



## **Intelligent and Effective Prediction of Various Diseases Prognosticating Naïve Bayes Classifier**

\*<sup>1</sup>T. Hannah Rose Esther, <sup>2</sup>G. Kannan, <sup>3</sup>Suresh Sagadevan, <sup>4</sup>L. Sharmila,

<sup>1</sup>Department of Information Technology, AMET University, Chennai, India

<sup>2</sup>Centre for Non-Destructive evaluation, AMET University, Chennai, India

<sup>3</sup>Department of Physics, AMET University, Chennai, India

<sup>4</sup>Department of Computer Science & Engg, Alpha College of Engineering, Chennai, India

**Abstract :** Prognosis of various diseases such as diabetes, cancer, and heart attack always tend to be a continuous problem that causes misleading assumptions and is intermittently accompanied by hasty effects. This system predicts better when it utilizes Naïve Bayes classifier. Using medical diagnosis which involves age, blood sugar, sex, blood pressure, and various other factors, it can predict the likelihood of patients getting these diseases. Medications and suggestions can be obtained from the data obtained. It enables significant knowledge, e.g. patterns, relationships between medical factors. It can be mobile accessible, user-friendly and a reliable source for both patient and the doctor.

**Keywords :** Prognosis, Naive Bayes classification, data mining, training datasets.

### **1. Introduction**

The prognosis of certain diseases is proved to be a imperative and sophisticated task in medicine. The prognosis of these diseases from various features or signs is a continuous problem is frequently accompanied by abrupt effects. Thus we attempt to assemble the expertise of several doctors and patient report composed in databases to aid the diagnosis procedure in regarded as a valuable assessment [1]. A Naive Bayes classification algorithm, for the derivation of decisive patterns from the heart disease, cancer and diabetics warehouses for prediction has been presented.

The relevance of the research is to promote a prototype of an Intelligent Heart Disease Prediction System employing data mining modeling technique Naïve Bayes. It can discover and extract hidden knowledge (patterns and relationships) associated with the disease from a historical patient database [2]. It can answer various queries for the diagnosis of various diseases and benefit healthcare practitioners for making impeccable prognosis which general human predictions cannot. By providing effective treatments, it also aids to reduce mis-predictions. To enhance perception and for gratification in judgment, it displays the predictable values in both graphical and tabular forms.

### **2. Related Work**

In “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, by Sellappan Palaniappan and Rafiah Awang, IEEE2008 a prototype prediction system for heart disease was developed using three classification modeling techniques for data mining [3]. Many hospital information systems are

significantly designed to help inventory management, patient billing, and generation of simple statistics. Some hospitals utilize decision support systems, but they are largely limited [4].

They can answer simple questions like “The average age of patients who had heart disease”, “How many surgeries occurred in hospital stay for longer than 10 days?”, “Identify the patients who are single, female and above 30 years old, and who have been treated for cancer.” However, they could not answer complicated queries like “Identifying the important preoperative predictors that may increase the length of hospital stay”, “The submitted patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?”, and “Given patient records, predict the prognosis of patients getting a heart disease.”[3] Clinical prognosis would be generally depending on doctors’ knowledge and background experience, than on the knowledge-rich data that are hidden within the database [5]. This practice causes biased results, excessive medical costs and errors which largely influence the quality of service to patients.

Wu, et al had proposed such that integration in clinical prognosis support (medical practices) with data-warehoused patient records could diminish the medical errors and thus enhance patient wellness, decrease unwanted practice variation, and improve the patient outcome [6]. This suggestion is a promising data modeling and analysis tool, e.g., data mining, which has the potential to generate and use a knowledge-rich environment (databases) which can help significantly to improve the quality of clinical decisions [6]. The system thus extracts hidden knowledge from a historical heart disease database. The models are trained and validated against a test dataset. Lift Chart and Classification Matrix methods are used to test and evaluate the effectiveness of the models. All three models are able to extract patterns in response to the predictable state.

The most effective model to predict patients with probability of having heart disease is Naive Bayes followed by Neural Network and Decision Trees. Five mining goals are prescribed based on business intelligence and data exploration. The goals are calculated against the trained models. All three models answer complex queries; each has its own strength implying ease of model interpretation, and access to detailed information and accuracy. Naive Bayes answers four out of the five goals; Decision Trees, three; and Neural Network, could answer only two [4]. Naive Bayes resulted in better prediction than Decision Trees as identified all the significant medical predictors. The relationship between attributes produced by Neural Network is more difficult to understand. The IHDPS employs CRISP-DM methodology to help build the mining models.

