Graphical User Interface developed for Diseases Prediction by Mean of Clustering and Apriori Algorithm

Shimpy Goyal
Student
Department Of Computer Science & Applications
Maharishi Dayanand University Rohtak 1240001
(Haryana) India

Rajender Singh Chhillar, PhD
Professor
Department Of Computer Science & Applications
Maharishi Dayanand University Rohtak 1240001
(Haryana) India

ABSTRACT
Disease prediction is one of the most important issues that we are facing today. A large number of patients struggling for their check up even for predictive disease like heart attack possibilities, kidney damage change and possibilities of lung problem. All these lies in predictive disease categories. They need not require very vast analysis if we can predict. This Research motivate to develop a console(GUI) on the basis of data mining which is used to analyze large volumes of data and extracts information that can be converted to useful knowledge. And overall predict a patient for their chances of disease. These techniques can be applied on predictive medical disease. This research papers which mainly concentrated on predicting kidney failure, heart disease. Experimental results will show that many of the rules help in the best prediction of heart disease and kidney failure which even helps doctors in their diagnosis decisions by using A-prior and k-mean algorithm. By the help of this algorithm it provide easy and efficient way in which we can find the stage of the kidney failure and heart disease. To swamp this problem the healthcare industry gathers enormous amounts of heart disease data which, grievously, are not "mined" to discover hidden information for effective decision making. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. So due to these condition even doctors not able to predict disease accurately. So there is need to develop a efficient decision making system which can predict the correct diseases with available data. So in this paper we are introducing the automated console to predict the diseases by mean of clustering & a-priori algorithm. This is web based convenient tool it can be used even in absence to doctors to predict diseases. Here, we consider almost 200 persons data to develop this automated console. Preliminary conclusions shows that it very effective tool to predict diseases.

Keywords
Data mining, kidney failure, heart disease**,** A-prior and k-mean Algorithm.

1. INTRODUCTION
Knowledge discovery in databases is well-defined process consisting of several distinct steps. Data mining is the technology, from which the data extracted to form useful information and can make a decisive work for user from huge databases. Quality of service resembles diagnosing patients correctly and administering treatments that are effective. Poor quality diagnosis decisions can lead to patient in danger which are therefore unacceptable. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care.

The availability of integrated information provide a good decision making analysis over the row data, there is a shift in the perception of clinicians view and patients diagnosis system from analytic study and visualization of diagnosis data for quantitative assessment of information with the supporting of all clinical diagnosis report data. Further it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Kidney failure and heart disease applies to a number of illnesses that affect the body at circulatory system, In that part heart and blood vessels resides. It has to deal only with the condition and the factors, which leads to such condition. Acute kidney injury means that your kidneys have suddenly stopped working. It is also called acute renal failure. Your kidneys remove waste products as well it helps balance water and salt and other minerals in blood. When it is being stopped. Waste products, fluids, and electrolytes build up in your body. This will greatly affect the body and make it quick illness and the condition might leads to life about to death.

1.1 The k-means algorithm:
The k-means algorithm is a simple iterative method to partition a given data set into a specified number of clusters, k. This algorithm has to be developed by several researchers and scientists across different disciplines. This algorithm operates on a set of d-dimensional vectors, $D = \{xi \mid i = 1. . . N\}$, where $xi \in Rd$ denotes the i th data point. This algorithm is initialized by picking k points in Rd as the initial k cluster representatives or "centroids". Techniques for selecting these initial seeds include sampling at random from the data set, collecting and make setting up them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. The simple way to understand K-means is:

- This Requires real-valued data.
- This include number of clusters present in the data.
- It results great when the clusters in the data are of approximately equal size.
- Attribute significance cannot be calculated
- Lacks explanation capabilities.

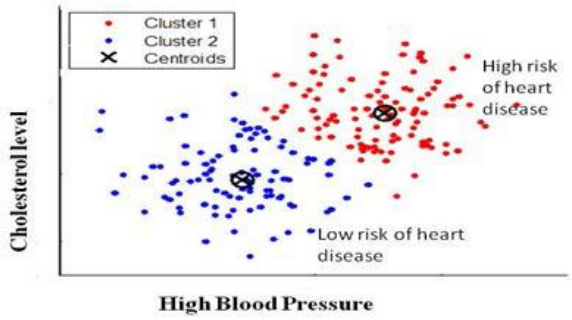


Fig 1. K-means Clustering for Heart Disease Patients

1.2 The Apriori algorithm

Apriori algorithm for association is proposed by R.Agarwal., in 1994. It finds out the relationships among item sets using two inputs-support and confidence. One of the most popular data mining approaches is to find frequent itemsets from a transaction dataset and derive association rules. Finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or item set are sorted in lexicographic order.

