. 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.

b. 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 on X. That is, the naive probability assigns an unknown sample X to the class Ci if and only if:

   a. \( P (Ci/X) > P(Cj/X) \) for all \( 1 \leq j \leq m \) and \( j \neq i \)

   b. 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,

   c. \( P (Ci/X) = \frac{P(X/Ci)P(Ci)}{P(X)} \)

   c. 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 \( P(C_i) = \frac{s_i}{s} \), where \( S_i \) is the number of training samples of class \( C_i \), and \( s \) is the total number of training samples.

### 5.2.1. Pseudo Code:

Calculate diagnosis="yes", diagnosis="no" probabilities \( P_{yes}, P_{no} \) from training input.

For Each Test Input Record

For Each Attribute

Calculate Category of Attribute Based On Categorical Division

Calculate Probabilities Of Diagnosis="Yes", Diagnosis="No" Corresponds To That Category \( P(Attr,Yes), P(Attr,No) \) From Training Input .

For Each Attribute

Calculate The Resultyes= Resultyes* \( P(Attr,Yes) \), Resultno= Resultno*\( P(Attr,No) \);

Calculate Resultyes= Resultyes *\( P_{yes} \)

Resultno= Resultno*\( P_{no} \);

If(Resultyes > Resultno) Then Diagnosis="Yes";Else Then Diagnosis ="No";

### 5.2.2. Formulae:

\( P_{yes} = \frac{\text{total number of yes}}{\text{total number of records}} \);

\( P_{no} = \frac{\text{total number of no}}{\text{total number of records}} \);

\( P(\text{attr,yes}) = \frac{\text{total number of yes in corresponding category}}{\text{total number of yes}} \);

\( P(\text{attr,no}) = \frac{\text{total number of yes in corresponding category}}{\text{total number of yes}} \);

### 6. Conclusion And Future Work:
The objective of our work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, six attributes were involved in predicting the heart disease. In our work, six attributes are reduced to four attributes which automatically reduces the number of tests to be taken by a patient. Subsequently, Naive Bayes Classifiers are used for calculating the efficiency of the proposed system. In our future work we wish to conduct experiments on large real time health datasets to predict the diseases like heart attack and compare the performance of our algorithm with other related algorithms.
References:


6. Cleveland database: http://archive.ics.uci.edu/ml/datasets/Heart+Disease