School of Engineering and Applied Science (SEAS), Ahmedabad University

B.Tech(ICT) Semester IV: Probability and Random Processes (MAT 202)

- Group No : S_B3
- Name (Roll No) :
  Nisarg Patel(AU1841048)
  Vismay Patel (AU1841071)
  Dhruv Panchal(AU1841081)
  Vardhan Shah (AU1841138)
- Project Title:Heart Disease Prediction

1 Introduction

1.1 Background

Write three detailed paragraphs as instructed below.

- The heart is the main part of our body. other organs of the body work properly due to the pumping of oxygenated blood in the whole body. A heart can have some type of disease Coronary heart disease(CHD), Coronary Artery Disease(CAD), Arythima etc which has symptoms of formation of the plague or the heart doesn't beat at the proper frequency etc caused due to consumption of alcohol, smoking or family history it can also be treated by several medicinal practices like ECG, changing lifestyle etc.[1] on the initial stage, it is hard to catch the disease. Healthcare industries are making efforts for diagnosing the disease but sometimes it takes time and it can also give a faulty result which is harmful to patients as an analytic result and a result which are predicted using the probabilistic method will helpful to configure and to make a decision by a person.around 17.9 million people die due to heart disease given by WHO.even through considered this disease as chronic but it can be avoided by regular checkups and test. [2]
- There are different no of researchers and scholars come up with different method and algorithm to analyze the disease, and to provide treatment accordingly. Humans beings are unable to arrange data if it's huge in size therefore such as data mining technique and data storing technique are used.[3] the problem to recognize the real relationship between heart disease and it's factor affecting is difficult to know therefore a dataset is used to classify a relation.[4]most of the researchers have used the dataset provided by courtesy of the Cleveland heart disease database[5] via UCI machine respiratory.[6,7].the data has been understood by correlation matrix, bar plot, histogram and then categorise variable into

columns.[8]MA. Jabbar, Priti Chandra, B.L.Deekshatulu proposed genetic algorithm for heart disease prediction.[9] Prediction of heart disease using the artificial neural network by two perceptrons one is sum function and another is the transfer function.[10] the author on its knowledge in field of biology select main features which to be applied on SVM, Naive Bayes, decision tree classification model. Avinash Golande mentions about the accuracy of the author's work with different machine learning techniques.[11]. Improving the accuracy of prediction of heart disease based on ensemble classification which improves the reliability of the weak algorithm and it's also focused on for the application to the medical dataset.[12]

- We have used the Bayesian network which is an acyclic graph where the nodes and connections are decided by author experience in biology and conditional probability table is evaluated by a dataset from another project.[13]Bayes theorem gives way to calculate the posterior probability of Parameter of heart disease prediction system with strong naive independence which called as class conditional independence.[14] We propose a Naive Bayes which is a straightforward method for developing classifiers: they figure the probability of every classification for a given sample and afterwards yield the class with the most noteworthy one. The way they get these probabilities is by utilizing Bayes' Theorem, which portrays the probability of an element, in light of earlier information on conditions that may be identified with that too future.[15].The Dataset has been taken from Cleveland UCI and further dividing data in test and train set by 80:20 ratio the confusion matrix for naive Bayes has been plotted to observe the output of predicted values and actual values.[7].by doing they have their original belief with addition to observation yield the new belief.

1.2 Motivation

Heart disease diagnosis is a complex task which requires highly expert professionals but doctors are not in extent with the populace. likewise, symptoms are ignored at an early stage so therefore diagnosis disease at an early stage becomes difficult. As the health care industry is "data-rich" which can't be handle manually. These a lot of information are significant to separate valuable data and create connections among the properties.through which we can develop a mechanism through data which helps along with medical treatment and diagnosis.

1.3 Problem Statement/ Case Study

- Sometimes it's difficult to reach to doctor and pretty costly also. The mild symptoms are Ignored by persons, therefore, a user requires a consultation to become aware or to change their lifestyle accordingly. The healthcare industry is a data-rich industry but less knowledge industry to utilize the data however to find link and pattern between the causes and symptom become helpful for better diagnosis. So heart disease prediction helps for the prediction of the heart diseases, symptoms and treatment required.

2 Data Acquisition

- No, our Special Assignment is not data-dependent.
- The Conditional probability Table is given in our base article by the data acquisition process performed by the author using the data collected during the research project entitled "Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina."

3 Probabilistic Model Used/ PRP Concept Used

- In the prediction of heart diseases we are using a bayesian belief network and the concepts of joint and conditonal probablity.

  Here we are considering 17 random variables in which there are 13 parent variables such as atherosclerosis,High BP, Family History,Serum selenium, adverse medicine etc and 4 child variables such as ECG,Angina Pectoris, Myocardial Infraction and Rapid Heartbeats.The probability of child variables and probability of heart diseases will differ if we increase or decrease the number of random variables. Here we consider only two states of the random variables either 0 or 1 where 1 depicts the good state and 0 depicts the bad state of the random variables. We take only two states because it is easy to expand the model if we add other parameters and it keeps the model simple because if we consider all the possible combinations of all the parameters the model will be more complex.

  The fundamental concept in bayesian networks is that probabilities can be assigned to the parameter values and through bayes theorem these probabilities can be updated given the new data.For the parameters without any parent prior probabilities for various states were entered.For the variables with one or more parents we calculate the posterior probabilities according to the conditional probability tables given and using joint probability distribution.

  In our Model we take we take 6 input variables which are all independent and we derive probabilities other parameters based on these inputs and by using bayes theorem and joint probability distribution.After that we calculate probability of heart diseases similarly and from this probability we derive 4 inferences like ECG,Heart Rate,Angina Pectoris and myocardial infraction. We know that Bayes theorem is given by:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

  The Joint probability of the variables can be calculated as

  The general expression is as:

$$P(x_1, x_2, ..., x_n) = \Pi_1^n P(x_i|Parent(X_i))$$

Now let us consider three random variables A,B,C where C is dependent on A and B.

Here A and B are independent random variables.And as C is has two parents the total combinations possible are 4.

So we can calculate the probability of C by using joint probability distribution.

$$P(C) = P(C|A,B)P(A)P(B)+P(C|-A,B)P(-A)P(B)+P(C|A,-B)P(A)P(-B)+P(C|-A,-B)P(-A)P(-B)$$

The posterior probabilities such as P(C/A,B) etc can be calculated through the conditional probability tables.