2. REVIEW OF LITERATURE

World health organization [1] presented a the ten leading causes of death by broad income group 2008.

Vikas Chaurasia *et al*.[2]"The objective of this research work is to predict more accurately the presence of heart disease with reduced number of features. Originally, thirteen attributes were involved in predicting the heart disease. Thirteen features are reduced to 11 features. Three classifiers that is Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. The empirical results show that they can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of incoming patients. This study will also work to identify those patients which needed special attention for treatment.

My Chau Tu *et al* [3] presented a The diagnosis of heart disease is important issue is to prompting many scientist to work on development of intelligent console for medical decision support systems to improve the ability of physicians.

M.Akhil jabbar *et al* [4] presented Experimental Results show that most of the classifier rules help in the best prediction of heart disease which even helps doctors in their diagnosis decisions.

N. Aditya Sundar *et al*[5] presented a training tool to train nurses and medical students to diagnose patients with heart disease. It is based web based user friendly application and can be used in hospitals if they have a data ware house for their hospital.In this paper we study the performances of the two classification data mining techniques by using various performance measures.

C Y HSU *et al* [6] presented Few studies have defined how the risk of hospital-acquired acute renal failure varies with the level of estimated glomerular filtration rate

Mohammed Abdul Khaleel *et al*[7]. Presented methodology to discover locally frequent diseases with the help of Apriori data mining technique.

Chris Ding *et al* [8] presented a Mapping data points into a multidimensional dimensional space via kernels technique, we show that solution for Kernel K-means is given by Kernel PCA. Through this learning algorithm, our results suggest effective techniques for K-means clustering. DNA gene expression and news groups are analyzed analytical illustrate the results.

K.R. Lakshmi *et al* [9] presented a The diagnosis of heart disease is a significant and tedious task in medicine. The study describes algorithmic discussion of the heart disease data set from Cleveland Heart Disease database, on line repository of large data sets.

Boshra Bahrami *et al*.[10] evaluate different classification techniques in heart disease diagnosis. Classifiers like J48 Decision Tree, K Nearest Neighbors(KNN), Naive Bayes(NB), and SMO are used to classify dataset. After classification, some performance evaluation measures like accuracy, precision, sensitivity, specificity, F-measure and area under ROC curve are evaluated and compared. The comparison results show that J48 Decision tree is the best classifier for heart disease diagnosis on the existing dataset.

3. PROPOSED ALGORITHM FOR GRAPHICAL USER INTERFACE BASED PATIENT MONITORING SYSTEM

By using the techniques, Apriori and K-means the heart and various other disease rate prediction is performed on the patient databases.

K-means

Step 1: Each data point is assigned to its closest centroid, with each broken randomly with closest cluster.This results in a partitioning of the data.

Step 2: Each cluster representative is relocated to the center (mean) of all data points assigned to it. If the data points come with a weighted probability, then the relocation is to the expectations (weighted mean) of the data partitions.

Apriori

Let the set of frequent item sets of size k be F_k and their candidates be C**k**. Apriori first scans the database and searches for frequent item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent item sets.

1. Generate Ck+1, candidates of frequent itemsets of size k+1, from the frequent itemsets of size k.

2. Scan the database and calculate the support of each candidate of frequent itemsets.

3. Add those item sets that satisfies the minimum support requirement to Fk+1.

Function apriori generates Ck+1 from Fk in the following two step process:

1. Join step: Generate RK+1, the initial candidates of frequent itemsets of size k + 1 by taking the union of the two frequent itemsets of size k, Pk and Qk that have the first k−1 elements in common.

RK+1 = Pk ∪ Qk ={item1 , item2, . . . , itemk−1, itemk , itemk' }

Pk = {item1 , item2, . . . , itemk−1, itemk }

Qk = {item1 , item2, . . . , itemk−1, itemk' }

where, item1 < item2 < · · · < itemk < itemk'_ .