In “Multi-class classification algorithm for optical diagnosis of cancer”, by A Gupta and S Gupta, December 2005 a project report development which uses direct multi-class spectroscopic diagnostic algorithm for identification of high-grade cancerous tissue sites from low-grade and precancerous and normal squamous tissue sites of human oral cavity. Here the use of recently formulated theory of Total Principal Component Regression (TCPR) for the development of multi-class classification algorithm for the optical diagnosis of cancer [7]. The autofluorescence spectral data that are acquired from patients screened for neoplasmas of oral cavity of government cancer hospital, Indore was used to train and validate the algorithm. The disadvantage of this project is that it doesn’t involve prediction. Only one disease was analyzed at high cost due to equipment cost.

In “Smart Home-based Health Platform for Behavioral Monitoring and alteration of Diabetics Patients” by Abdelsalam Helal, Mark Schmalz, and Diane J. Cook, 2007 a smart home-based software architecture that aid in behavioral monitoring for diabetes patients was designed. Medical researchers and practitioners have desired the ability to continuously and automatically monitor the patients with diabetics [8]. Additionally, the use of monitoring data to influence treatment dosage or regimen in real-time is an important objective of behavior modification practice. Thus these capabilities can contribute to the healthcare systems, to conquer key obstacles to bring acceptable quality-of-service at a reasonable unit cost per patient.

### **3. Materials and Methods**

#### **3.1 Naive bayes classification**

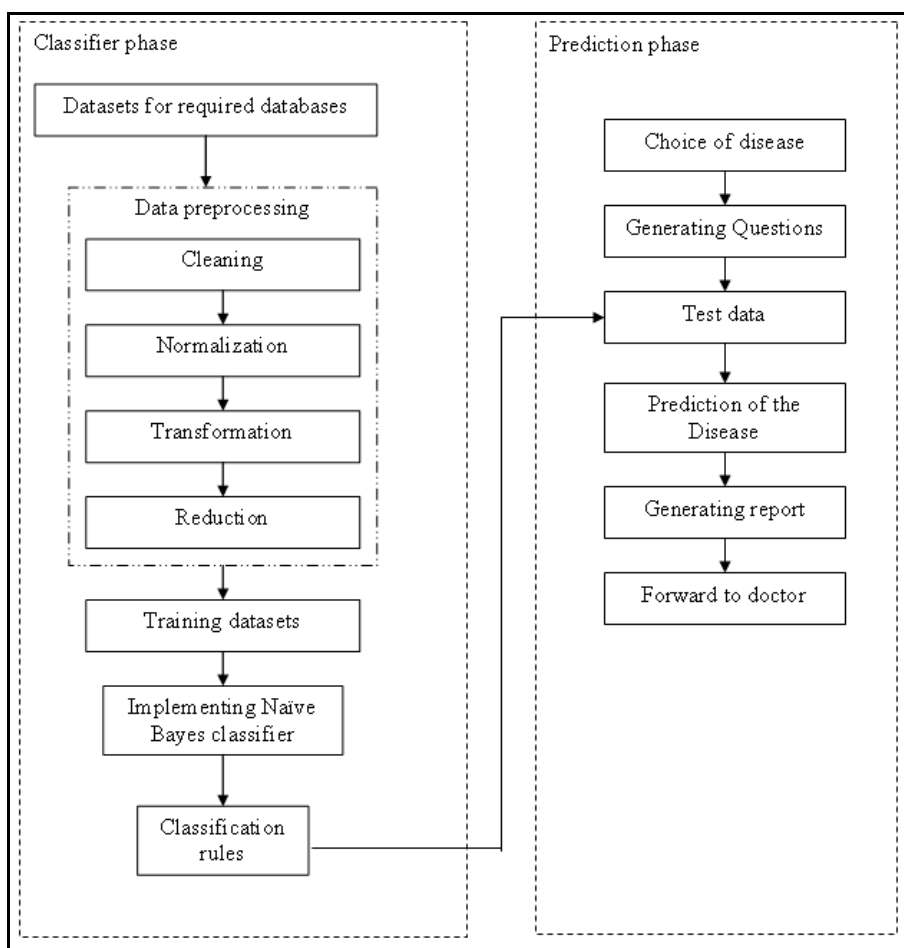
The order should be possible by discovering the rate of affectability and specificity. Affectability (additionally called genuine positive rate or the review rate in a few fields) measures the specific proportion of effectively distinguished genuine positives. Specificity (at times called the genuine negative rate) measures the specific proportion of effectively distinguished genuine negatives. These two measures are firmly identified