Similarly if a random variable has 'k' parents then the probability of that variable can be calculated in the similar manner but in this case there will be 2^k combinations and all the combinations will be given in the conditional probability table.

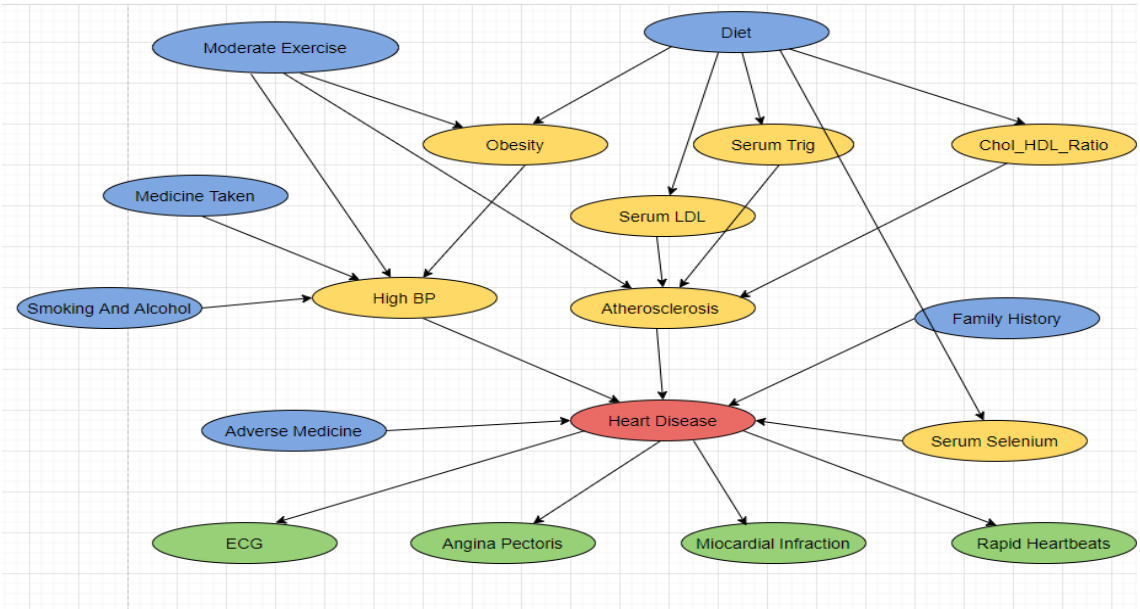

Fig: Block Diagram

4 Pseudo Code/ Algorithm

- **Data given**

  → Conditional Probability Table evaluated for parameter's. for 2^n different cases where no of child variable it depend on.

- **Mathematical Representation**

  →Finding the conditional probability of parent variable with children to derive result which parameter will affect or not.

  → $P(x_1, x_2, ..., x_n) = \Pi_1^n P(x_i|Parent(X_i))$

  Part 1

  Function : decimal to binary

  INPUT : number int,n int ***no of variable***

  OUTPUT : int array[] ***binary no***

  array[2^n]

  cnt = 0

  while(number>0 and cnt<n)

  Arr[cnt] = number%2

  number=number/2

  Cnt = cnt+1

  Return array

  Part 2

- For finding conditional probability of parameter

  INPUT :n int ***no of parent variable***,array float[] ***parent probability due to child variables*** ,array float[] probability array of 2^n cases

  OUTPUT :answer float ***joint conditional probability***

  array[2^n]

  for i in 1..2^n

  array = decimaltobinary(i,2^n)

  answer = answer + child_variable1[array[2^n-1]]*child_variable2[array[2^n-2]]...child_variablen[array0]*parent_variable[i]

Return answer

Loop Iteration's

INPUT : n =2,MEXC[],EXEC[],Obesity[]

OUTPUT :

Table . Conditional probability of Obesity given Moderate Exercise and Diet.

Abbreviation Used: Obesity (OB), Moderate Exercise (MEXC), Diet (DIET) , Probability (P).

States Used: MEXC(No, Yes), DIET (Bad, Good).

MEXC	DIET	P(OB)	P(-OB)
NO	BAD	0.6	0.4
NO	GOOD	0.1	0.9
YES	BAD	0.1	0.9
YES	GOOD	0.05	0.95

when i=0

array=decimaltobinary(0,2); i.e array={0,0}

ans = ans + MEXC[0]*DIET[0]*Obesity[0]

when i =1

arr=decimaltobinary(1,2); i.e arr[] = {0,1}

ans = ans + MEXC[0]*DIET[1]*Obesity[1]

when i =2

arr=decimaltobinary(2,2); i.e arr[] = {0,1}

ans = ans + MEXC[1]*DIET[0]*Obesity[2]

when i=3

arr=decimaltobinary(3,2); i.e arr[] ={1,1}

ans = ans + MEXC[1]*DIET[1]*Obesity[3]

Simillarly we can calculate probabilities of other parameter dependent on distinct variables.

5 Coding and Simulation

5.1 Simulation Framework

In our Framework mainly six controlling parameters are taken into consideration which are taken as input from user. Also they are independent of each other. Mainly the values are taken in such a way that it represents the probability of that particular parameter. Here for example let say one of the parameter is probability of smoking and alcohol, so the input value of this is taken in such a way that how frequent is he/she does smoke or take alcohol if value is near to 1 depicts that he/she is more frequent addicted to smoking and alcohol. Similarly value nearer 0 implies that he/she does rarely consumption or alcohol or smokes.Likewise the values of all others controlling parameters are taken. Further there are 7 parameters which are dependent on any of these 6 controlling parameters which are used to derive final chances of heart related disease.