2. Next step: Check if all the item sets of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this requirement from Rk+1. This is because any subset of size k of Ck+1 that is not frequent cannot be a subset of a frequent item set of size k + 1. Function subset finds all the candidates of the frequent item sets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most kmax+1 times when the maximum size of frequent item sets is set at k-max.

4.EMPRICAL EVALUATION OF PROPOSED ALGORITHM TO FIND ANALYTICAL RESULT OF PREDICTION

We are integrating both of above to find more analytical result of prediction

Assignment 1

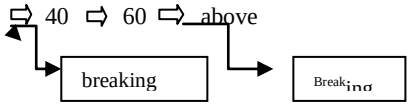
Disease is categorized on the basis of following factors.

1) Age Scaling Factor.
2) Sex Probability Factor
3) CP Scaling Factor
4) Bp Scaling Factor
5) Chol Scaling Factor
6) Fbs Probability Factor
7) Res Probability Factor
8) Thal Scaling Factor
9) Exang Probability Factor
10) Ca Null

Range

Age ⇨ 20 100
Lowest Highest

Age Follows two breaking point.



Breaking Point

a) 0to 40⇨ very less factor ⇨ NIL
b) 40to 90 ⇨ Average
c) 90 to above⇨ high

40 to 60 ⇨ linear growth.

60 to above ⇨exponential growth

Linear growth= 100(Highest)/ 100 –Age (Age 100)

Must valid before 90.

TABLE1. A General Occurrence of a Particular Case of Problem

	A_1 A_2 A_3	C_1	B_1	C_2 C_3	T_1
T_1	x	x	x	x	
T_2	x	x	x		x
T_3	x		x	x	
T_4	x	x	x	x	x
T_5	x	x		x	
T_6	x		x		
T_7	x	x	x	x	x
T_8	x	x	x	x	x
T_9	x	x		x	x
T_{10}	x	x	x	x	

Implementing A- priori

Age ⟶ { A_1 A_2 A_3 }

Cp ⟶ { C_1 }

Bp ⟶ { B_1 }

Chol ⟶ { C_2\C_3 }

Thal ⟶ { T_1 }

1) Exponential growth = $100 * e^{0.1*Age}/10$

$= 10* e^{0.1*Age}$

Weight Code Factor

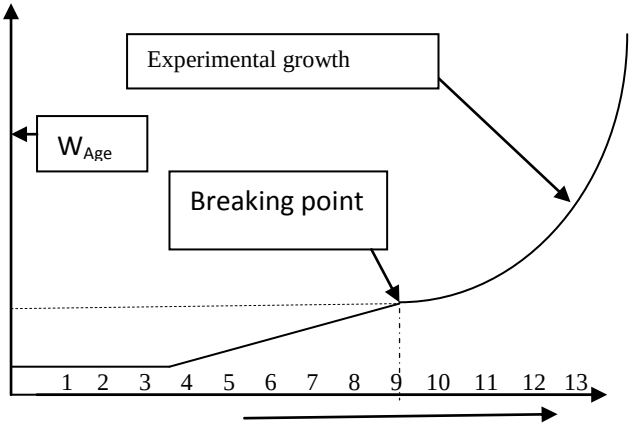


Fig 2. Weight versues Age

0 to 40 40- 90 90-100

$[^{W}_{A}]^{40}_{0}$ $[^{W}_{A}]^{90}_{40}$ $[^{W}_{A}]^{100}_{90}$

Fixed Linear Exponential

W_{AL} 1 1.6to 10/2 (division) 24.59to 26.91

Division of 3

2) Sex (Male/Female)

Male ⟶ 0.7

Female ⟶ 0.3

$B_{1=}$ { T_1 T_2 T_3 T_4 T_5 T_6 T_7 T_8 T_9 T_{10} }

$C_{1=}$ { T_2 , T_4, T_5, T_7 , T_8 , T_9 , T_{10} }

C_2 = { T_1 , T_4 , T_7 , T_{10} }

C_3 = {T_3 , T_5 , T_8, T_9 }

$T_{1=}$ {T_2 , T_4, , T_7 , T_8 , T_9 }

$C_{1\cap}$ $C_{2=\varphi}$

$T_{1=}$ {T_2 , T_4, , T_7 , T_8 , T_9 }

$C_{1\cap}$ $C_{2=\varphi}$

3) CP

Linear coding

Linear coding = 4/ 4-cp+1

CP_L ⟶ 1 , CP_H ⟶ 4

4) bp(80-220)

Linear growth = 220/220-bp

BP_L= 220/220-80= 1.5 , **BP_H** =220/220-219= 220

5) Chol

Static growth

Linear growth

Exponential growth

Static growth <200 below

$W_{CH=1}$

Above 200> Exponential growth = $400*e^{0.1*chol}/100$

$= 4*e^{0.1*chol}$

Falts <200 lower

200-240 high (Border line)

240 High

6) Fbs

Fetal bovine serum

Probability (No much impact on heart)

0.2 to each.