with the ideas of sort I and sort II blunders [9]. A sort I blunder (or mistake of the principal kind) will be the off base disposal of a genuine invalid speculation. It speaks to false positive. Typically a sort I mistake drives one to reason that an assumed impact or relationship exists when in reality it doesn't. Cases of sort I mistakes incorporate an after-effect of the tests that demonstrates a patient to have a specific malady in actuality the patient does not acquire the sickness. A sort II blunder is the inability to dismiss a false invalid theory. As for the non-invalid speculation, it speaks to a false negative. Cases of sort II mistakes are blood test which neglects to recognize the sickness it was intended to identify, in a patient who has the ailment.

This paper introduces a novel characterization plan to enhance arrangement execution when couple preparing information is accessible. In the proposed plot, the information is portrayed utilizing the discretized measurable elements. The grouping in a classifier mix system and hypothetically breaks down the execution. Encourage, arrangement technique is proposed to total the Naive Bayes (NB) intensions of the associated information. We additionally show an examination on forecast of mistake affectability of the accumulation methodologies. Countless are done on substantial scale certifiable datasets to assess the proposed system.

#### 4. Architecture of Proposed System

The proposed system contemplates various propositions such as doctor details and prescriptions. Each disease will have different specialists for analyzing the disease. The details of doctors with respect to the disease, along with their location will be given. Cost and effort of visiting the doctor in the initial or unknown state could be avoided since the medications will be prescribed.



**Fig.1 System Architecture**

First phase (classifier phase): Data preprocessing processes raw data to qualify and quantify it for another processing procedure. It is most commonly used for initial data mining practice. Data preprocessing mutates the data into a reformed format that will be processed regarding the purpose of the user. In a customer relationship management (CRM) context, data preprocessing is a component of Web mining [10]. Web usage logs are preprocessed to extract relational sets of data called transactions, which has groups of URL references.

User sessions are referred to track and to identify the user, the Web sites requested and their order and duration spent on each one. Training data is used for research of the process; we monitor the demeanor of the random process for some time, collecting a large number of samples.

The system architecture can be viewed as two phases. Figure 1 presents the two phases of the research.

#### 4.1. Analyzing the Data set

A **data set** (or **dataset**) is a collection of data, usually presented in the form of a table [11]. Each column serves for a particular variable. Each row correlates to a member of the data set in question respectively. It index data for each and every variable. For example heart disease involves height, weight and age for a patient or values of random numbers. Each cell is known as a datum. The data set compose the data for individual members, in each row [12].

The values are usually numbers that are real numbers or integers. For example depicting a person's height in centimeters, but may not consist of numerical values, for example depicting a person's name or community. Commonly, values are any of the kinds described as a level of measurement. For each and every variable, the values will normally be the same. However, there may likewise be "missing values", which need to be indicated in some way.

The "Diagnosis" attribute was classified as an attribute that could be predicted with value "1" for patients with disease and the value is "0" for patients without disease. The attribute "PatientID" is used as the key; the rest are input attributes which are updated frequently. It can be assumed that the problems like missing data, duplicate data, and inconsistent data have all been determined.

For our research purpose we get data set from .dat file, and our file reader program which extracts necessary data from them to be giving as the input for Naïve Bayes. The attributes that are available for each disease are explained in the following:

Consider Heart attack,

Key attribute

1. Patient ID – Patient's unique identification number

Input attributes

1. Age (Years)
2. Sex (value 1: Male; value 0: Female)
3. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type 2 angina, value 3: non-angina pain; value 4: asymptomatic)

Angina is chest pain that exists when an area of your heart muscle doesn't get enough oxygen-rich blood [8]. Angina may feel like pressure or squeezing in your chest. The pain also may occur in your shoulders, arms, neck, jaw, or back[1, 3]. It feels like indigestion. An underlying heart problem can cause Angina. Angina can be considered as a symptom of an underlying coronary artery disease (CAD). A fatty material called plaque over break on the inner walls of the coronary arteries.

The increase of plaque in the arteries is the condition called atherosclerosis (ATH-er-o-skler-O-sis) as shown in figure 2.