5.2 Reproduced Figures

- The tool used is JAVA for coding which includes Jframe for EndUser input and JFree chart to generate graph.
- Reproduced Figure-1

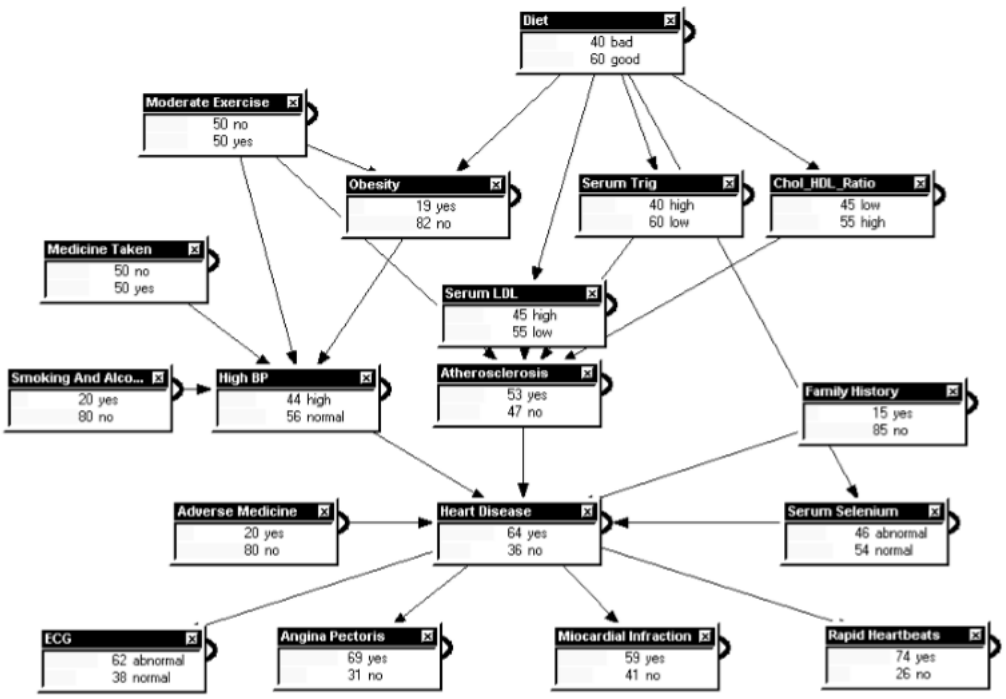

Fig: 1B

The Figure(1B) depicts the Probability of heart disease. Thus the model calculated calculated the probability of heart disease based on the values of input probability.

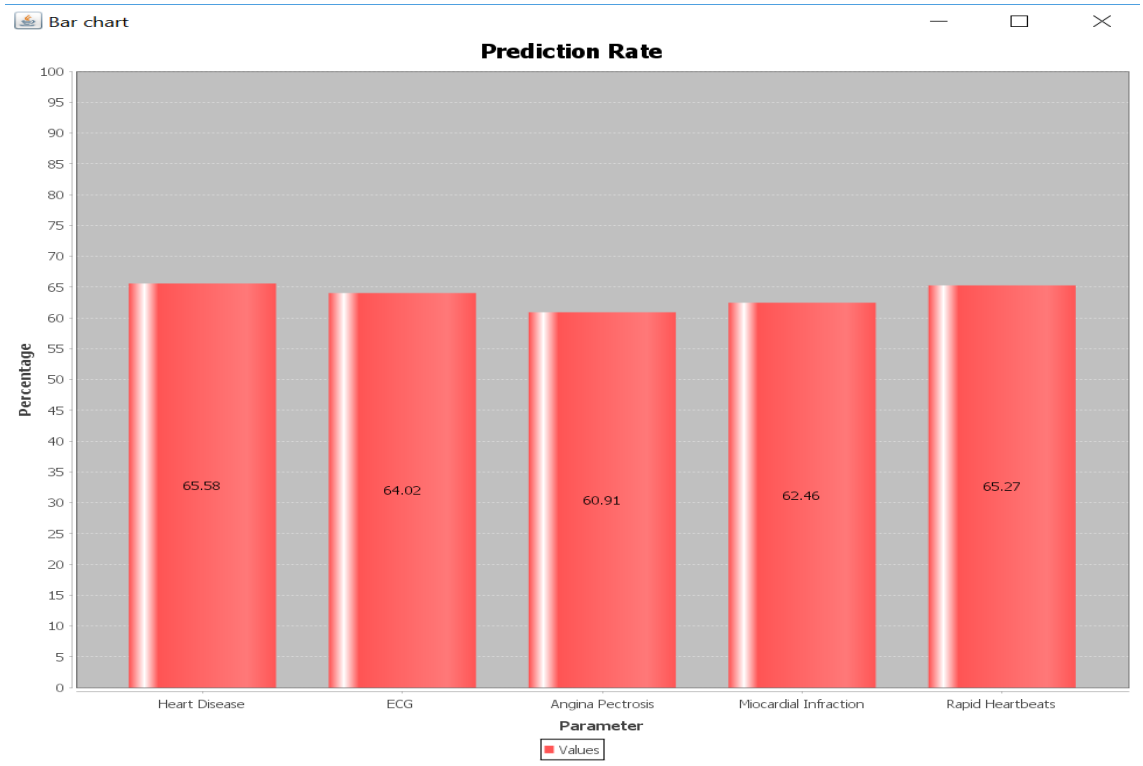

Fig: 1R

This graph shows the plot of the percentage versus parameters of heart diseases. On y-axis we are considering the percentage and on x-axis we are considering the parameters of heart diseases.

Inference Derived : When probability of medicine taken is 0.5 , smoking and alcohol is 0.2 , family history of having heart disease is 0.15, adverse medicine taken is 0.2, diet is 0.6 and exercise is 0.5 we found that the percentage of heart disease is approx 65 percent which subjects to a moderate risk of heart disease.