7) RES

Probability 0.3 to each.

8) Thal (80-220) ⇨ 20 extra for saving exponential time .

Linear coding = 220(H)/ 220- Thal.

$W_{Thal}^{(L)=1.3}$

$W_{Thal}^{(H)=220}$ (consider)

9) Exang

P.F (not much big impact)

W_{exang}=0.25

11) Ca = Null (not consider)

5. SIMULATION MODEL

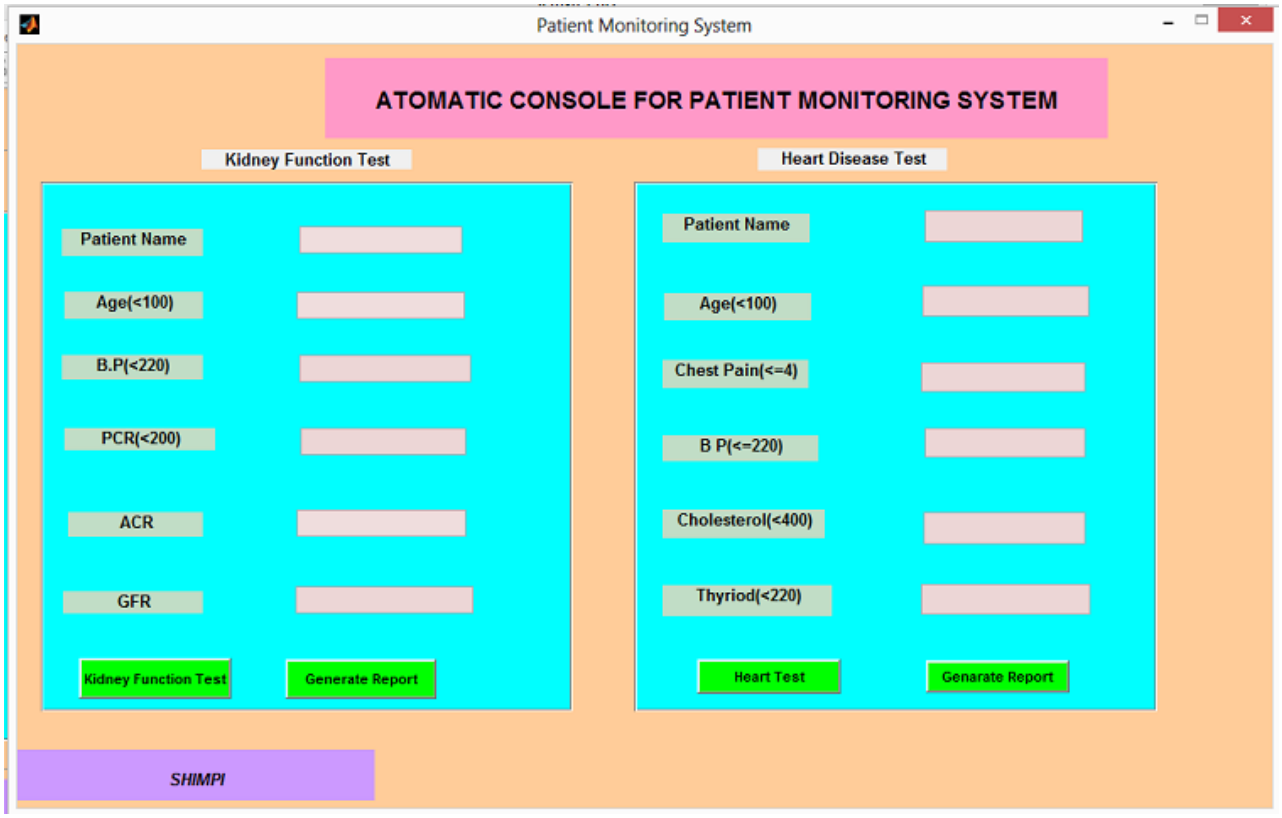


Fig 3:-Graphical User Interface developed for diseases prediction by mean of clustering & apriori algorithm