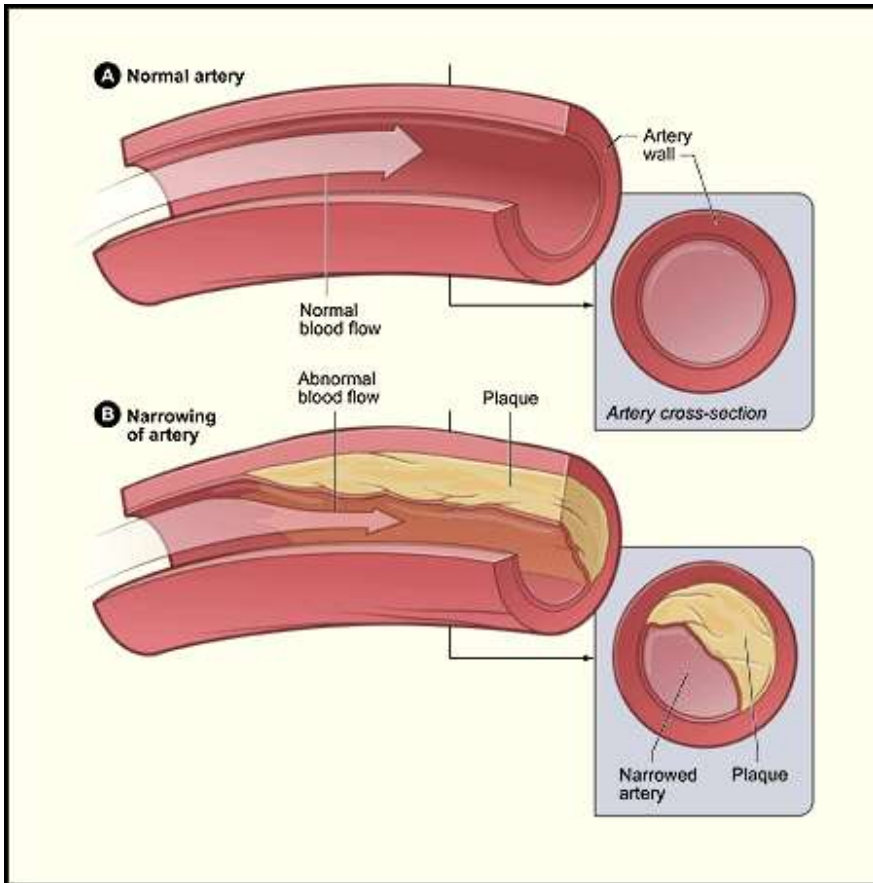


Fig.2 Atherosclerosis

Table.1 Heart Disease Datasets

Age	Se	CP	TBP	SC	FB	REC	thalac	exan	O	slop	C	Tha	H	W	R
63.	1.0	1.0	145.	233.	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	16	50	0
67.	1.0	4.0	160.	286.	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	15	70	1
40.	1.0	4.0	110.	167.	0.0	2.0	114.0	1.0	2.0	2.0	0.0	7.0	14	54	1
37.	1.0	3.0	130.	250.	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	16	58	0
41.	0.0	2.0	130.	204.	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	15	60	0

Table.2 Cancer Datasets

CT	UCSi	UCSp	MA	SECS	BN	BC	NN	mitoses	R
04.0	01.0	01.0	01.0	02.0	01.0	02.0	01.0	01.0	2
04.0	01.0	01.0	01.0	02.0	01.0	03.0	01.0	01.0	2
10.0	07.0	07.0	06.0	04.0	10.0	04.0	01.0	02.0	4
06.0	01.0	01.0	01.0	02.0	01.0	03.0	01.0	01.0	2
07.0	03.0	02.0	10.0	5.0	10.0	05.0	04.0	04.0	4

Table.3 Diabetics Datasets

No.	Sex	PGC	DBP	TSFT	2hr-SI	BMI	DPF	Age	R
1	1.0	148.0	72.0	35.0	000.0	33.6	0.627	50.0	1
2	0.0	085.0	66.0	29.0	000.0	26.6	0.351	31.0	0
3	1.0	183.0	64.0	00.0	000.0	23.3	0.672	32.0	1
4	0.0	89.0	23.0	23.0	094.0	28.1	0.167	21.0	0
5	0.0	078.0	50.0	32.0	088.0	31.0	0.248	26.0	1