- Reproduced Figure-2

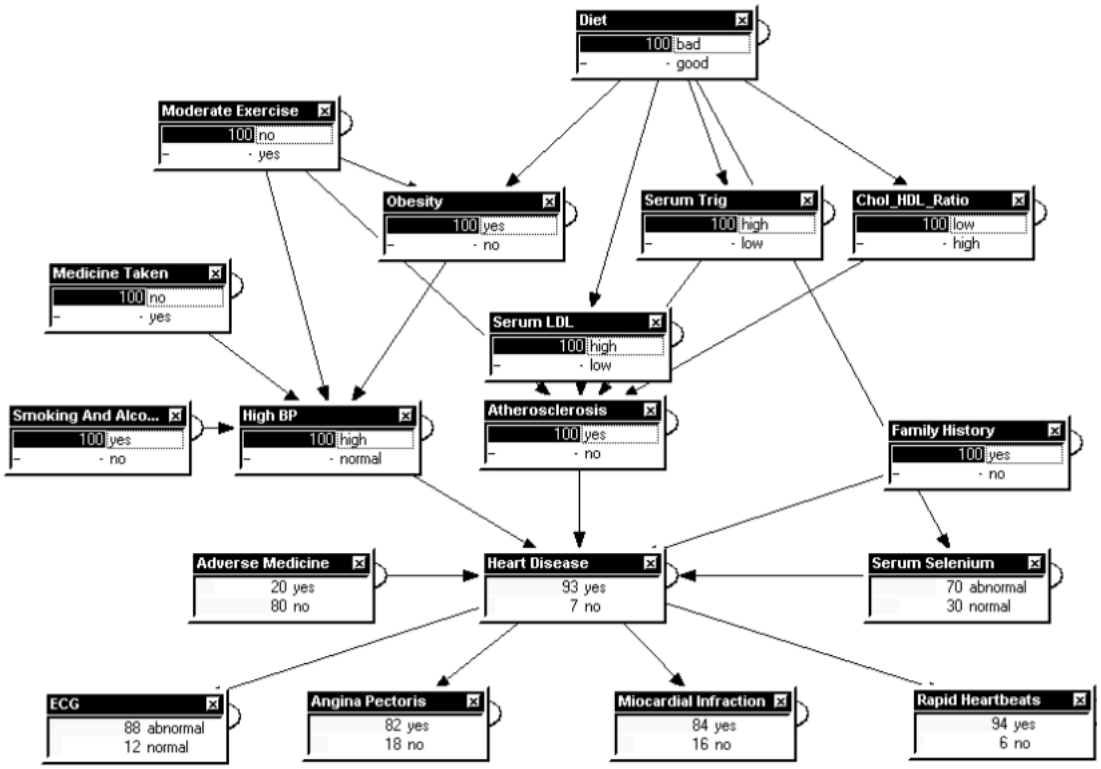

Fig: 2B

In the figure 2B all other controlling parameters except adverse medicine and serum selenium are 100 percent.

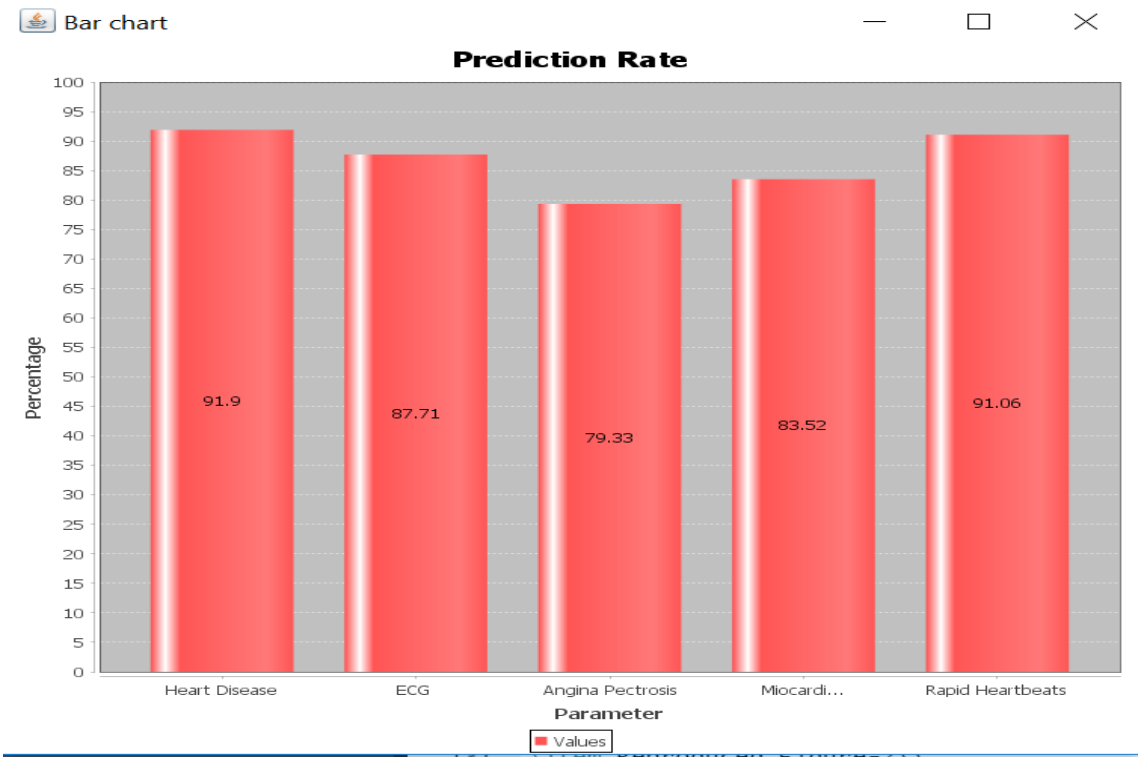

Fig: 2R

Inference derived : When all the inputs were at max then the probability of the heart disease was updated from 65 percent to 91 percent.

- Reproduced Figure-3

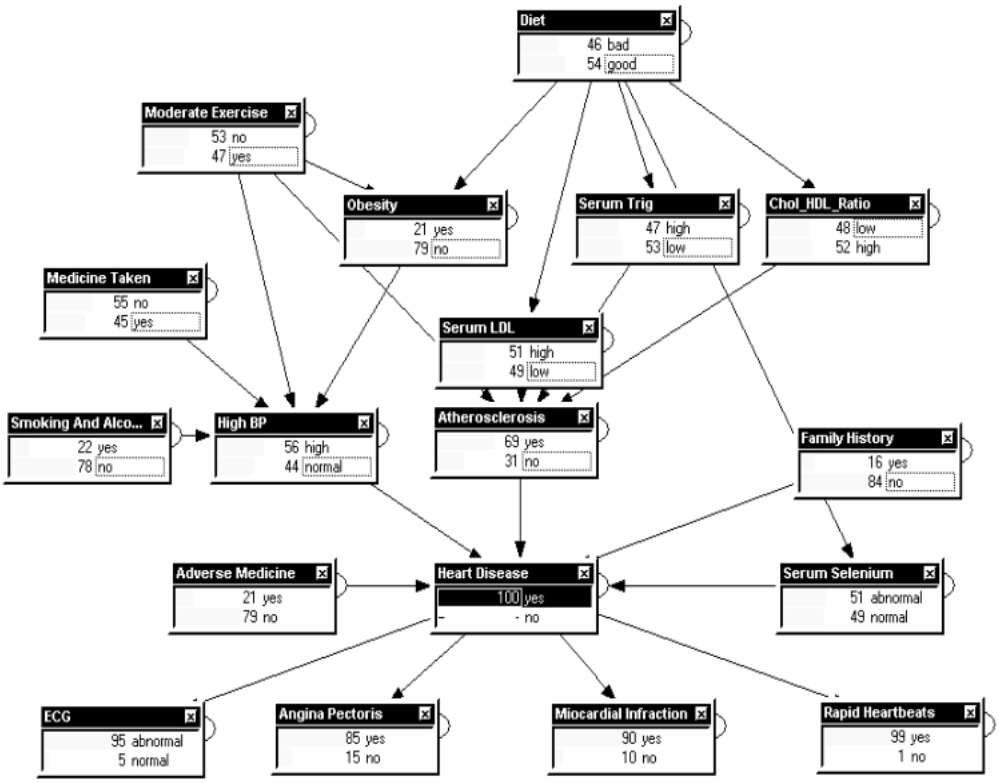

Fig: 3B This figure shows the probability of the child parameters of the heart disease when heart disease is 100 percent.