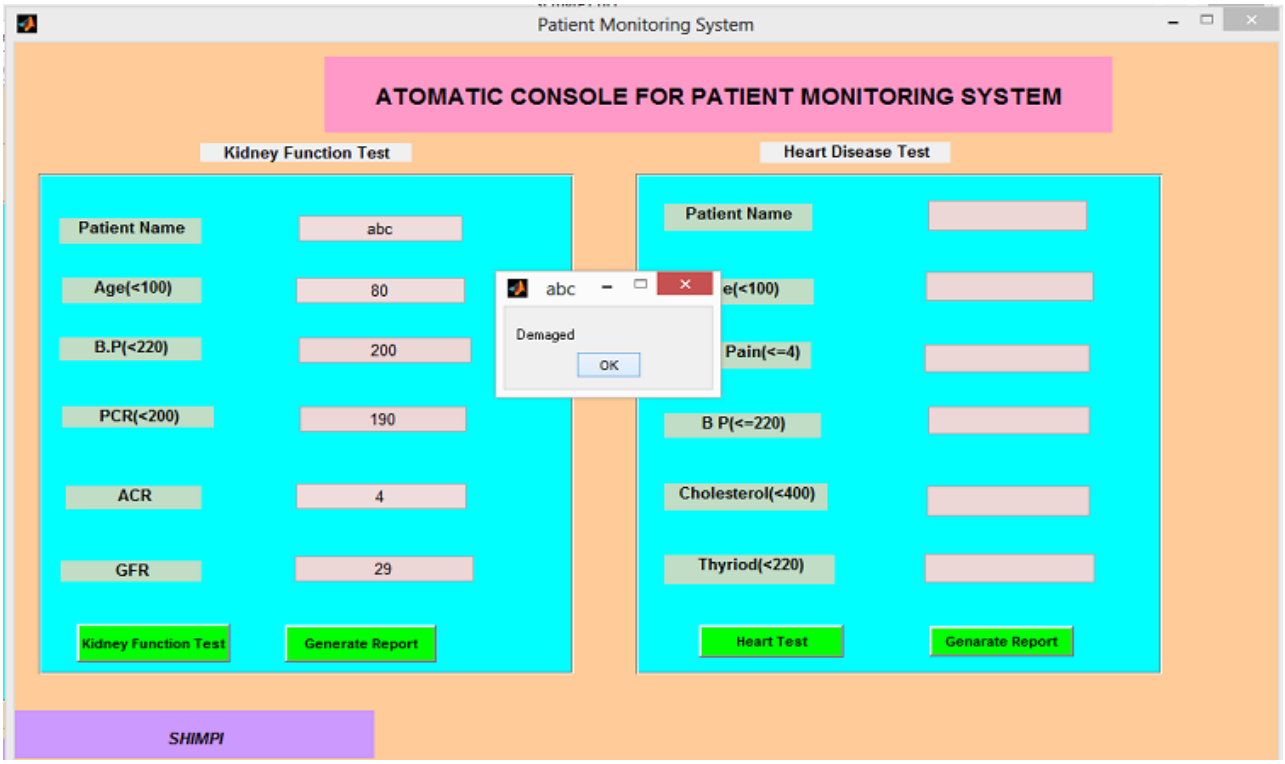


Fig4:-Kidney Function Result

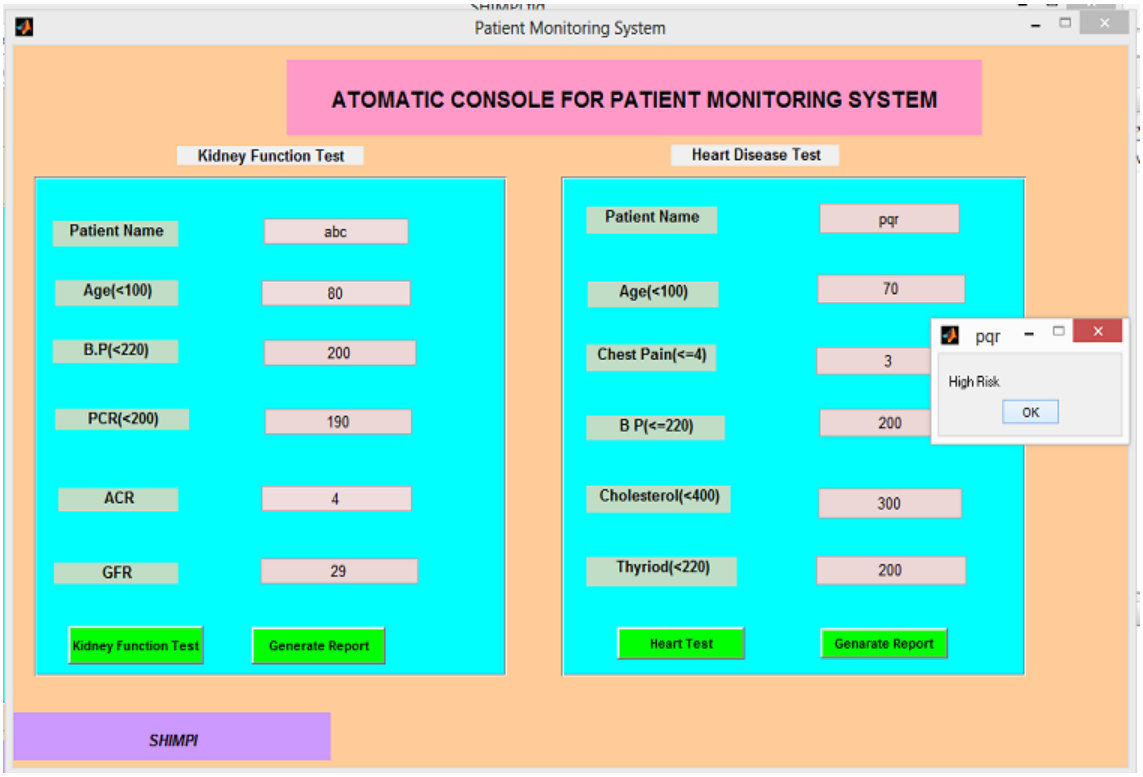


Fig5:-Heart Diseases Test

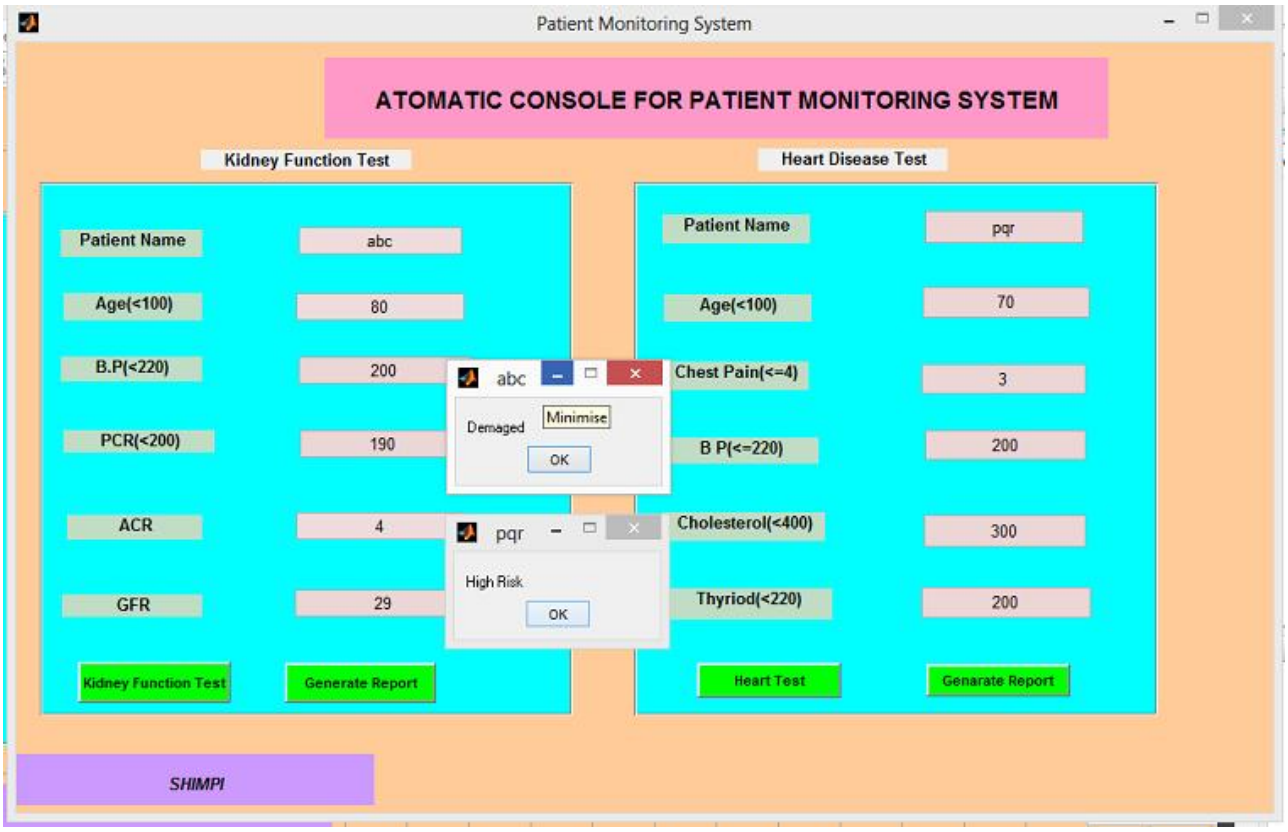


Fig 6:-Kidney Function & Heart Disease Test

Table 2:- Testing Under simulation

Serial No.	Patient Name	Age	Chest Pain	Bp	Cholestrol	Thyriod	Actual Report	Our Mdal Test	Match
1	M1	99	4	180	380	200	Hishest risk	Highest	yes
2	M2	32	2	150	350	200	Highest risk	Medium	no
3	M3	2	4	110	150	50	Low risk	Low	yes
4	M4	82	3	170	399	219	Highest risk	Highest	yes
5	M5	26	4	210	300	200	Medium risk	Highest	no
6	M6	24	1	123	200	100	Low risk	Low	yes
7	M7	26	1	120	200	100	Low risk	Low	yes

The above result shows the good accuracy to detect the heart disease

Probability of detecting the disease as per actual report

P (e) =N(e)/N(s)=5/7= .71

In term of percentage 71 % which is very exciting for our model.

6. CONCLUSION

This console has greatly predicts the person for predictive disease. A very clear categorization over their effectiveness. It provide the user or patient to monitor them self for their report. It also capable to provide the suggested medicine if find problem and also suggest to consult a doctor in critical situation.