Sample Screenshots

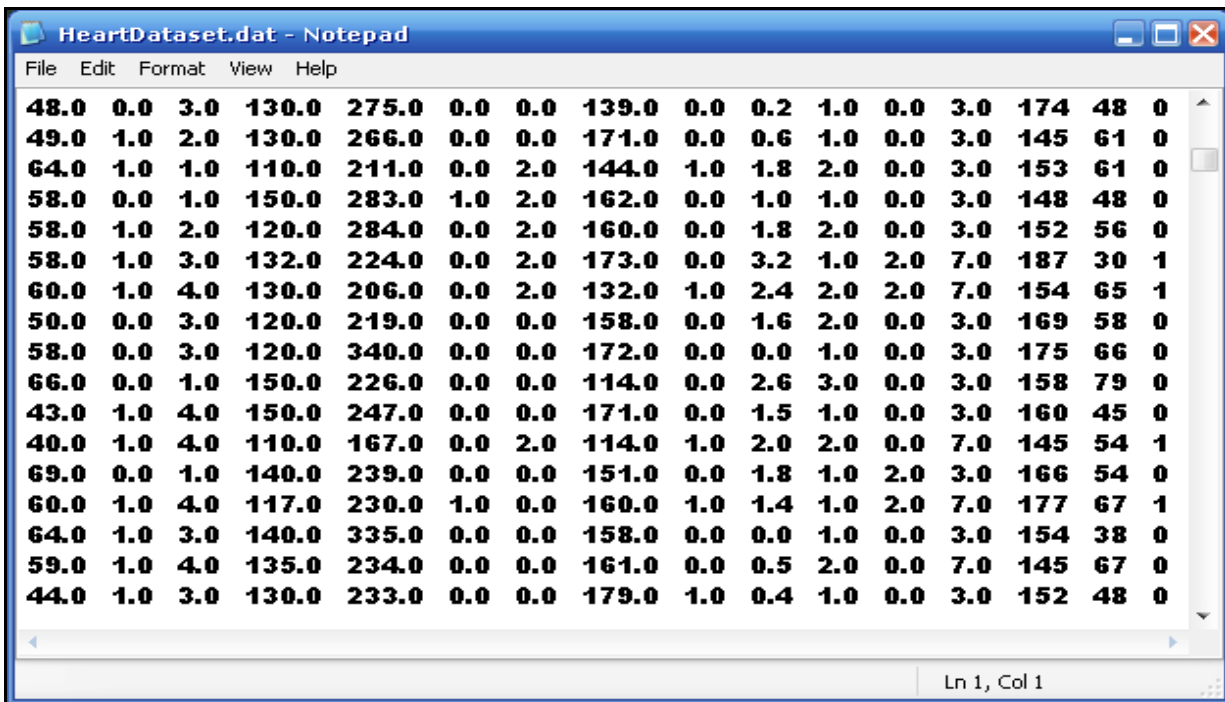


Fig.3 Heart Disease Data Sets

4.2 Naives Bayes Implementation in Mining

Bayes' Theorem diagnoses the probability of an event occurrence, given the probability of another event that has already occurred. If B as the dependent event and A serve as the prior event, Bayes' theorem can be declared as follows in formulas 1 and 2.

Bayes' Theorem

$$\text{Prob (B given A)} = \text{Prob (A and B)}/\text{Prob (A)} \tag{1}$$

$$P(X | C_i) = \prod_{k=1} P(x_k | C_i) \tag{2}$$

To calculate the probability of B given the value A, the algorithm computes the number of cases where A and B occur together and divide it by the number of cases where A occurs alone[1][6]. First we should appraise the mean  $\mu$  and standard deviation  $\sigma$  values for the numerical attributes  $X_i, i=1...n$  – the  $i^{\text{th}}$  measurement, n number of measurements

$$\mu = \frac{\sum_{i=1}^n X_i}{n} \tag{3}$$

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n - 1} \tag{4}$$

Then the result will be applied to the following formula

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

Table.4 Example

Age	TBP	Cholesterol	Diagnosis
63	145	233	0
87	160	286	1
37	130	210	0
57	140	192	0
44	120	263	0
56	130	256	1
58	132	224	1
41	105	198	0
60	130	206	1
77	118	219	1

The values of mean and standard deviation are calculated as follows:

$$\mu = \frac{\sum_{i=1}^n X_i}{n} \quad (3)$$

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \mu)^2}{n-1} \quad (4)$$

$$\mu_{\text{Age}_1} = 67.6$$

$$\mu_{\text{Age}_0} = 48.4$$

$$\mu_{\text{TBP}_1} = 134$$

$$\mu_{\text{TBP}_0} = 148$$

$$\mu_{\text{Ch}_1} = 218.6$$

$$\mu_{\text{Ch}_0} = 217.2$$

$$\sigma_{\text{Age}_1} = 13.69$$

$$\sigma_{\text{Age}_0} = 11.082$$

$$\sigma_{\text{TBP}_1} = 15.556$$

$$\sigma_{\text{TBP}_0} = 33.279$$

$$\sigma_{\text{Ch}_1} = 63.6459$$

$$\sigma_{\text{Ch}_0} = 28.2259$$

Thus the mean and standard deviation are calculated. Applying this in the formula of probability for input Age=66 TBP= 150 Cholesterol = 210 we get:

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

$$f(\text{Age}=66|\text{DV}=1) = (1/(13.68)\Gamma(2*3.14))e^{-(66-67.6)^2/(2*(13.68^2))} = 0.02895$$

$$f(\text{Age}=66|\text{DV}=0) = (1/(11.08)\Gamma(2*3.14))e^{-(66-48.4)^2/(2*(11.08^2))} = 0.010199$$

$$f(\text{TBP}=150|\text{DV}=1) = (1/(15.556)\Gamma(2*3.14))e^{-(150-134)^2/(2*(15.556^2))} = 0.01511$$

$$f(\text{TBP}=150|\text{DV}=0) = (1/(33.279)\Gamma(2*3.14))e^{-(150-148)^2/(2*(33.279^2))} = 0.011966$$

$$f(\text{Ch}=210|\text{DV}=1) = (1/(1052.2)\Gamma(2*3.14))e^{-(210-218.6)^2/(2*(1052.2^2))} = 0.006246$$

$$f(\text{Ch}=210|\text{DV}=0) = (1/(28.2259)\Gamma(2*3.14))e^{-(210-217.2)^2/(2*(28.2259^2))} = 0.136816$$

For nominal attributes we have

$$P(1)=5/10=0.5$$

$$P(0)=5/10=0.5$$

By applying this to the final formula we get:

$$P(E|1)= 0.2895 * 0.01511 * 0.006246 * 0.5 = 1.3662*10^{-6}$$

$$P(E|0)= 0.010199 * 0.011966 * 0.136816 * 0.5 = 8.3487*10^{-6}$$

P(1) is the relative probability of the person having heart disease

P(0) is the relative probability of the person without heart disease

IF  $(P(1) > P(0))$ ,

then the Result is 1 (person has heart disease).

Else

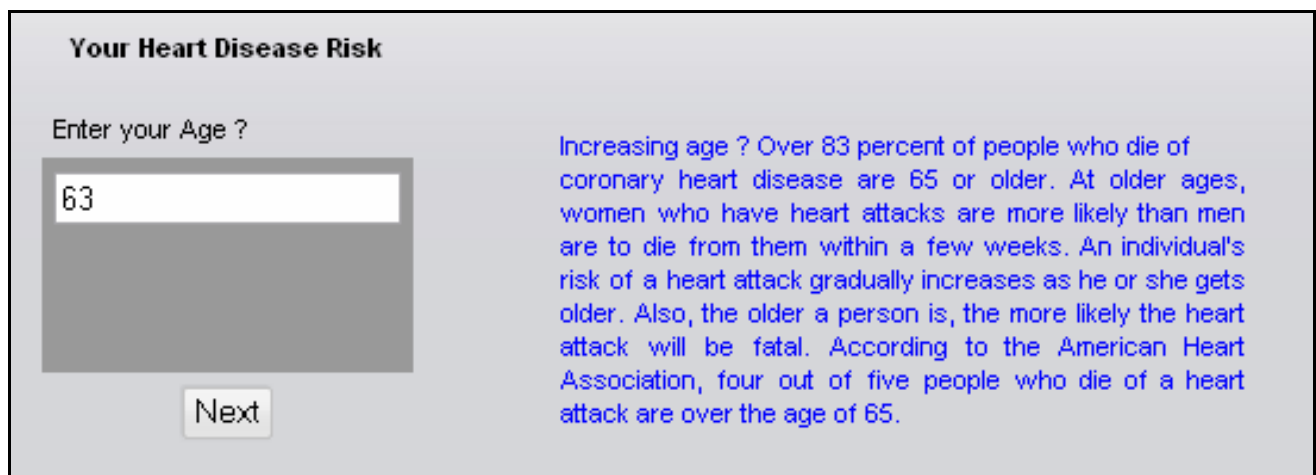
Result is 0 (person doesn't have heart disease).

Thus from the above calculation  $P(1) < P(0)$

Therefore the person has very less chance of having heart disease

### 4.3. Designing the Questionnaire

Questionnaires are cheaper and more effective over some other types of medical analytics. They require very less training from the questioner either as verbal or telephone surveys, and generally have standardized relevant answers that make it simple and efficient to compile data.