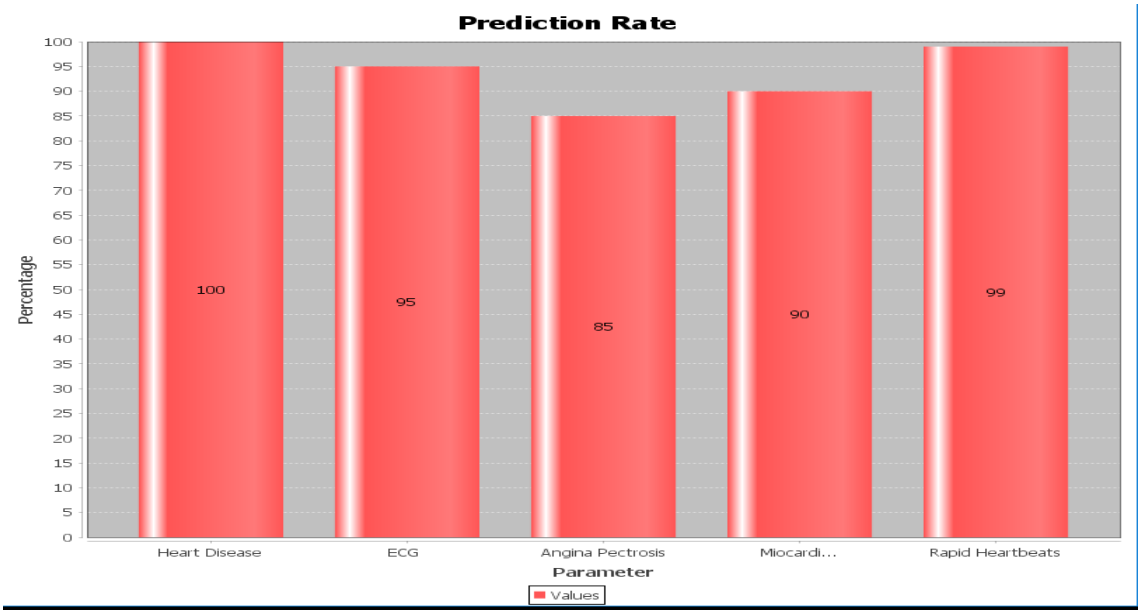

Fig: 3R

Inference Derived : When the heart disease was at max the child parameters ECG was at 95 percent , angina pectrosis was at 85, Miocardial infractions was at 90 and rapid heartbeats was at 99. So by the above observations we can say that ECG, Miocardial infractions, rapid heartbeats are the most sensitive tools to take care off.

- Reproduced Figure-4

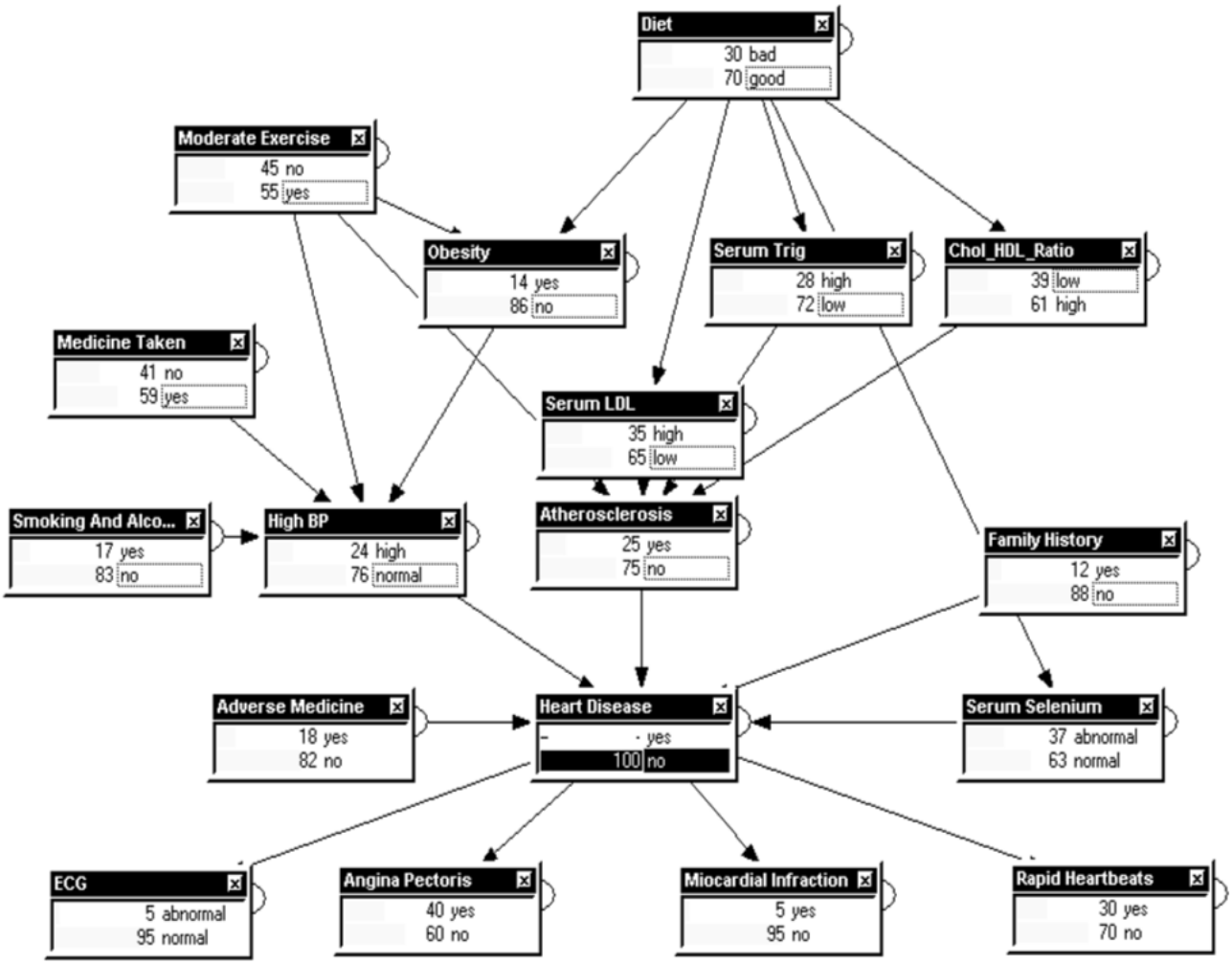

Fig: 4B This figure shows the probability of the child parameters of the heart disease when heart disease is absent.