7. FUTURE SCOPE

A large population needs a great demand of doctor. But their deficiency create problem so console plays very important role for some extent. Facilitate the users to predict them self even if they are at remote location and very hard to reach doctors regularly. Less cost and time saving if we integrate it to web portals

8. REFERENCES

[1] World Health Organization. 2008 May 2011]; Available from.http://www.who.int/mediacentre/factsheets/fs310_2008.pdf

[2] Vikas Chaurasia, Saurabh Pal Data Mining Approach to Detect Heart Dieses International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739 © Helvetic Editions LTD, Switzerland www.elvedit.com

[3] My Chau Tu, Dongil Shin, Dongkyoo Shin ,"Effective Diagnosis of Heart Disease through Bagging Approach", 2nd International Conference on Biomedical Engineering and Informatics,2009.

[4] M.Akhil jabbar, Dr.Priti Chandra, Dr.B.L Deekshatulu " Heart Disease Prediction System using Associative Classification and Genetic Algorithm". ICECIT, 2012

[5] N. Aditya Sundar, P. Pushpa Latha, M. Rama Chandra ," Performance analysis of classification data mining techniques over heart disease data base," International journal of engineering science & advanced technology Volume-2, Issue-3, 470 – 478 .

[6] C Y Hsu, J D Ordoñez "The risk of acute renal failure in patients with chronic kidney disease". 2 April 2008

[7] Mohammed Abdul Khaleel, Sateesh Kumar Pradhan," Finding Locally Frequent Diseases Using Modified Apriori Algorithm," International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013 .

[8] Chris Ding ,Xiaofeng He," K-means Clustering via Principal Component Analysis,Chris", Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[9] K.R. Lakshmi, M.Veera Krishna, S.Prem Kumar " Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability",International Journal of Scientific and Research Publications.

[10] Boshra Bahrami, Mirsaeid Hosseini Shirvani "Prediction and Diagnosis of Heart Disease by Data Mining Techniques" Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February - 2015