**Your Heart Disease Risk**

Enter your Age ?

63

Next

Increasing age ? Over 83 percent of people who die of coronary heart disease are 65 or older. At older ages, women who have heart attacks are more likely than men are to die from them within a few weeks. An individual's risk of a heart attack gradually increases as he or she gets older. Also, the older a person is, the more likely the heart attack will be fatal. According to the American Heart Association, four out of five people who die of a heart attack are over the age of 65.

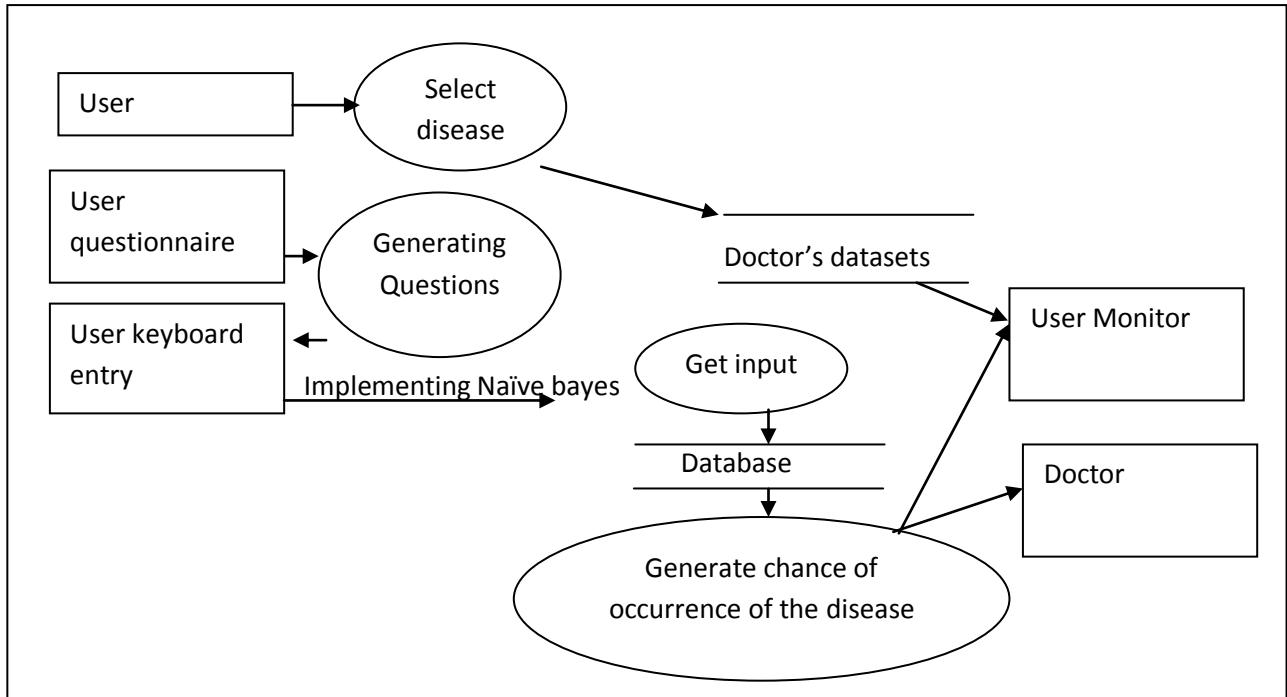
Fig.4 Sample question

### 5. Data flow diagram

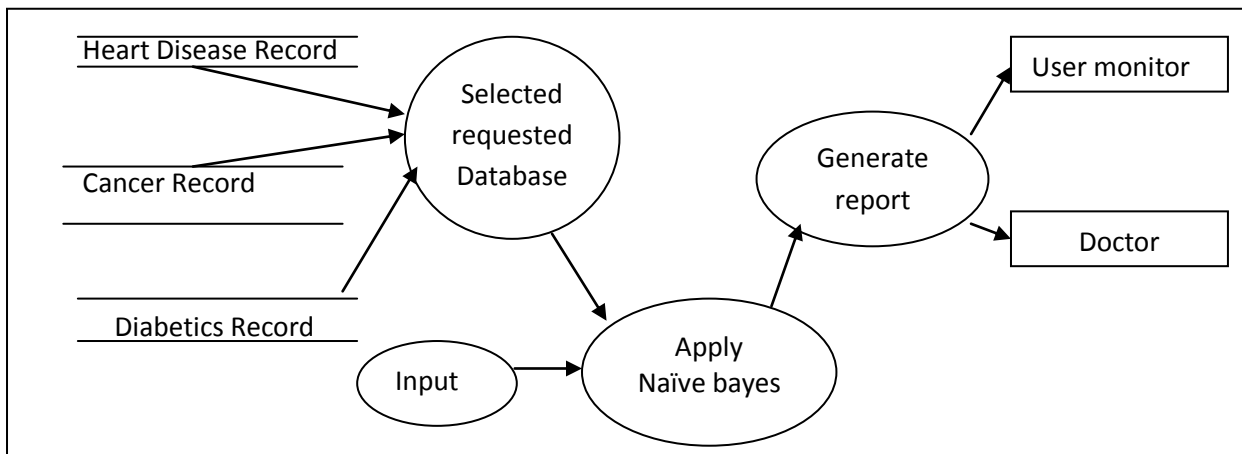
Figure 5 shows a level 1 data flow diagram which shows the user's choice and keyboard entry which is the filling of the questionnaire are the input given to the system.

As shown in Figure 5 DFD level-1 the user selects the type of disease for which the person wants to check the chance of occurrence of the disease. The questionnaire will be automatically generated by the client system according to the type of disease as requested by the client. This questionnaire will be shown to the client, in which the client can enter the data as yes or no questions. Thus the input got from the client will be used for the calculation of chance of occurrence of heart disease. Thus the generated data can be sent to the client along with the doctor's information. This generated file could be sent to the doctor if necessary.





**Fig.5 Data Flow Diagram level 1**



**Fig.6 Data Flow Diagram level 2**

Figure 6 shows DFD level 2 which is the expansion of the bubble “Generate chance of occurrence of disease”. Here the selected database and the input given by the client are taken and naïve bays classifier is applied to it. The report is generated accordingly and given to the client. This report can be dispatched to the doctor as per the client request.

## 6. Results and Discussion

The data entered from the client side, is mined employing Naïve Bayes classifier technique and the occurrence of the required disease is found. Thus the incidental of occurrence of these three diseases can be predicted employing this technique. The results are generated according to the given input. The following example depicts the obtained result in Figure 7. The required result is obtained in the .pdf format file in the server side for future reference.

**1. User Information**

**1.1. Personal Detail**

Userid	1