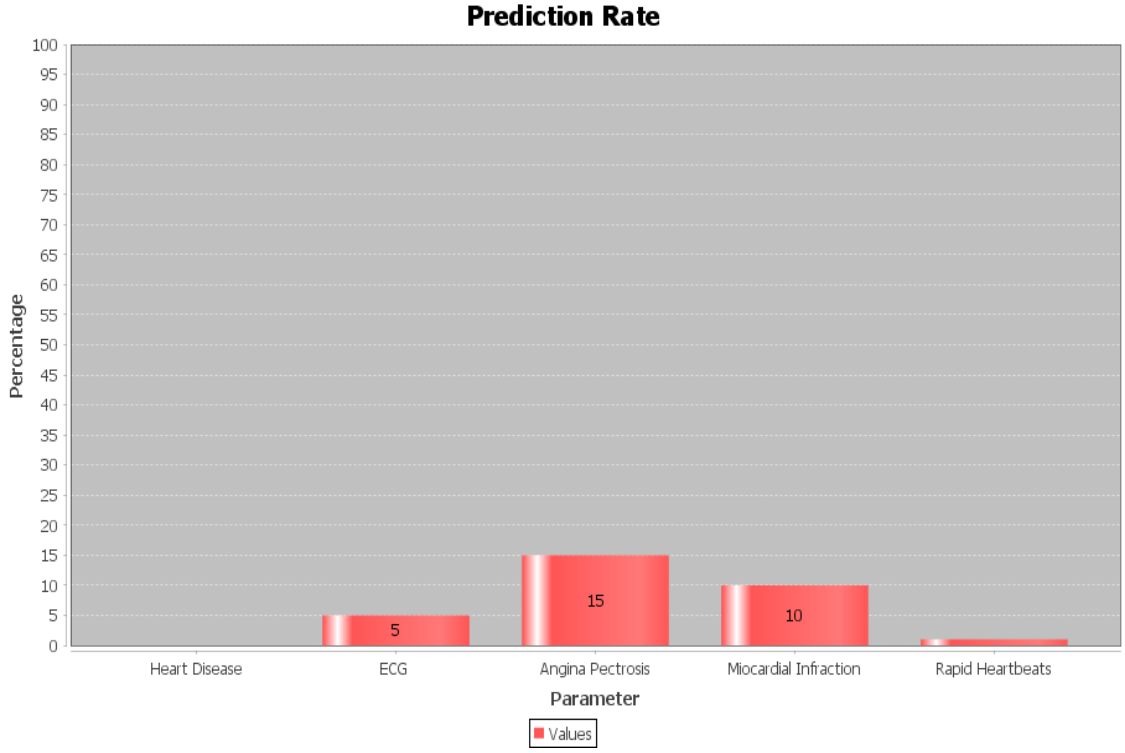

Fig: 4R

Inference Derived : When the heart disease was at zero the child parameters ECG was at 5 percent , angina pectrosis was at 15, Miocardial infractions was at 10 and rapid heartbeats was near to zero. So by the above observations we can say miocardial infractions and ECG gets more affected.

- Reproduced Figure-5

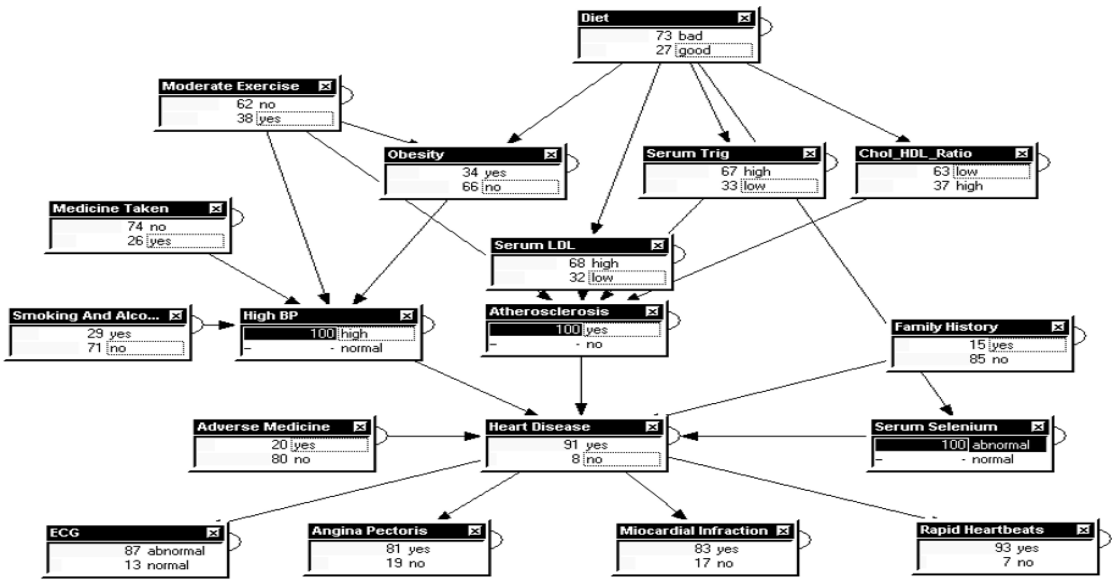

Fig: 5B

This figure shows the probability of the heart disease when high bp, atherosclerosis and serum selenium are 100 percent.