UserName	aa
phone No	aa

**1.2. Report Time :**  
Mon Aug 10 03:32:30 IST 2009

**1.3. Entered Details:**

Age	60
Sex	Male
Chest Pain	Asymptomatic
Blood Pressure	130
Cholesterol	206
Blood Sugar	<120
Restecg	Ventricular_hypertrophy
Thalach	132
Exang	yes
Oldpeak	2.4
Slope	flat

CA	2.0
Thal	reversible_defect
Height	154
Weight	65

**1.4. Prediction Result:**

Report	Has heart disease
--------	-------------------

**Fig.7 PDF Format Document**

**6. Conclusion**

This paper can be further enhanced and expanded. For example, it can incorporate other medical attributes. It can also consolidate other data mining techniques, e.g., Time Series, Clustering and Association Rules. Continuous data could be relatively instead of just categorical data. Another technique is to use Text

Mining by image mining to analyze and mine enormous amount of image data available in healthcare databases. A new challenge would be to assimilate data mining and text mining.

## References

1. Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", *Infection Control and Hospital Epidemiology*, 25(8), 690–695, 2004.
2. Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", *Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.*
3. Sellappan, P., Chua, S.L.: "Model-based Healthcare Decision Support System", *Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, 2005*
4. Sellappan Palaniappan, Rafiah Awang.: "Intelligent Heart Disease Prediction System Using Data Mining Techniques", *AICCSA, 2008*
5. Fayyad, U: "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases", *Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.*
6. Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", *Journal Healthcare Information Management. 16(4), 50-55, 2002.*
7. Thuraisingham, B.: "A Primer for Understanding and Applying Data Mining", *IT Professional, 28-31, 2000.*
8. S.K. Majumder, A. Gupta, S. Gupta, N. Ghosh, P.K. Gupta "Multi-class classification algorithm for optical diagnosis of oral cancer "
9. Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases",<http://mllearn.ics.uci.edu/databases/heart-disease/>, 2004.
10. Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: "CRISP-DM 1.0: Step by step data mining guide", *SPSS, 1-78, 2000.*
11. Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", *New York: John Wiley, 2003.*
12. Charly, K.: "Data Mining for the Enterprise", *31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.*

\*\*\*\*\*