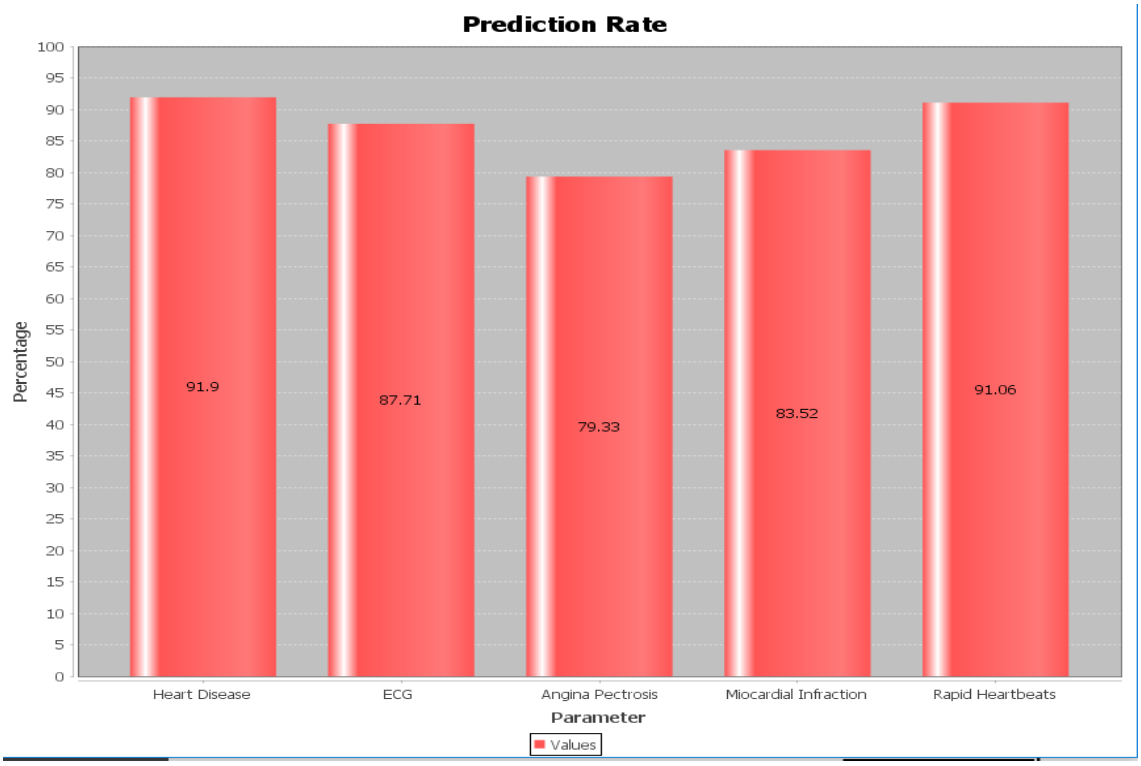

Fig: 5R

Inference Derived : In this case the probability of the heart disease was found to be 92 percent.

- Reproduced Figure-6

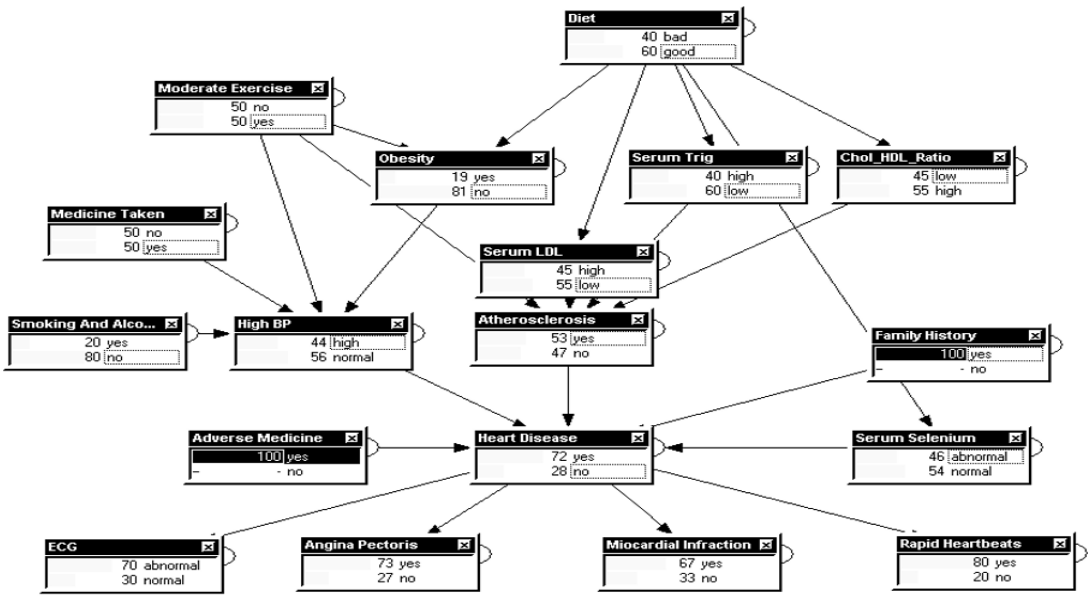

Fig: 6B

This Figure shows the probability of heart disease when adverse medicine and family history were 100 percent.

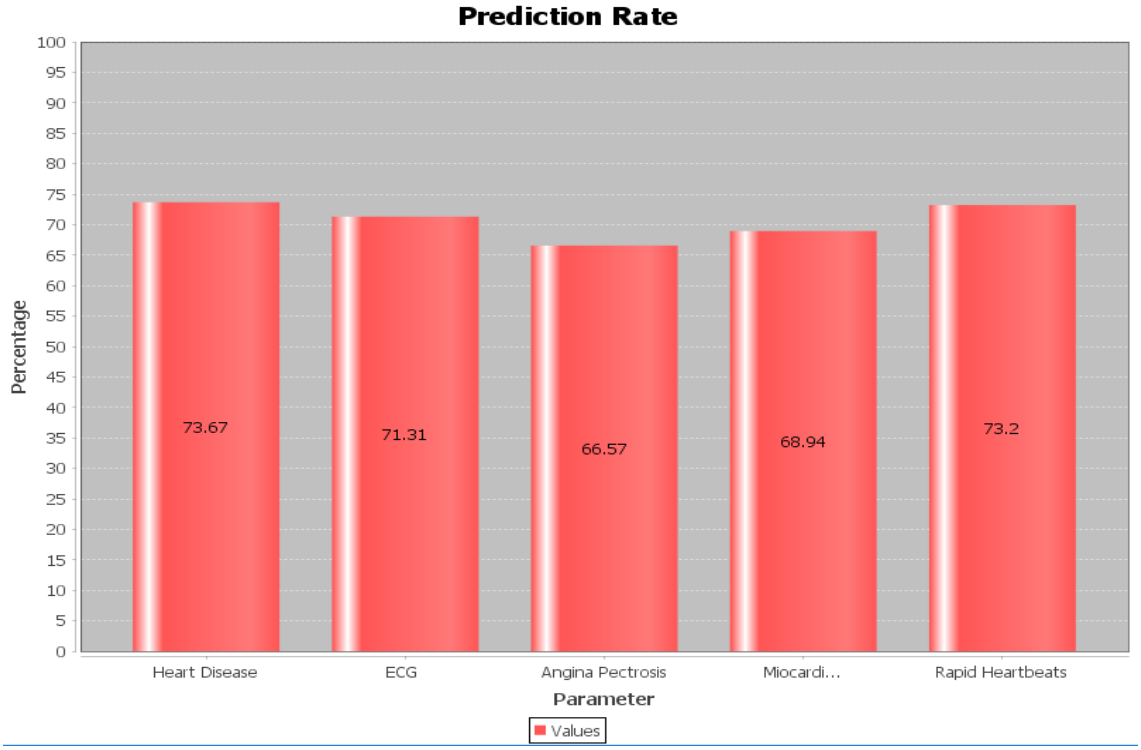

Fig: 6R

Inference Derived : In this case the probability of the heart disease was found to be 73.6 percent.

6 Inference Analysis/ Comparison

- to find the dependency of beleif between the parameter we have generated 7 cases which shows change in probabilitiy the all predecessor of heart disease are set high except adverse medicine and serum selenium is kept normal of adverse health case which shows high risk of heart disease and adjoining we further calculate heart disease probability of a person with good health. the output is expected as of low risk of heart disease but the two factor adverse medicine and serum selenium kept so low risk in second identifies that this two parameter doesn't contribute much in heart disease hence this factor only should be seen when advance diagnosis is to be taken.

  In case of heart disease present, all symptoms and treatment show significance high value but in absence of heart disease the Angina Pectoris and Rapid Heartbeats does not change accordingly with heart disease hence the ECG and Myocardial infarction is specificto heart disease and rapid heart beat and angina pectoris are not sensitive,while ECG is very sensetive can be used as diagonstic tool .our model has also some limitation since it's not considering the inhibitor factor which lower down the effect hence much change is not detected.

  The Parent Variable of Heart Disease are Adverse Medicine,High BP, Atherosclerosis,Family History,Serum Selenium The Heart disease show significant high value when Serum Selenium,Atherosclerosis and High BP are high.but when adverse medicine and family history are at highest value bu the heart

disease parameter probability is quite low compare to earlier.Adverse Medicine and Family History do not contribute significantly towards heartdisease as compared to Atherosclerosis, High BP and Serum Selenium.

- By the help of our heart disease prediction modelling the doctors can use it to predict the possibility of whether a person will have heart disease or not by simply inserting the probability of various parameters that a person do in their daily life or by inserting the probability of other parameters like cholestrol level etc from their report and if the probability of heart disease is more than 0.6 then the doctors can treat and prescribe the medicines to the patient accordingly.Doctors can also see the major affecting factors and they can concentrate to cure that factor affectively.

7 Contribution of team members

7.1 Technical contribution of all team members

Enlist the technical contribution of members in the table.

Tasks	Nisarg	Vismay	Dhruv	Vardhan
Modelling	Done	Done	Done	Done
Coding	Done	Done	Done	Done
Inference	Done	Done	Done	Done

7.2 Non-Technical contribution of all team members

Enlist the non-technical contribution of members in the table.

Tasks	Nisarg	Vismay	Dhruv	Vardhan
Mathematical Work	Done	Done	Done	Done
C-Map	Done	Done	Done	Done
Abstract writing	Done	Done	Done	Done
Report Writng	Done	Done	Done	Done

References

[1] "Glossary of Heart Disease Terms," , Jan. 2009. Accessed on: Feb. , 2020. [Online]. Available: https://www.webmd.com/heart-disease/glossary-heart-disease-terms

[2] "Cardiovascular Diseases,". Accessed on: Mar. , 2020. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[3] ShusakuTsumoto," Problems with Mining Medical Data", 0-7695- 0792-1 I00@ 2000 IEEE.

[4] Carlos Ordonez, "Improving Heart Disease Prediction Using Constrained Association Rules," *Seminar Presentation at University of Tokyo*, 2004.

[5] Aha, D., and Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database." *University of California* 3.1 (1988): 3–2.

[6] Margaret Wanjiru,"Data Science For Good — Machine Learning for Heart Disease Prediction" , Aug 19,2019 Accesed on Feb. ,2020.[Online]. Available :https://medium.com/@wanjirumaggie45/data-science-for-good-machine-learning-for-heart-disease-prediction-289234651fed

[7] Shubankar Rawat, "Heart Diseases Prediction," for *Cleveland Heart Disease(UCI Repository) dataset — classification with various models.*,2019

[8] Karan Bhanot,"Predicting presence of Heart Diseases using Machine Learning" ,Feb 13,2019 Accesed on Feb 2020, Available on :https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c

[9] MA.jabbar, Priti Chandra, B.L.Deekshatulu,"Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm",7 May 2015

[10] Shahid Mehmood Awan,Muhammad Usama Riaz, Abdul Ghaffar Khan," Prediction of Heart Disease using Artificial Neural Network" ,October 2018

[11] Avinash Golande, Pavan Kumar T,"Heart Disease Prediction Using Effective Machine Learning Techniques",June 2019

[12] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informatics in Medicine Unlocked,* 02-Jul-2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291481830217X. [Accessed: 09-Mar-2020].

[13] Jayanta K. Ghosh and Marco Valtorta , "Probabilistic Bayesian Network Model Building of Heart Disease[1] ,"

Dr. Kailash Mathu"Cholesterol, Selected Minerals and Health Status of the Elderly in South Carolina." ,Nov 30 1999.

[14] K.Vembandasamy R.Sasipriya and E.Deepa "Heart Diseases Detection Using Naive Bayes Algorithm",*International Journal of Innovative Science, Engineering Technology, Vol. 2.*9 Sep 2015.

[15] Mrs. N.Deepa,"HEART DISEASE PREDICTION SYSTEM USING NAIVE BAYES" *International Journal of Pure and Applied Mathematics Volume 119.* Nov 16,2018